IN SILICO SUBSTRATE BINDING PROFILING FOR SARS-CoV-2 MAIN PROTEASE (M^{PRO}) USING HEXAPEPTIDE SUBSTRATES.

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

MASTER OF SCIENCE OF RHODES UNIVERSITY

by

Coursework and Thesis

in

Bioinformatics and Computational Molecular Biology

Research Unit in Bioinformatics (RUBi)

Department of Biochemistry and Microbiology

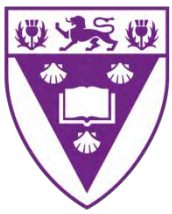
Faculty of Science

by

Sophakama Zabo

15z3394

February 2022



RHODES UNIVERSITY

Where leaders learn

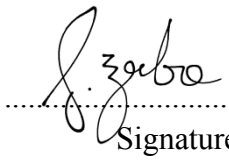
ABSTRACT

COVID-19, as a disease resulting from SARS-CoV-2 infection, and a pandemic has had a devastating effect on the world. There are limited effective measures that control the spread and treatment of COVID-19 illness. The homodimeric cysteine main protease (M^{pro}) is crucial to the life cycle of the virus, as it cleaves the large polyproteins 1a and 1ab into matured, functional non-structural proteins. The M^{pro} exhibits high degrees of conservation in sequence, structure and specificity across coronavirus species, making it an ideal drug target. The M^{pro} substrate-binding profiles remain, despite the resolution of its recognition sequence and cleavage points (Leu-Gln↓(Ser/Ala/Gly)). In this study, a series of hexapeptide sequences containing the appropriate recognition sequence and cleavage points were generated and screened against the M^{pro} to study these binding profiles, and to further be the basis for efficiency-driven drug design. A multi-conformer hexapeptide substrate library comprising optimised 81000 models of 810 unique sequences was generated using RDKit within the context of python. Terminal capping with ACE and NMe was effected using SMILES and SMARTS matching. Multiple hexapeptides were complexed with chain B of crystallographic M^{pro} (PDS ID: 6XHM), following the validation of chain B for this purpose using AutoDock Vina at high levels of exhaustiveness (480). The resulting Vina scores ranged between -8.7 and -7.0 $kcal.mol^{-1}$, and the reproducibility of best poses was validated through redocking. Ligand efficiency indices were calculated to identify substrate residues with high binding efficiency at their respective positions, revealing Val (P3), Ala (P1′); and Gly and Ala (P2′ and P3′) as leading efficient binders. Binding efficiencies were lowered by molecular weight. Substrate recognition was assessed by mapping of binding subsites, and M^{pro} specificity was evaluated through the resolution of intermolecular interaction at the binding interface. Molecular dynamics simulations for 20 ns were performed to assess the stability and behaviour of 132 M^{pro} systems complexed with KLQ*** substrates. Principal component analysis (PCA), was performed to assess

protein motions and conformational changes during the simulations. A strategy was formulated to classify and evaluate relations in the M^{pro} PCA motions, revealing four main clades of similarity. Similarity within a clade (Group 2) and dissimilarity between clades were confirmed. Trajectory visualisation revealed complex stability, substrate unbinding and dimer dissociation for various M^{pro} systems.

DECLARATION

I, Sophakama Zabo, declare that this thesis submitted to Rhodes University is my original work and has never been submitted for a degree or diploma at this or any other institution.



..... February 2022

Signature Date

DEDICATION

This mini-thesis is dedicated to the loving memory of my grandmother,

VUYISWA

Whose life was a great inspiration.

ACKNOWLEDGEMENTS

To God be all the glory, for the great things He has done. I sing my praises to the Lord, God Almighty, for His daily blessings, the cognition and strength He bestowed upon me.

I would like to thank the National Research Fund and LINK-SA Trust for funding my studies and the Centre for high performance computing (CHPC) for providing the computational resources to perform this study.

I express my heartfelt gratitude to my mom, Ms Zintombi Zabo, whose immeasurable love, care and support provided great sustenance through the duration of this endeavour. I thank and greatly appreciate you!

I acknowledge and thank the members of the CMCDD and RUBi groups for the endless motivation, support, assistance, constructive discussions and warm embrace. I also thank my lecturers for their guidance and support.

To my classmates, Amira, Lillian, Mbunge, Prudence, Thomas and Nabeelah, thank you so much for the best time one could get during a pandemic. Your presence and friendship during this period was a great blessing I will cherish forever. A special thanks to Nabeelah Dudha, Yolanda Novokoza, Nobathembu Ndzengu and Amira Damji.

My sincerest gratitude and appreciation to my supervisor Professor Kevin Lobb for all the advice, guidance, patience, kindness, and understanding he has shown and given me for the duration of this project. Thank you for always availing yourselves in times of need; guiding and teaching me scientific methodologies and research approaches with patience and enthusiasm.

Thank you!

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

WEBSERVERS AND SOFTWARE TOOLS USED

AutoDock Vina

AutoDockTools

BIOVIA Discovery Studio 2020 Client

CHPC

Grace (Xmgrace)

GROMACS

JupyterHub

OpenBabel

Pandas

Perl

PyMOL

Python

RCSB Protein Data Bank

RDKit

RStudio

Seaborn

Visual Molecular Dynamics

WPS Spreadsheet

XTB software

LIST OF ABBREVIATIONS AND ACRONYMS

2D	Two dimension
$3CL^{pro}$	chymotrypsin-like protease
3D	Three dimension
ACE	acetyl
ACE2	Angiotensin-converting enzyme 2
ADMET	Absorption, Distribution, Metabolism, Elimination and Toxicity
ARDS	Acute Respiratory Distress Syndrome
BEI	binding efficiency index
BFGS	Broyden-Fletcher-Goldfarb-Shanno
CADD	computer-aided discovery/design
CHPC	Centre for high performance computing
COVID-19	Coronavirus Disease 2019
CPU	Central Processing Unit
CT	Computed Tomography
C-terminus	carboxyl-terminus
DMV	Double-membrane vesicles
ER	Endoplasmic reticulum
ERGIC	ER–Golgi intermediate compartment
G + C	guanine + cytosine
GA	genetic algorithm
GPU	Graphics Processing Unit
GROMACS	GROningen MAchine for Chemical Simulation
HA	heavy atoms
ICU	Intensive care unit
IMV	invasive mechanical ventilation
LE	Ligand efficiency
LELP	logP/ligand efficiency
LLE	ligand-lipophilicity efficiency
MC	Monte Carlo
MD	Molecular dynamics
MERS-CoV	Middle East Respiratory Syndrome Coronavirus
mol	MDL Molfile
MPI	Message Passage Interface
M^{pro}	main protease
mRNA	messenger ribonucleic acid
NIV	non-invasive
NME	methylamine
NMR	nuclear magnetic resonance
NPT	constant Number of atoms, constant Pressure and constant Temperature

nsp	non-structural protein
N-terminus	amino-terminus
NVT	constant Number of atoms, Volume and Temperature
ORFs	Open Reading Frames
PBC	periodic boundary conditions
PBS	Portable Batch System
PCA	Principal Component Analysis
PDB	Protein Data Bank
PDBQT	Protein Data Bank, Partial Charge (Q), & Atom Type (T)
PEI	percentage efficiency index
pp1a	polyprotein 1a
pp1ab	polyprotein 1ab
PSA	Polar Surface Area
QSAR	Quantitative Structure-Activity Relationship
Rg	Radius of gyration
RMSD	Root Mean Square Deviation
RMSF	Root Mean Square Fluctuation
RNA	ribonucleic acid
SARS-CoV	Severe Acute Respiratory Syndrome Coronavirus
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus-2
SD	standard deviation
SDF	Standard Database Format
SEI	surface-binding efficiency index
SMARTS	SMILES arbitrary target specification
SMILES	Simplified Molecular Input Line Entry System
SO	Swarm optimization
TDT	THOR Data Tree
TMPRSS2	Transmembrane protease serine 2
UFF	Universal Force Field
UTR	Untranslated Regions
VMD	Visual Molecular Dynamics
WHO	World Health Organization
XYZ	XMOL molecule model
α	alpha
β	beta
γ	gamma
δ	delta

TABLE OF AMINO ACIDS

Amino Acid	Three letter code	One letter code
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic Acid	Asp	D
Cysteine	Cys	C
Glutamic Acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

CHAPTER ONE

LITERATURE REVIEW

1.1 BACKGROUND

The novel coronavirus disease 2019 (COVID-19) is a pulmonary disease that is caused by the infection of a virus called severe acute respiratory syndrome-coronavirus 2 (SARS-CoV-2) (Liu *et al.*, 2020). Since the emergence of COVID-19 in December 2019 out of Wuhan, China, the disease has spread globally at a rapid rate and was quickly declared a public health emergency of international concern and shortly after, a pandemic by March 2020, due to high transmission rates and fatalities associated with the disease (Khan *et al.*, 2020; Zhang *et al.*, 2020). As of February 2022, there have been more than 396 million infection cases reported globally, along with more than 5.74 million confirmed deaths and approximately 315 million recovered cases across 216 countries, areas or territories (WHO, 2022; Statista, 2022).

Since the initial outbreak of the coronavirus pandemic, COVID-19 has demonstrated a massive impact on the global economy, exerted a great strain on public health resources and predominantly threatened public health and the livelihood of millions of people. Globally, countries were compelled to implement national lockdowns as means to control the spread of COVID-19. These lockdowns entailed the suspension of mass gatherings and the enforcement of social isolation, including academic progression, religious assemblies, the closure of non-essential business sectors, border shutdowns and travel restrictions (Nicola *et al.*, 2020). Subsequently, the economic disruption and/or inactivity landed many countries in critical economic instabilities, accompanied by hikes in unemployment rates and economic recession, and even the decline of the standard of living for people living in specific regions located in Asia, Africa, Europe and South America (Buheji *et al.*, 2020).

According to economists, the threat of the COVID-19 crisis to global financial stability is such that has the potential to be extremely adverse if the pandemic persists and could lead to a global financial crisis similar to, or even worse than the financial crisis of 2007–2009, and this will leave millions in devastating living conditions with little to no resources to support their livelihood (International Monetary Fund and Capital Markets Department, 2020; Adrian and Natalucci, 2020a; Adrian and Natalucci, 2020b; Bhuiyan *et al.*, 2020). Developing

countries are very vulnerable to these adverse economic conditions, and the socio-economic impact of the COVID-19 pandemic is already amplified in these countries (Bhuiyan *et al.*, 2020; Abouzzohour, 2020). Moreover, being low-income countries, a majority of developing countries lack the basic health care facilities required to combat the outbreak of COVID-19 (Khan *et al.*, 2020).

1.2 SEVERE ACUTE RESPIRATORY SYNDROME-CORONAVIRUS 2

Coronaviruses are a diverse group of RNA viruses that cause respiratory diseases in birds, humans and other higher mammals (Perlman and McIntosh, 2015; Milewska *et al.*, 2020) and which belong to the subfamily *Coronavirinae*, in the family Coronaviridae and the order Nidovirales (Mousavizadeh and Ghasemi, 2020; Kumar *et al.*, 2020a). Coronaviridae is a diverse family of enveloped viruses consisting of large single-stranded, positive-sense RNA genomes of around 27–32 kb (Mousavizadeh and Ghasemi, 2020; Tu *et al.*, 2020) which are typically composed of a 5′-methylguanosine cap at the beginning, a 3′-poly-A tail at the end, and a total of 6-10 genes in between (Tu *et al.*, 2020). The viral genome is characterised as having a high frequency of genomic recombination and mutation (Khan *et al.*, 2020) and is the largest among RNA viruses, with G + C contents varying from 32% to 43% (Mousavizadeh and Ghasemi, 2020).

There are four main classes of coronaviruses namely alpha, beta, gamma, and delta (Shereen *et al.,* 2020). The alpha and betacoronaviruses are believed to infect humans and mammals, whereas the delta and gammacoronaviruses seem to infect bird species (Cascella *et al.*, 2020). SARS-CoV-2 belongs to the betacoronavirus class, together with severe acute respiratory syndrome (SARS) coronavirus (SARS-CoV) and the Middle East respiratory syndrome (MERS) coronavirus (MERS-CoV). SARS-CoV-2 shares 82% RNA genome identity to that of SARS-CoV, making both viruses members of the clade b of the genus *Betacoronavirus* and hence, have similar names (Liu *et al.*, 2020; Zhang *et al.,* 2020). The RNA genome identity between SARS-CoV-2 and MERS-CoV is about 50% (Kim *et al.*, 2020). The enveloped viral particles of coronaviruses are minute in size, ranging between 65–125 nm in diameter (Shereen *et al.*, 2020). Coronaviruses are sensitive to ultraviolet rays and heat, with high temperatures decreasing replication and/or activity at about 27°C. On the contrary, some species have shown resistance to cold temperatures even below 0°C (Cascella *et al.*, 2020). The inactivation temperature of SARS-CoV-2 is yet to be well elucidated. In addition, these viruses can be effectively inactivated by lipid solvents including ether (75%), ethanol,

chlorine-containing disinfectant, peroxyacetic acid, and chloroform (but not chlorhexidine) (Cascella *et al.*, 2020).

The structure of the SARS-CoV-2 virion is spherical or elliptic and often exhibit pleomorphism (Mousavizadeh and Ghasemi, 2020; Cascella *et al.,* 2020). Similar to other coronaviruses, the SARS-CoV-2 virions have a crown-like appearance under an electron microscope due to the presence of the club-shaped glycoprotein projections referred to as the spike protein (figure 1.1) (Mousavizadeh and Ghasemi, 2020; Cascella *et al.*, 2020). The name coronavirus is owing to this crown-like appearance (*coronam* is the Latin term for crown) (Cascella *et al.*, 2020).

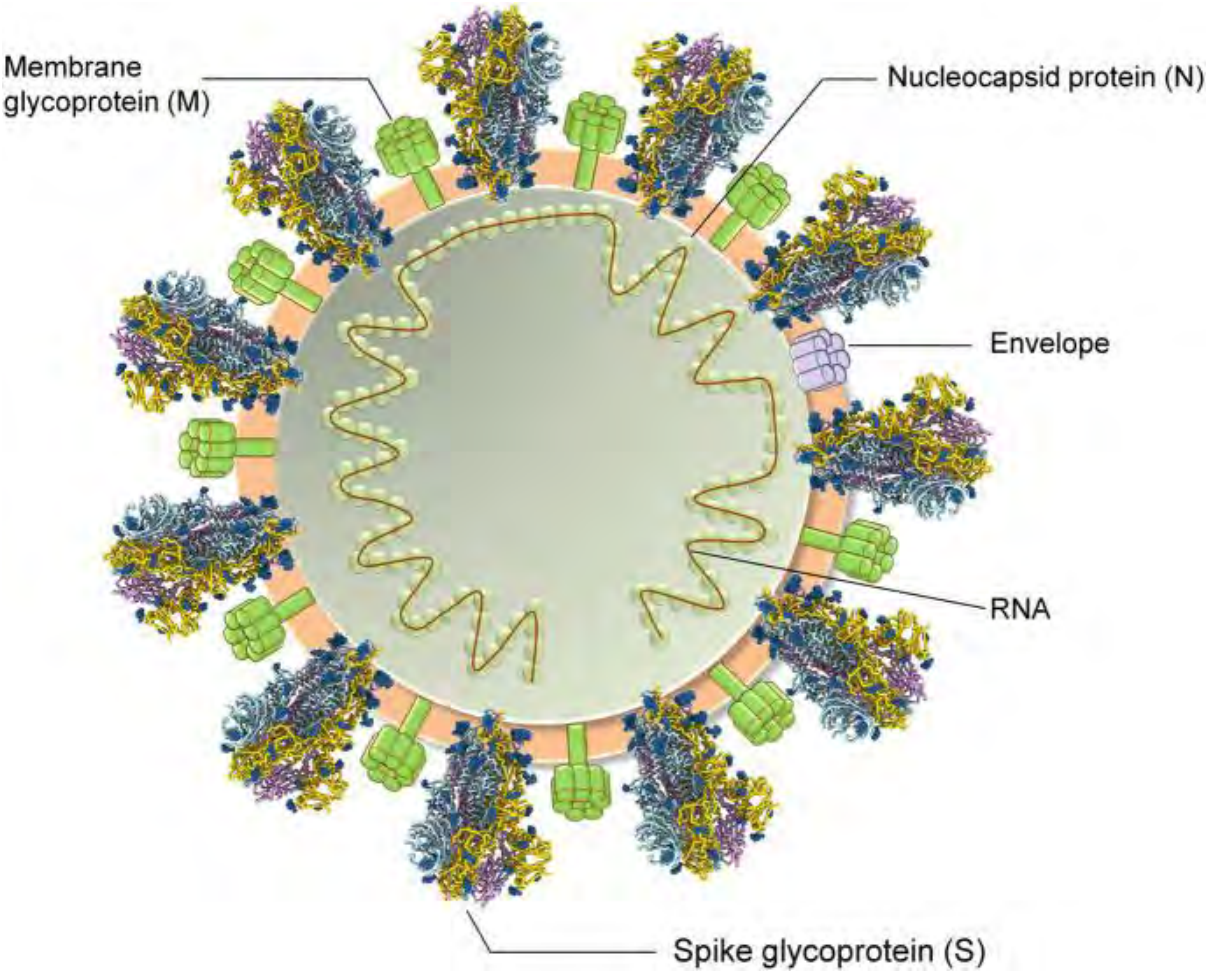


Figure 1.1. The structure of SARS-CoV-2. SARS-CoV-2 has surface viral proteins, namely, spike glycoprotein (S), which mediates interaction with cell surface receptor ACE2. The viral membrane glycoprotein (M) and envelope (E) of SARS-CoV-2 are embedded in the host membrane-derived lipid bilayer encapsulating the helical nucleocapsid comprising viral RNA. Adapted with permission from Kumar *et al.*, 2020b.

There are three main structural proteins on the coronavirus membrane. The spike protein is a homotrimeric, type I membrane glycoprotein that constitutes peplomers that protrude the viral surface (Mousavizadeh and Ghasemi, 2020; Walls *et al.*, 2020). Like in many other

coronaviruses, the majority of the spike protein is exposed to the exterior of the virion, followed by a short transmembrane domain and a short cytoplasmic tail rich in cysteine residues at the C-terminus (Woo *et al.*, 2010). Being the prominent feature of the viral membrane, the spike proteins are the main inducers of the neutralising antibodies (Mousavizadeh and Ghasemi, 2020). More importantly, the spike protein plays a crucial role in viral infection into host cells, as it mediates the fusion process between viral and host membranes and ensures efficient cell entry of coronavirus particles (Alsaadi and Jones, 2019). The details of this process and the interactions involved will be discussed in the next section.

The membrane glycoprotein is a type III transmembrane glycoprotein and is the most abundant glycoprotein in the coronavirus membrane (Alsaadi and Jones, 2019). The protein spans the membrane bilayer three times, with the short N-terminal domain on the exterior of the virion, and the long C-terminal domain inside the cytoplasm of the virion (Mousavizadeh and Ghasemi, 2020). The membrane glycoprotein is believed to play a crucial role in the intracellular formation of virions, particularly the budding process of coronaviruses (Mousavizadeh and Ghasemi, 2020; Bianchi *et al.*, 2020). During assembly of the authentic virions, the membrane glycoprotein interacts with itself, the nucleocapsid protein, envelope protein and the spike protein (Alsaadi and Jones, 2019). Moreover, the activity of membrane glycoprotein is independent of the spike protein. According to Mousavizadeh and Ghasemi (2020), the coronavirus replicates and forms spikeless (devoid of spike protein) non-infectious virions containing membrane glycoproteins when exposed to tunicamycin.

The envelope protein is a small hydrophobic integral membrane protein which is generally a minor component of the virus membrane in all coronaviruses groups (Alsaadi and Jones, 2019). The protein has an N-terminal domain, a long α-helical transmembrane domain and a C-terminal hydrophilic domain (Alsaadi and Jones, 2019). The envelope protein is crucial to the pathogenicity of SARS-CoV-2 as it promotes viral assembly and release (Cascella *et al.*, 2020), achieved through the induction of membrane curvature which leads to membrane scission of the budding virus particle and its eventual release. The membrane curvature induced by the envelope protein is established such that the co-expression of membrane and envelope proteins (and even the spike proteins, if spike protein co-expression took place) is sufficient for the efficient formation of viral particles (Alsaadi and Jones, 2019).

The viral membrane encapsulates the single-stranded RNA associated with a nucleoprotein within a capsid composed of matrix protein (Mousavizadeh and Ghasemi, 2020). The genome

is 29 891 nucleotides long with a G + C content of 38% and encodes 9 860 amino acids (Guo *et al.*, 2020). The viral RNA genome comprises of two flanking, untranslated regions (UTR) and open reading frames (ORFs) arranged in the order: 5′-replicase (ORF1ab) - structural proteins [Spike (S) - Envelope (E) - Membrane (M) - Nucleocapsid (N)]−3′ and nonstructural ORFs (figure 1.2) (Wu *et al.*, 2020b; Guo *et al.*, 2020). In a typical coronavirus genome, there can be at least six ORFs (Mousavizadeh and Ghasemi, 2020). The SARS-CoV-2 genome encodes at least 27 proteins, which include 16 non-structural proteins (nsp1-10, nsp12-16), the four structural proteins(S, E, M, and N) and 8 accessory proteins (ORF3a, ORF3b, ORF6, ORF7a, ORF7b, ORF8, ORF9b, and ORF14) (Guo *et al.*, 2020; Mousavizadeh and Ghasemi, 2020).

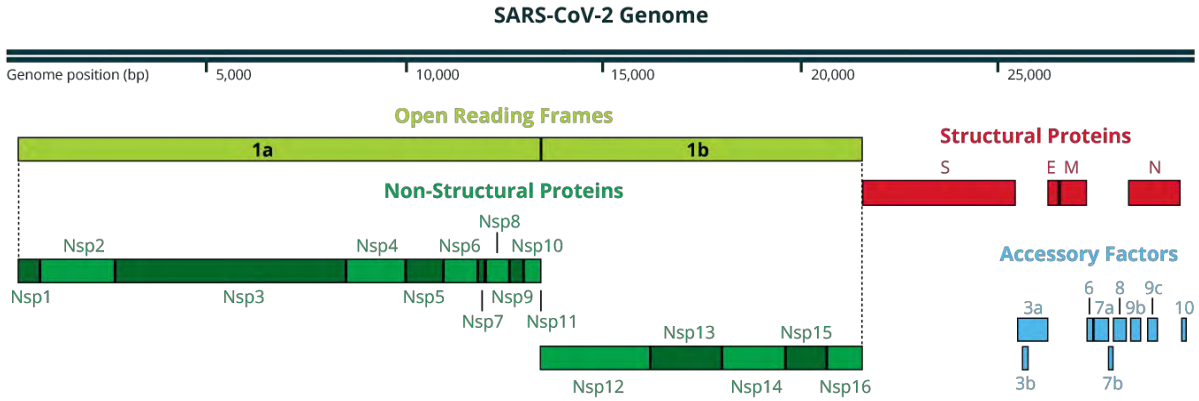


Figure 1.2. The genomic arrangement of SARS-CoV-2. The size of the coronavirus genome ranges from 26 to 32 kb and comprises 6–11 open reading frames (ORFs) encoding 9680 amino acid polyprotein. The first ORF comprises approximately 67% of the genome that encodes 16 nonstructural proteins (nsps), whereas the remaining ORFs encode for accessory and structural proteins. The nsps includes two viral cysteine proteases, including papain-like protease (nsp3), chymotrypsin-like, 3C-like, or main protease (nsp5), RNA-dependent RNA polymerase (nsp12), helicase (nsp13), and others likely to be involved in the transcription and replication of SARS-CoV-2. In addition to nsps, the genome encodes for four major structural proteins including spike surface glycoprotein (S), membrane, nucleocapsid protein (N), envelope (E) and accessory proteins like ORFs. Adapted from Boster, 2020 and with permission from Kumar *et al.*, 2020b.

The first ORFs (OFR1a/b) comprise two-thirds of the SARS-CoV-2 genome and encode two large proteins, polyproteins 1a (pp1a) and 1ab (pp1ab). A frameshift between ORF1a and ORF1b guides the synthesis of pp1a and pp1ab which are subsequently processed by virally encoded chymotrypsin-like protease ($3CL^{pro}$) or main protease (M^{pro}) and one or two papain-like protease into non-structural proteins (Chen *et al.*, 2020). Apart from ORF1a and ORF1b, other ORFs on the one-third of the genome near the 3′-terminus encode the four main structural proteins (ORF10, ORF11), while the remaining ORFs encode accessory proteins, such as 3a/b protein, and 4a/b protein (Mousavizadeh and Ghasemi, 2020, Chen *et*

al., 2020; Cascella *et al.*, 2020). Different species of coronaviruses present special structural and accessory proteins translated by dedicated subgenomic RNAs (Chen *et al.*, 2020; Cascella *et al.*, 2020).

The viral genome also serves as the template for replication and transcription. These processes are mediated by nsp12, which displays RNA-dependent RNA polymerase activity (Kim *et al.*, 2020). Once efficient host infection is established, the transcription commences through the replication-transcription complex in double-membrane vesicles and *via* the synthesis of subgenomic RNAs sequences (Cascella *et al.*, 2020). During transcription, negative-sense RNA intermediates are synthesised as the templates for the synthesis of positive-sense genomic RNA and subgenomic RNAs. The genomic RNA is packaged by the structural proteins to assemble progeny virions, while the shorter subgenomic RNAs encode conserved structural proteins and several accessory proteins (Kim *et al.*, 2020). Transcription termination occurs at transcription regulatory sequences, located between the ORFs (Cascella *et al.*, 2020).

1.3 THE PATHOLOGY AND VIRULENCE MECHANISMS OF SEVERE ACUTE RESPIRATORY SYNDROME-CORONAVIRUS 2

The first cases of the COVID-19 disease were presumed to spread *via* animal-to-human transmission since they were associated with the Huanan Seafood Wholesale Market of Wuhan and the fact that betacoronaviruses were known to infect higher mammals. Nonetheless, the subsequent cases were not directly linked to the market and the contagion mechanism was concluded to primarily involve human-to-human transmission, and symptomatic people were the most frequent source of the COVID-19 spread. It was shortly discovered that presymptomatic and asymptomatic individuals contributed to the spread of the disease, accounting for about 80% of COVID-19 transmission (Cascella *et al.*, 2020). It was also established that close contact is essential for successful SARS-CoV-2 transmission, but aerosol transmission is also possible in case of protracted exposure to elevated aerosol concentrations in closed spaces (Sironi *et al.*, 2020). Other possible modes of transmission include contact with contaminated objects and surfaces such as plastic (2-3 days), stainless steel (2-3 days), cardboard (1 day) copper (up to 4 hours) (Orleans and Manchikanti, 2020; Cascella *et al.*, 2020).

1.3.1 SEVERE ACUTE RESPIRATORY SYNDROME-CORONAVIRUS 2 INFECTION

Much like other respiratory pathogens, efficient SARS-CoV-2 infection occurs *via* spraying

respiratory droplets (5-10 μm in diameter) from infected individuals through their cough or sneeze (Cascella *et al.*, 2020). Once inhaled, the virus particles are transported to the airway where they invade the airway epithelial cells. To enter host cells, coronaviruses first bind to a cell membrane receptor for viral attachment, subsequently enter endosomes, and eventually fuse viral and lysosomal membranes (Shang *et al.*, 2020). The surface-anchored spike protein mediates the entire process of host cell invasion. In mature virions, the spike protein presents as a trimer with two functional subunits, S1 and S2. To facilitate the fusion of viral and host membranes, the spike protein requires proteolytic activation at the S1/S2 boundary for S1 to dissociate and allow S2 to undergo the essential structural change. The host proteases that mediate this entry-activating proteolysis include the surface-anchored serine protease, the transmembrane protease serine 2 (TMPRSS2), and the lysosomal protease cathepsins. S1 contains a receptor-binding domain that recognises angiotensin-converting enzyme 2 (ACE2) as its specific receptor. The receptor-binding domain is constantly changing conformations to evade immune response. The binding of S1 to ACE2 facilitates viral attachment to the surface of host cells. S2 is further cleaved at the S2' site and activated by TMPRSS2 in a process called protein priming. Together, these actions result in viral-host membrane fusion and contribute to the rapid spread of COVID-19, as well as the severe clinical manifestation of the SARS-CoV-2 exhibited by infected individuals (figure 1.3) (Hoffman *et al.*, 2020; Shang *et al.*, 2020; Guo *et al.*, 2020). Furthermore, the basic reproduction number (R_0) for SARS-CoV-2 is 2.2, meaning that each patient transmits the infection to an additional 2.2 individuals (Cascella *et al.*, 2020).

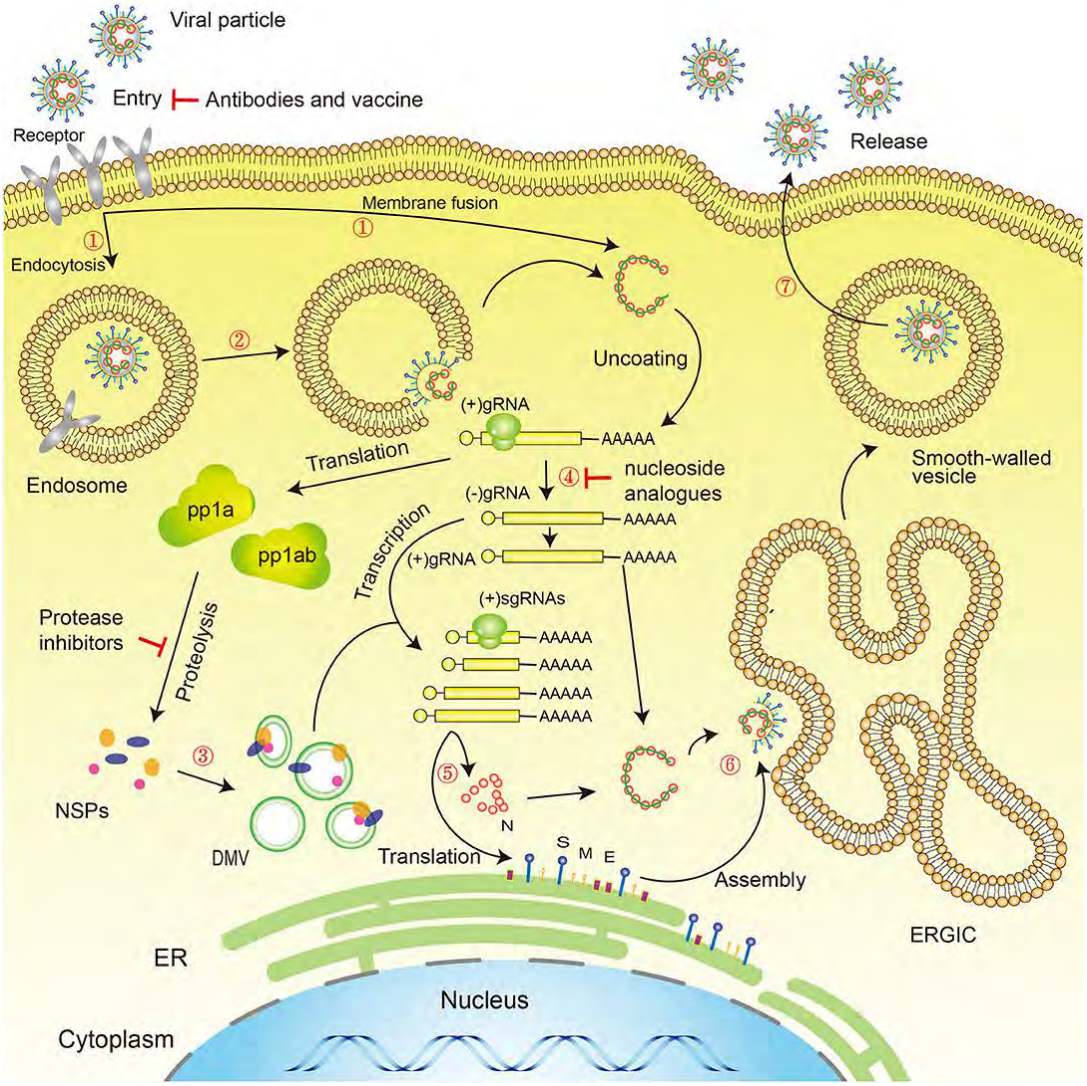


Figure 1.3. Novel coronavirus life cycle. Life cycle: **(1)** First, the virus binds to receptors on the surface of the host cell through the S-protein and is endocytosed or directly fused with the host cell membrane into the cell; **(2)** Next, the lysosome degrades the lipid membrane and protein envelope on the exterior of the virus (endocytosis only); **(3)** Viral RNA is released into the cell, where *ORF1a* and *ORF1ab* are translated into pp1a and pp1ab, which in turn are cleaved by proteases encoded by *ORF1a* to produce multiple NSPs, forming the replication/transcription complex; **(4)** At the same time as the previous step, viral RNA continues to use the cell for replication; **(5)** The replicated viral RNA undergoes discontinuous transcription under the action of the replication/transcription complex to produce subgenomic RNA, which is translated into structural proteins in the cell's endoplasmic reticulum; **(6)** The resulting structural proteins assemble in the ER-Golgi intermediate compartment (ERGIC) to form the nucleocapsid and viral envelope; **(7)** Finally, smooth-walled vesicles containing the nascent virus particles fuse with the cell membrane, releasing the virus particles from the infected cell. **S**, Spike protein; **M**, Membrane protein; **E**, Envelope protein; **N**, Nucleocapsid protein; **NSPs**, Non-structural proteins; **DMV**, Double-membrane vesicles; **ER**, Endoplasmic reticulum; **ERGIC**, ER–Golgi intermediate compartment. Adapted with permission (under the terms of the Creative Commons Attribution License (CC BY)) from Guo *et al.*, 2020.

1.3.2 CLINICAL MANIFESTATIONS OF COVID-19

The clinical spectrum of COVID-19 varies from asymptomatic or paucisymptomatic forms to severe clinical conditions characterized by respiratory failure that necessitates mechanical ventilation and support in an ICU, to multi-organ and systemic manifestations in terms of

sepsis, septic shock, and multiple organ dysfunction syndromes (Cascella *et al.*, 2020). Between infection and symptoms, the incubation period is generally within 3 to 5 days, and even up to 2 weeks depending on the amount of virus that initially entered the body and the general physical health of the infected person (Guo *et al.*, 2020).

Typical clinical symptoms of COVID-19 include fever, fatigue, malaise, dry cough and dyspnea, while atypical clinical symptoms include expectoration, headache, hemoptysis, nausea, vomiting, and diarrhoea (Cascella *et al.*, 2020; Guo *et al.*, 2020). There were reported cases of chemosensory dysfunction, such as smell and taste impairment, associated with COVID-19 in patients displaying flu-like symptoms (Yan *et al.*, 2020). Confirmed mild cases present with symptoms of low fever, dry cough, mild fatigue, sore throat, nasal congestion, headache, muscle pain or other symptoms, without presenting with pneumonia, and the majority of these cases recover after 1 week (Guo *et al.*, 2020; Wang *et al.*, 2020; Cascella *et al.*, 2020).

Severe cases of COVID-19 are characterised by a fever associated with severe dyspnea, respiratory distress, and tachypnea, such that the respiratory rate can increase to 30 breaths/min or more. Hypoxia (blood oxygen saturation $\leq$ 93%) and Acute Respiratory Distress Syndrome (ARDS) (PaO2/FiO2 $\leq$ 100) are other common clinical conditions associated with severe cases. Furthermore, chest imaging using chest radiograph, Computed Tomography (CT) scans, or lung ultrasound scans revealed increases in pulmonary infiltrates, exceeding 50% within 24 to 48 hours (Wang *et al.*, 2020; Cascella *et al.*, 2020). Other radiological features of severe cases include patients presenting with bilateral pulmonary injury characterised by ground-glass opacities in X-ray scans, as well as the CT scans showing SARS-CoV-2 distribution in the subpleural and lobular zones, with the two possibly merged into a sheet or progressing to into bilobal diffuse opacities (Guo *et al.*, 2020).

In critical COVID-19 cases, the patients generally present with hypoxemia, respiratory failure, septic shock, and/or multiple organ dysfunction or failure. After a week of presenting dyspnea, the individual rapidly progresses to ARDS accompanied by septic shock, metabolic acidosis and coagulopathy. With septic shock, the patients usually suffer from persistent hypotension despite volume resuscitation. This clinical condition is associated with increased mortality, circulatory, and cellular/metabolic abnormalities such as serum lactate levels rising even greater than 2 mmol/L. Extrapulmonary manifestations and systemic complications are also prevalent in critical cases, demonstrated through injuries to the kidney, heart, and other

organs, and even multiple organ failure. These clinical manifestations suggest that SARS-CoV-2 infection, in addition to affecting the respiratory organs, also have clinical presentations that involve invasion of other organs. Of note, some tissue in select organs, such as renal tubular cells, Leydig cells, and cells in seminiferous ducts in testis, is very permissive to SARS-CoV infection due to high expression of ACE2, allowing for direct viral attachment and invasion of such cells and the subsequent damage to the kidneys and testicular tissue of a patient. Studies have confirmed that renal insufficiency is common in patients with COVID-19, which may be one of the main causes of COVID-19 eventually leading to multiple organ failure and even death (Guo *et al.*, 2020; Cascella *et al.*, 2020).

1.3.3 CURRENT TREATMENT/MANAGEMENT OF COVID-19

At present, there are three protective COVID-19 vaccines for primary and booster vaccinations, approved or authorised by the Centre for Disease Control and Prevention. These vaccines include the BNT162b2, mRNA-1273 and Ad26.COV2.S, which were developed by Pfizer-BioNTech, Moderna and Johnson & Johnson/Janssen, respectively. The BNT162b2 vaccine is a nucleoside-modified RNA vaccine that induces immune response and antibody production against the wild-type and beta variant of SARS-CoV-2, by expressing the full-length prefusion spike protein (Liu *et al.*, 2021; Falsey *et al.*, 2021). The mRNA-1273 vaccine is a lipid nanoparticle-encapsulated mRNA-based vaccine that encodes the prefusion stabilized full-length spike protein of SARS-CoV-2, resulting in an immune response that protects against SARS-CoV-2 infection and lowers the severity of COVID-19 symptoms (Baden *et al.*, 2021). The Ad26.COV2.S vaccine is a recombinant, replication-incompetent adenovirus serotype 26 (Ad26) vector encoding a full-length and stabilized SARS-CoV-2 spike protein (Sadoff *et al.*, 2021). The Ad26.COV2.S vaccine has been shown to have high risks of adverse events and is thus less preferable, in comparison to the BNT162b2 and mRNA-1273 vaccines.

In addition to the vaccines, there is one FDA-approved drug for the treatment of COVID-19 symptoms called remdesivir, or Veklury (trade name). Remdesivir targets the RNA-dependent, RNA polymerase (nsp12) and exhibits inhibitory activity against SARS-CoV, MERS-CoV and SARS-CoV-2 *in vitro* (Beigel *et al.*, 2020). The drug is suitable for children, paediatric patients and adults with SARS-CoV-2 infection, whether hospitalised or not. Remdesivir is administered to soothe all symptomatic manifestations of COVID-19, including mild-to-moderate cases, high risk for progression to severe COVID-19, and severe cases.

There are additional treatments and therapeutic strategies in place to support symptomatic cases of COVID-19 and were designed to primarily address respiratory impairments. These strategies were implemented even before the development and authorisation of vaccines. They help mitigate tissue injury and damage in extrapulmonary manifestations of COVID-19. Intensive care is crucial for dealing with complicated cases of the disease. The first steps in addressing respiratory impairment, hypoxia and ARDS in severe cases incorporate oxygen therapy. Oxygen therapy involves the administration of oxygen using adaptable techniques such as non-invasive (NIV) and invasive mechanical ventilation (IMV) therapy, Heated humidified high-flow therapy, Continuous Positive Airway Pressure therapy, Intubation and Protective Mechanical Ventilation. In critical cases and some severe cases, healthcare experts employ several pharmaceutical therapies to treat adverse cases of COVID-19 clinical manifestations, such as COVID-19 induced ARDS, as well as limit the spread and direct extension of the virus to adjacent organs. These therapies include the use of corticosteroids, the administration of antiviral and immunomodulatory drugs, serotherapy (plasma and antibody therapies), the administration of anticoagulant agents and inflammation inhibitors. These therapies have demonstrated effectiveness in treating COVID-19 symptoms and effect recovery in patients.

Nonetheless, prevention is currently the best strategy to limit the spread of COVID-19. Preventive strategies focus on social distancing, the isolation of patients and the careful clinical care to an infected patient. The WHO and other organizations have issued general recommendations for the public and healthcare personnel to implement during social distancing, medical isolations and quarantines. Despite the implementation of preventative measures, the number of cases continues to rise and COVID-19 continues to claim lives. The ultimate measure for SARS-CoV-2 epidemic control and prevention will be the use of protective vaccines that confer long-term immunity even against multiple variants and strains, as well as therapeutic drugs against SARS-CoV-2 infection and COVID-19 symptoms. Efforts in finding such vaccines and drugs are still ongoing. More compounds are being proposed as potential treatments against COVID-19 (Nhean *et al.*, 2021; Awadasseid *et al.*, 2021). The current COVID-19 management measures, although effective to a certain extent, highlight the urgent need for the development of broad-spectrum antiviral chemotherapies, which specifically target highly conserved proteins, to fight infections against the novel SARS-CoV-2 variants and other coronaviruses (Cascella *et al.*, 2020; Guo *et al.*, 2020; Aleem *et al.*, 2021).

1.4 SARS-CoV-2 MAIN PROTEASE

The SARS-CoV-2 main protease (M^{pro}), also referred to as chymotrypsin-like protease ($3CL^{pro}$), is a cysteine protease and a non-structural protein (nsp5), encoded by OFR1a/b (Ullrich and Nitsche, 2020). As in other coronaviruses, M^{pro} is a homodimer and two subunits are arranged almost perpendicular to each other. Each monomer of M^{pro} consists of 306 amino acid residues, comprising 13 strands and 11 helices distributed among three distinct structural domains (I, II and III), and a long loop (residues 185-200) joining domains II and III (Khan *et al.*, 2020; Wu *et al*, 2020a; Chang, 2010; Zhang *et al.*, 2020). Domains I (8-101 amino acid residues) and II (102-184 amino acid residues) are mainly beta-barrels and display resemblance to chymotrypsin, whereas domain III (201-306 amino acid residues) primarily comprises alpha helices (Zhang *et al.*, 2020). On each monomer, there is a catalytic dyad (His41 and Cys145) situated on the cleft of the chymotrypsin-like double beta-barrel fold between domains I and II (figure 1.4) (Goyal and Goyal, 2020; Ullrich and Nitsche, 2020). The individual monomers are enzymatically inactive and require dimerization for functionality (Goyal and Goyal, 2020).

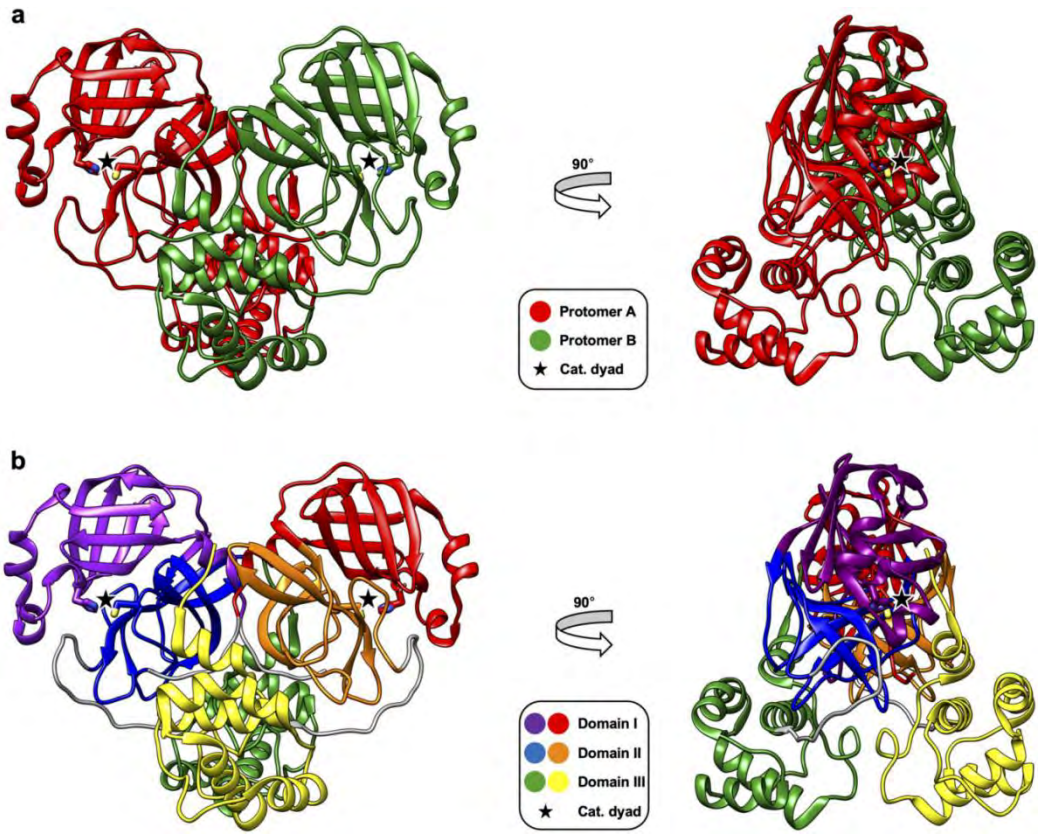


Figure 1.4. The 3D structure of the SARS-CoV-2 M^{pro}. X-ray crystal structure of the M^{pro} homodimer of SARS-CoV-2 (PDB: 6Y2E). Residues of the catalytic dyad (His41/Cys145) are indicated. (a) Monomers are indicated. (b) Domains of each monomer are indicated. Adapted with permission from Ullrich and Nitsche, 2020.

Unlike other cysteine and serine proteases with catalytic triads, M^{pro} consists of a water molecule that occupies the place of the third catalytic residue in the active site (Ullrich and Nitsche, 2020). In addition to the catalytic dyad, there are two deeply buried subsites (S1 and S2) and three shallow subsites (S3-S5) (figure 1.5). The amino acid residues in subsites S1 and S2 participate in hydrophobic and electrostatic interactions, whereas the residues in the shallow subsites S3-S5 tolerate different functionalities (Khan *et al.*, 2020).

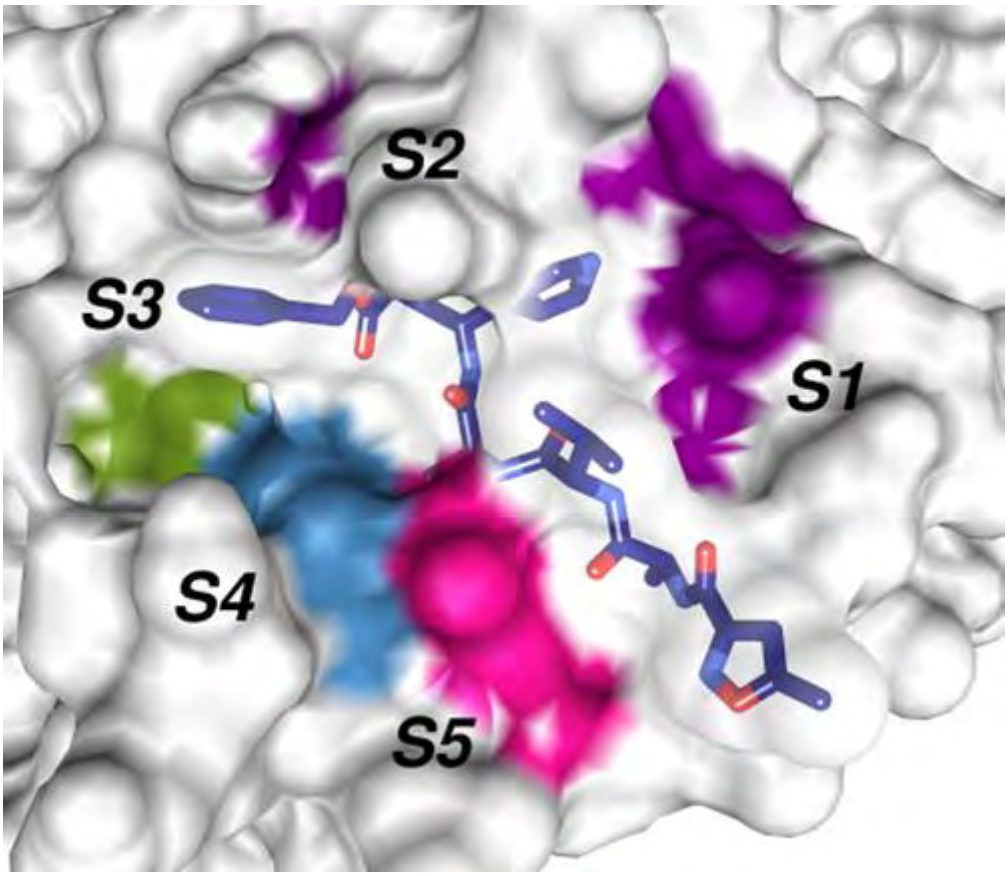


Figure 1.5. The substrate-binding subsites of the SARS-CoV-2 M^{pro}. The surface of SARS-CoV-2 M^{pro}, showing the substrate-binding subsites, colour-coded as follows: purple site S1 and S2, olive green site S3, blue site S4, pink site S5. Adapted with permission from Khan *et al.*, 2020.

The M^{pro} plays a vital role in cleaving the polyproteins translated by the ORF1a and ORF1ab. The M^{pro} is first autocleaved from pp1ab (nps5) to produce a mature protease and then proceeds to cleave downstream nsps at 11 different sites of the pp1ab to release nsp4-nsp16 (Khan *et al.*, 2020). The recognition sequence at most sites was found to be Leu-Gln↓(Ser/Ala/Gly) (↓ shows the cleavage site) (Goyal and Goyal, 2020). The M^{pro} is vital to the life cycle of coronaviruses, as it directly mediates the maturation of the nsps which is essential for viral replication (Khan *et al.*, 2020). Proteolysis mediated by M^{pro} is believed to involve proton abstraction at the cysteine side chain by the histidine's imidazole,

resulting in a thiolate nucleophile which in turn, attacks the amide bond of the substrate. The N-terminal peptide product is released by proton abstraction from histidine before the release of the C-terminal peptide product *via* the hydrolysis of the thioester, and ultimately the dyad is restored (Ullrich and Nitsche, 2020).

1.5 PROBLEM STATEMENT

Despite having zoonotic reservoirs, the virus has exponential transmission rates owing to its efficient human-to-human transmission allowing the pulmonary disease to be widespread. Similar to its predecessors, SARS-CoV and MERS-CoV, SARS-CoV-2 attacks the lower respiratory system to cause viral pneumonia, but it may also affect the gastrointestinal system, heart, kidney, liver, and central nervous system leading to multiple organ failure. These conditions can be fatal, particularly in patients with underlying cardiovascular diseases. Confirmed cases continue to rise rapidly around the world, despite the availability of vaccines. As of February 2022, there are more than 10 billion doses of vaccines administered, yet, transmissions continue to rise (WHO, 2022). Infections after vaccination also contribute to these increasing COVID-19 cases due to the emergence of SARS-CoV-2 variants that evade immunological defences. Indeed, the vaccines have been shown to mitigate the severe morbidity associated with COVID-19 and lower the mortality rate, which in turn alleviate the burden and strain on public health services. The emergence of new SARS-CoV-2 variants (mediated by structural protein mutations, especially the spike protein), however, lower the efficacy of these vaccines as indicated by increased virulence and advanced evasive mechanisms against host immunological defence employed by these coronaviruses (Aleem *et al.*, 2021; Azgari *et al.*, 2021; Chen *et al.*, 2021). There are currently few drugs (FDA-approved and promising candidates) that fight off SARS-CoV-2 infection and provide therapeutic relief from COVID-19 illness. There are no vaccines that confer long-term immunity and consistently elicit immunological protection against a broad range of SARS-CoV-2 variants. Optimal pharmacological measures to control the spread and SARS-CoV-2-related morbidity are yet to be discovered and developed.

Nevertheless, there has been an influx of biological data relating to SARS-CoV-2, as well as the availability of biological data of closely related species, such as SARS-CoV and MERS-CoV, in biological databases. These data contribute to the efforts to elucidate the pathogenicity of SARS-CoV-2, as well as its biology. Despite this data and knowledge influx, there is still a knowledge gap in terms of the proteolytic mechanism employed by the

SARS-CoV-2 M^{pro}. Recent attempts in elucidating the proteolytic activity of SARS-CoV-2 M^{pro} with the intent to develop or suggest potential antiviral agents were comparative studies based on genomic and structural conservation (Ullrich and Nitsche, 2020; Zhang *et al.*, 2020), while others utilised experimental approaches (Rut *et al.*, 2020) and further groups utilised different substrates as opposed to oligopeptide substrates conventionally used to study substrate specificity and the proteolytic mechanism of the M^{pro} (Swiderek and Moliner, 2020).

In the past, viral proteases have proven to be excellent drug targets that have led to the development of effective drugs against chronic infections, like human immunodeficiency virus (HIV) or hepatitis C virus (HCV), which employ aspartyl and serine proteases, respectively (Ullrich and Nitsche, 2020). Due to its participation in cleaving replication-essential enzymes, such as RNA-dependent RNA polymerase or nsp12, the M^{pro} is thus placed in the viral replication cycle, an essential process for SARS-CoV-2 pathogenesis (Ullrich and Nitsche, 2020). The RNA-dependent RNA polymerase cannot fully function before its proteolytic release from pp1ab (Ullrich and Nitsche, 2020). The functional importance of M^{pro} makes it an attractive target for antiviral drug discovery against SARS-CoV-2, as its inhibition could potentially hinder the viral replication cycle, and overall, stall the production of infectious SARS-CoV-2 virions (Khan *et al.*, 2020; Ullrich and Nitsche, 2020). In addition, the structural architecture of M^{pro} is highly conserved across various coronaviruses, despite the extensive mutagenesis that coronaviruses are subject to (Goyal and Goyal, 2020). Mutations in key proteins are frequently detrimental to viruses. Thus, the development of M^{pro} inhibitors will create broad-spectrum antiviral therapeutic agents against SARS-CoV-2 and other coronaviruses, while reducing the risk of mutation-mediated drug resistance in future deadly viral strains (Khan *et al.*, 2020; Goyal and Goyal, 2020). There are no human proteases with an analogous substrate specificity to the protein and therefore, M^{pro} inhibitors are more likely to be harmless to the patients (Goyal and Goyal, 2020). Furthermore, a study outlined the role of M^{pro} in counteracting the host innate immune response by acting on post-translational modifications of host proteins across various coronaviruses (Lei *et a*l., 2018), thus, M^{pro} inhibition is of vital importance in preventing replication and proliferation of SARS-CoV-2 and ultimately, in the fight against COVID-19. Understanding the functionality of SARS-CoV-2 M^{pro} will provide insight into the physiology of the coronavirus and also provide a rational approach in the development of effective antiviral chemotherapy against COVID-19.

1.7 AIM AND OBJECTIVES

The principal aim of the present study was to profile peptide substrate binding onto SARS-CoV-2 M^{pro} and explore the binding interactions in the protease-peptide complexes to gain better insight into the underlying proteolytic mechanism using Bioinformatics approaches.

The specific objectives for this study were to:

1) To generate a library of potential hexapeptide substrates and calculate protein-peptide complexes using molecular docking, in the context that the conformational search space for a hexapeptide will be large

2) To profile substrate binding, by assessing binding efficiencies of the substrate residues and evaluating the binding modes

3) To assess the protein behaviour and stability in the complex systems using molecular dynamics simulations

CHAPTER TWO

GENERATION OF THE VIRTUAL MULTI-CONFORMER HEXAPEPTIDE LIBRARY

COVID-19 has been spreading devastation across the world through the disruption of social, economic, and political stability. Having claimed more than 5 million lives since its emergence in 2019, the disease has proven to be the deadliest in recent history and has resulted in a pandemic of extraordinary proportions with a severe negative impact on public health and the livelihood of people (WHO, 2022). The disease is caused by the infection of the SARS-CoV-2 coronavirus, which manifests in deadly pneumonia-like symptoms. The virus consists of a 30 kilo-base RNA genome that encodes about 9 860 amino acids that form the composition of at least 27 proteins (Guo *et al.*, 2020; Mousavizadeh and Ghasemi, 2020). The ORF1a and ORF1b in the first two-thirds of the RNA genome encodes two polyproteins, pp1a and pp1ab, which are subsequently processed by the M^{pro} and a papain-like protease into non-structural proteins (nsps) (Chen *et al.*, 2020). The preferred recognition sequence for M^{pro} was determined to be Leu-Gln↓ (Ser/Ala/Gly) (↓ shows the cleavage site) (Goyal and Goyal, 2020). The M^{pro} is vital to the life cycle of coronaviruses, as it directly mediates the maturation of the nsps which is essential for viral replication and assembly (Khan *et al.*, 2020;). This chapter details the generation of the multi-conformer hexapeptide library based on the reported substrate specificity of SARS-CoV-2 M^{pro} according to the findings of Ullrich and Nitsche (2020). High occurrence amino acids were used to generate the peptide substrates. Each hexapeptide was constructed to contain the recognition sequence and the cleavage site. Terminal capping was performed to increase structural stability in the substrates.

2.1 INTRODUCTION

The investigation, discovery and testing of natural substrates are fundamental to the biochemical characterisation of any protein. Natural substrates reveal the function of the protein and show the overall biological importance and physiological relevance to cellular

homeostasis (Venkatraman *et al.*, 2009). In many instances, substrates and their products provide better insight into the cellular pathways the protein catalyses, as well as cascades of cellular processes that rely on the regulation of the protein (Grigalunas *et al.*, 2020; Venkatraman *et al.*, 2009). In the context of proteases, substrate specificity which highly relates to protease function is understood through the investigation of the peptide substrates and the examination of their products. Substrate specificity underpins the elucidation of the mechanism of proteolysis which instructs the assignment of the protease to site-specific proteolysis (Qi *et al.*, 2017). In turn, substrate specificity substantially aids in deciphering the biological importance of the products to the physiology of the organism (Hara *et al.*, 2017; Johnson and Chen, 2017). Since proteases are often associated with the development and progression of diseases, the knowledge of their substrates and cleavage preferences become fundamental to the rational design of therapeutic molecules that modulate protease activity (Uliana *et al.*, 2021).

Generally, the use of substrate libraries has been crucial to the characterization and profiling of substrate specificity, and this has provided information for the elucidation of protein function and catalytic mechanisms (Boulware and Daugherty, 2006). Since the pioneering of peptide synthesis, synthetic peptides have been used to create cleavage preference profiles for specific proteases; these profiles help in identifying the preferred cleavage sites and in the characterisation of their linear recognition sequence specificities. In addition, these shed light on subsite preferences and also aid in revealing the underlying molecular modes of action (Ivry *et al.*, 2018; Biniossek *et al.*, 2016; Zhou *et al.*, 2020; Vizovišek *et al*, 2018). These profiles also allow the identification of the structural origins of protease specificity and promiscuity (Biniossek *et al.*, 2016). As a result, several pharmacological successes have been possible due to this elucidation of substrate specificity and the subsequent exploitation of the promiscuity revealed by the substrate specificity profiles. Peptide substrates have always served as a strong basis for rational drug design and drug discovery, and have led to effective and efficient chemotherapies (Grigalunas *et al.*, 2020; Ullrich and Nitsche, 2020).

2.2 RDKit

RDKit is a powerful open-source software suite for cheminformatics, computational chemistry, and predictive modelling. The software toolkit was developed to support the construction of predictive models for ADMET (absorption, distribution, metabolism, elimination and toxicity) and biological activity. RDKit supports various queries of

computer-aided studies and machine learning like substructure searching, canonical Simplified Molecular Input Line Entry System (SMILES), chirality support, chemical transformations, chemical reactions, and even molecular serialization (Landrum, 2013a). The general implementation of RDKit in computational modelling involves the use of SMILES arbitrary target specification (SMARTS) and/or SMILES as inputs which generate models as output that can be stored in MDL Molfile (mol), Standard Database Format (SDF), and THOR Data Tree (TDT) files (Landrum, 2013a).

RDKit supports the construction of 2D and 3D molecular structures. It is very powerful at 2D depiction of structures and also accommodates constrained depiction and mimicry of 3D coordinates. Moreover, the toolkit accommodates the conversion of 2D molecules to 3D, together with the conformational analysis via distance geometry that accompanies the process. RDKit uses the Universal Force Field (UFF) implementation to clean up structures and optimise geometries and conformations (Landrum, 2013a; Landrum, 2013b). Other implementations within RDKit include Fingerprinting (Daylight-like, circular, atom pairs, topological torsions, "MACCS keys", etc.); similarity/diversity picking (include fuzzy similarity); 2-D pharmacophores; Gasteiger-Marsili charges; hierarchical subgraph/fragment analysis and Hierarchical RECAP implementation (Landrum, 2013a).

2.3 SMILES

SMILES is a linear notation language for entering and representing chemical structures and reactions (Daylight, 2019a; Gasteiger *et al.*, 2018). The SMILES language is a typographical method that represents molecular structure by a linear string of symbols (i.e. printable characters), like natural language (Weininger, 1988; Daylight, 2019a). SMILES notation is an efficient alternative to conventional conversion tables as it requires less storage space (50% to 70% less space) and it is a linguistic construct that can be integrated into other languages designed for the storage of chemical information and chemical intelligence (Daylight, 2019a). SMILES denotes a molecular structure as a graph with optional chiral indications detailing the description of a molecule in the manner in which they are drawn by chemists (Weininger, 1988; Daylight, 2019a). SMILES provides a platform for the accurate and unique specification of molecules which can be used with chemical databases due to storage efficiency (Weininger, 1988; Daylight, 2019a).

2.4 SMARTS

SMARTS is a substructure identification language that allows one to specify substructures using rules that are compatible with SMILES (Gasteiger *et al.*, 2018; Daylight, 2019b). SMARTS specifies substructures using the same linear strings used to specify chemical structures in SMILES. However, the specification in SMARTS extends to include logical operators and special atomic and bond symbols which allows for the SMARTS atoms and bonds to be more general (Daylight, 2019b). The difference in semantics between SMILES and SMARTS expressions allude to the interpretation of the specific expression. The SMILES string is interpreted as a molecule and the resultant molecule is what is subject to substructure searching. Alternatively, the SMARTS string is interpreted as a pattern (which denotes a substructure) and matched against a molecule (Daylight, 2019b). Both SMILES and SMARTS provide a fast and efficient way to store and query chemical structures (Gasteiger *et al.*, 2018).

2.5 METHODOLOGY

As there were no available oligopeptide libraries with peptide chains demonstrating the recognition sequence required by SARS-CoV-2 M^{pro} for proteolytic cleavage, a peptide library was generated using RDKit (v. 2019.09.1), powered by Python on a server (Landrum, 2013). With the findings of Ullrich and Nitsche (2020), hexapeptides were generated using the MolFromSequence method. The selection of the constituent amino acids was based on their high occurrence frequency in the recognition sequence and cleavage site.

Acetyl and methylamine constructs were generated to constitute terminal caps using MolFromSMILES functionality. SMARTS patterns were utilised to identify the alpha carbon atoms (α-carbons) of terminal amino acids. The C- and N- terminal α-carbons were replaced with acetyl (ACE) and methylamine (NME) constructs to perform terminal capping (Penkler *et al.*, 2017). Subsequently, hydrogen atoms were removed and added back to ensure that the valency of each atom was satisfied. The 3D coordinates of the atoms were generated using the EmbedMolecule method. The molecules were optimised using the implemented UFFOptimizeMolecule. A structural conformational search was performed to generate conformers. The resulting conformers were stored in an SDF file.

2.6 RESULTS AND DISCUSSION

The generation of the peptide library was based on the findings of Ullrich and Nitsche (2020).

Their study sought to elucidate the substrate specificity of the SARS-CoV-2 M^{pro}. Their approach was to analyse the polyprotein 1a/b sequence to identify the M^{pro} recognition sequence or cleavage sites, provided that the P2 to P1′ residues of these sites display the highest degree of conservation in closely related SARS-CoV and MERS-CoV viruses (Ullrich and Nitsche, 2020). Their approach was motivated by the evident high conservation of the RNA genomes across SARS-CoV-2, SARS-CoV and MERS-CoV, together with the high degree of structural similarity and conservation of the active site which they observed in superimposed main protease structures belonging to SARS-CoV-2, SARS-CoV and MERS-CoV (Goyal and Goyal, 2020; Liu *et al.*, 2020; Zhang *et al.,* 2020; Ullrich and Nitsche, 2020). In addition, these conserved P2 to P1′ residues are crucial in determining substrate specificity and they follow a similar pattern across the coronavirus species (figure 2.1). The P2 position tolerates small hydrophobic amino acids with a clear preference for leucine (figure 2.1; Ullrich and Nitsche, 2020). The P1 position is always occupied by the highly conserved Glutamine (GLN) which is present in all polyprotein cleavage sites of SARS-CoV-2, SARS-CoV and MERS-CoV (figure 2.1; Ullrich and Nitsche, 2020). The P1′ position tolerates small amino acids such as serine or alanine (figure 2.1; Ullrich and Nitsche, 2020).

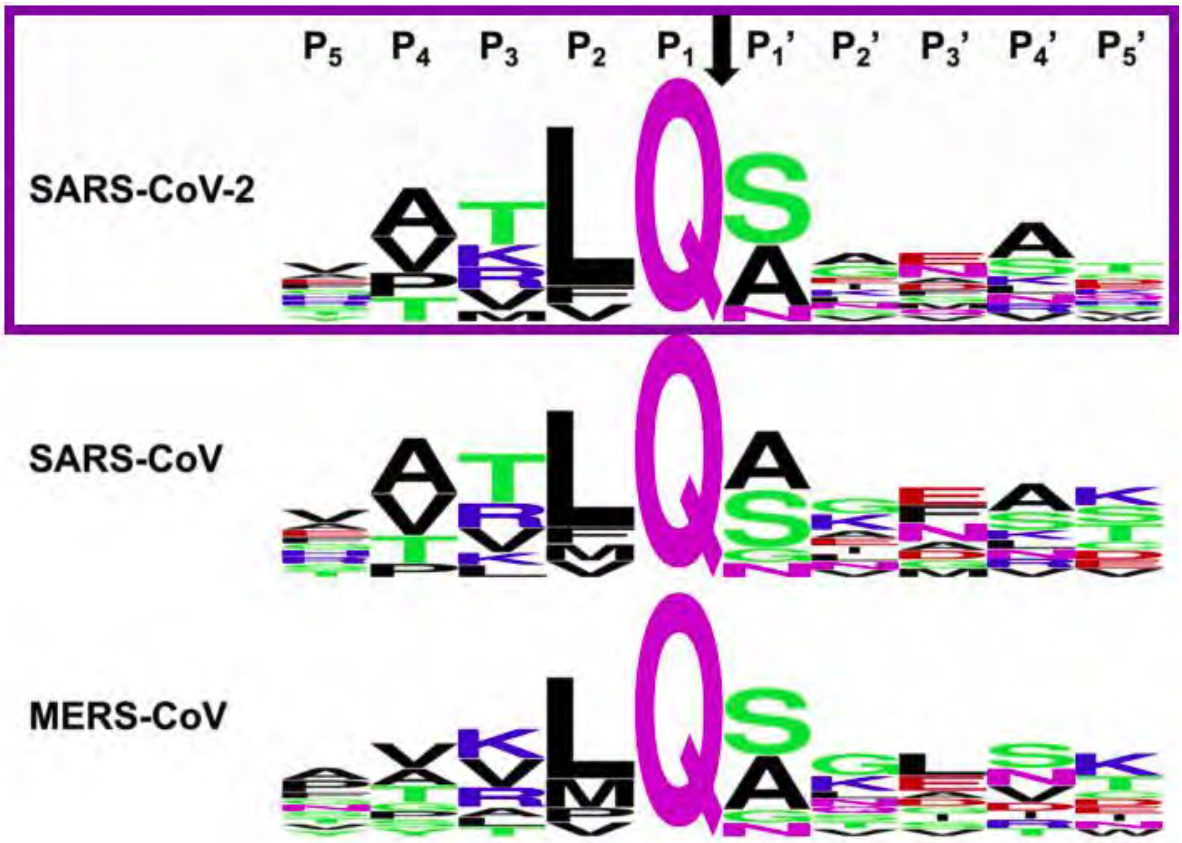


Figure 2.1. Polyprotein cleavage sites recognised by M^{pro} of SARS-CoV-2, SARS-CoV and MERS-CoV. Peptide sequences cover residues P5 to P5′ according to the nomenclature of Schechter and Berger (1967). 81 Data were generated from pp1ab polyprotein sequences reported in the UniProt database with the accession codes P0DTD1 (SARS-CoV-2), P0C6X7 (SARS-CoV) and K9N7C7 (MERS-CoV). The consensus sequence covering all cleavage sites was plotted using WebLogo. Adapted with permission from Ullrich and Nitsche (2020).

The generation of the peptide library was carried out using RDKit and Python scripting (Landrum, 2013). The initial plan was to generate a peptide library of octapeptide substrates utilising all the amino acids that occur in P4 to P4′. The resultant library consisted of 102 060 octapeptides. Due to time and computational (limited access to the large queue at the supercomputer) constraints, hexapeptides (P3 to P3′) were prioritised and most importantly, the focus was placed on amino acids that displayed high frequencies of occurrence in their respective position of the cleavage site. The selection of the constituent amino acids was based on their high occurrence frequency in the recognition sequence and cleavage site. In essence, each hexapeptide (P3-P3′) consisted of the both the recognition sequences (P2-P1′) and the cleavage site (P1-P1′), equally divided in the C- and N- terminal products of the substrates (***). Despite this reduction of amino acids in the peptide substrates, variability in chemical properties was maintained in P3, P2′ and P3′ (table 2.1; Crooks *et al.*, 2004).

Table 2.1: The amino acid residues in the polyprotein cleavage sites recognised by SARS-CoV-2 M^{pro} used in the construction of hexapeptides.

Residue position	Amino acids
P3	Thr; Arg; Lys; Val; Met
P2	Leu
P1	Gln
P1′	Ser; Ala
P2′	Ala; Gly; Glu; Thr; Leu; Asn; Ser; Val
P3′	Glu; Asn; Ala; Asp; Phe; Gly; Met; Gln; Val

Green: Polar; Blue: Basic; Red: Acidic; Black: Hydrophobic; Purple: Neutral

Python scripting through nested loops was employed to generate the different hexapeptides; each consisting of unique amino acid combinations of the cleavage site. A total of 810 capped hexapeptides were generated and each hexapeptide consisted of 100 different structural conformations. The purpose of the structural conformational search was to create a multi-conformer library. In ligand-based drug design, multi-conformer libraries are essential for predicting the bioactive conformations of ligands in the absence of the structural model of the receptor, especially for ligands with rotatable bonds (Yongye *et al.*, 2010). In this study, the variation in conformation was intended to supplement the conformation generation within molecular docking procedures.

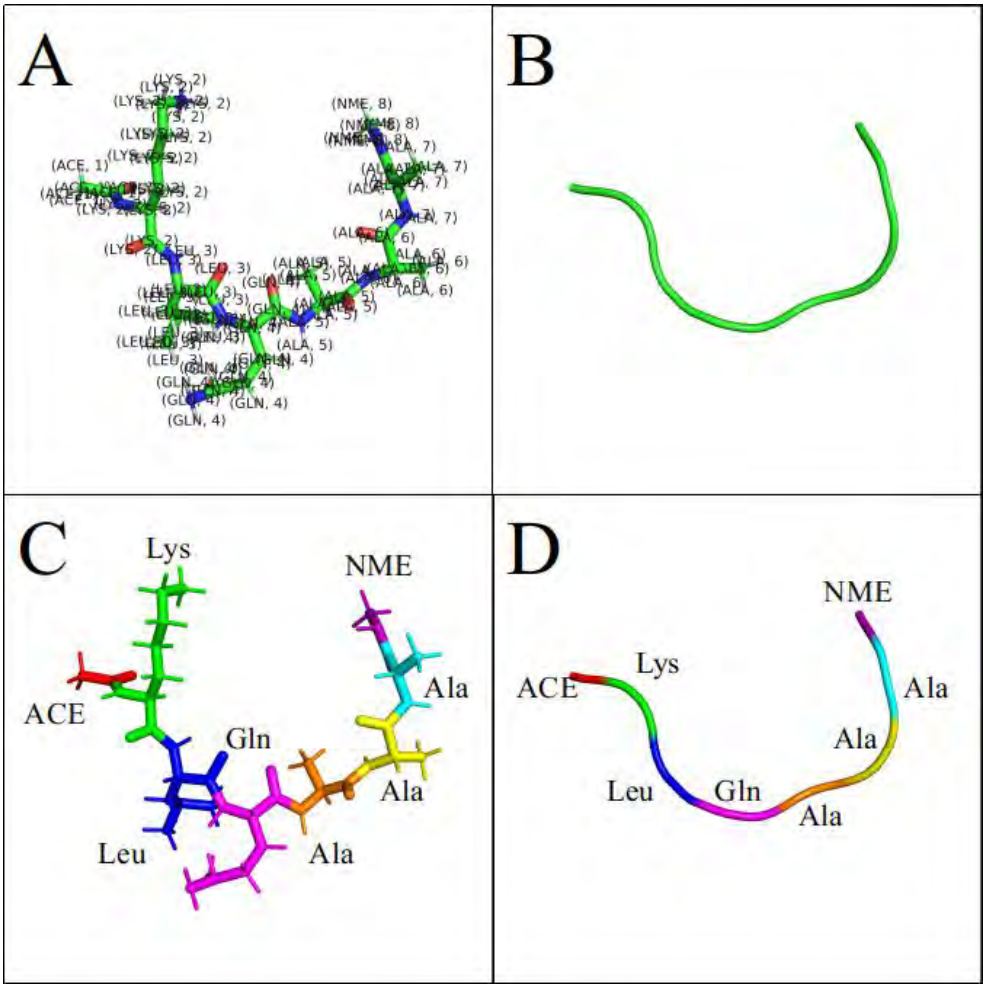


Figure 2.2. The terminal capping of the hexapeptides. The 3D structure of the Lys-Leu-Gln-Ala-Ala-Ala (KLQAAA) substrate capped with acetyl (ACE) and methylamine (NME) in the C- and N-termini, respectively. **A**) shows the stick representation of the capped hexapeptide presented in its atomic composition, where green represents carbons; red represents oxygens; blue represents nitrogens and grey represents hydrogens. **B**) shows the cartoon representation of the substrate backbone. **C)** shows the typical, colour-coded amino acid composition of a hexapeptide showing the residues and caps in different colours. **D)** shows a cartoon representation of a typical, colour-coded amino acid composition of a hexapeptide showing the residues and caps in different colours. The image was generated using PyMOL.

Terminal capping using acetyl and methylamine groups can increase the peptide stability towards its protease, which in turn can improve their affinity for specific biological targets (figure 2.2; Fang *et al.*, 2011). According to Penkler *et al.* (2017), the use of ACE and NME for capping peptide chains allow the structures to simulate being part of a protein. To perform the terminal capping shown in Figure 2.3, the SMARTS strings '[$(OC(=O)CN)]' and '[$(NCC(=O))]' were applied to each hexapeptide to search and identify the C- and N-terminal α-carbons. Once identified, these α-carbons were replaced with the α-carbons consisting of the ACE and NME constructs for the C- and N-termini, respectively (figure 2.3B & D). This was all automated using python scripts which are listed in Appendix A.

Figure 2.2 shows the successful implementation of terminal capping. Although the structures of terminal residues were modified during capping, amino acid compositions were restored

and conventional peptide representations (such as the 3D structure) were conserved, as indicated by the cartoon representation in figures 2.2B and 2.2D. In this context, terminal capping was carried out to safeguard the stability of the hexapeptides and prevent helical ends from fraying during the dynamic simulation studies (Zuo *et al.*, 2012; Zuo *et al.*, 2014). Furthermore, capped termini prevent the misfolding of the protein and peptide chains - serving as a protective group, and this avoids the disruption of the molecular structure in receptor targets (Andreasen *et al.*, 2014; Hernik-Magoń *et al.*, 2017; Lee et al., 2007).

2.5 CHAPTER SUMMARY

In this chapter, the objective to generate suitable peptide substrates for SARS-CoV-2 M^{pro} was fulfilled. A multi-conformer hexapeptide library was generated based on the SARS-CoV-2 M^{pro} substrate specificity and recognition informed by the work of Ullrich and Nitsche (2020). The hexapeptides consisted of the recognition sequence and cleavage site for SARS-CoV-2 M^{pro}. Only amino acids displaying high occurrence frequencies in their respective positions were prioritised. A total of 810 unique hexapeptides were generated, each with 100 conformers using Python and SMILES in RDKit (a total of 81 000 conformers). The constituent residues were equally divided between the C and N-terminal products (P3-P3′). The C- and N- terminal α-carbons were replaced with ACE and NME constructs to constitute terminal capping using SMARTS. Terminal capping was performed to increase the stability of the substrates in dynamic environments. The hexapeptide conformers were stored in SDF files and used in molecular docking studies detailed in the next chapter.

CHAPTER THREE

MOLECULAR DOCKING OF MULTI-CONFORMER HEXAPEPTIDE LIBRARY

The SARS-CoV-2 M^{pro} is essential to the life cycle of the coronavirus, as it directly mediates the maturation of the non-structural proteins (nsps) which are essential for viral replication and assembly (Khan *et al.*, 2020). The M^{pro} plays a key role in the processing of the polyprotein to form mature nsps. The M^{pro} is first autocleaved from pp1ab (nsp5) to produce mature protease, and then proceeds to cleave downstream nsps at 11 different sites of the pp1ab to release nsp4-nsp16 (Khan *et al.*, 2020). The recognition sequence at most sites was determined to be Leu-Gln↓(Ser/Ala/Gly) (↓ shows the cleavage site) (Goyal and Goyal, 2020). This chapter details the profiling of substrate binding of the hexapeptides (from the multi-conformer peptide library) onto M^{pro} via molecular docking. A suitable M^{pro} crystal structure was selected and prepared for molecular docking alongside the conformers of the substrates. Substrate binding was characterized through docking results and the calculation of ligand efficiencies. Substrate recognition and specificity were profiled via subsite mapping of the protease-peptide interface, and through the assessment of ligand-receptor molecular interactions at the active site of complexed structures.

3.1 INTRODUCTION

The biological significance of protein function and regulation is realized when the protein makes direct physical interaction with other molecules (Du *et al.*, 2016). These direct interactions occur as a result of molecular recognition, where the protein interacts with binding partners through non-covalent interactions to form specific complexes (Du *et al.*, 2016). Molecular recognition is mainly characterized by specificity and affinity. Protein-ligand binding has been a topic of study for many years and is certainly an imperative step for drug discovery. To gain better insight into protein function, thorough elucidation of

the mechanisms governing protein-ligand interactions is required, together with the full description, characterization and quantification of the energetics that facilitates the formation of complexes. Consequently, this in-depth understanding of the physicochemical mechanisms of protein-ligand binding, and analysis of structural data optimizes rational drug design and facilitates the discovery, design and development of new drugs based on the detailed information about molecular recognition and interaction (Du *et al.*, 2016).

To date, three main models explain the mechanisms of protein-ligand binding. These include the "lock-and-key", "induced fit" and "conformational selection" models. Fisher proposed the "lock-and-key" model in 1894, which depicts the ligand and protein/enzyme as rigid structures wherein their binding interfaces perfectly complement one another. While the model explains the substrate specificity that proteins exhibit, and also emphasizes the importance of shape complementarity between the two structures, it does not account for the structural differences the protein and ligands exhibit in unbound and bound states (Tobi and Bahar, 2005; Du *et al.*, 2016). This has led to the “induced fit” model proposed by Koshland in 1958, to account for this type of plasticity of proteins (Tobi and Bahar, 2005). The model explains that an interacting ligand induces a conformational change in the flexible binding site of the protein and thus, mediates protein-ligand binding and interaction (Du *et al.*, 2016). This model takes into account the flexibility of the ligand-binding site, and also explains the substrate recognition that proteins exhibit (Tobi and Bahar, 2005; Du *et al.*, 2016). However, the "induced fit" model is seemingly suitable for proteins that show minor conformational changes after the ligand binding. Furthermore, both models depict a protein as a single, stable conformation under given conditions (Du *et al.*, 2016). The "conformational selection" model later emerged to take into account this inherent flexibility of proteins. The "conformational selection" model postulates the native state of a protein exists as a vast ensemble of closely related conformational states or substates. These substates coexist in equilibrium, and the ligand only binds selectively to the most suitable substate, thus shifting the equilibrium towards this substate and consequently towards the formation of a complex. This suggests that the unbound protein can sample with a certain probability the same conformation as that of the ligand-bound state (Tobi and Bahar, 2005; Du *et al.*, 2016)

3.2 MOLECULAR DOCKING

3.2.1 COMPUTER-AIDED DRUG DISCOVERY/DESIGN

Advancements in computer hardware, software, and algorithms have led to the optimization

of the drug discovery process, where computation rather than experimentation reduces time and costs (Lin *et al.*, 2020). Over the past three decades, the implementation of computer-aided discovery/design (CADD) methods in drug discovery has helped accelerate the process and contributed to many breakthroughs in the development and rapid availability of novel therapeutic agents. CADD methods largely contribute to the optimization of virtual screening techniques, which then allow rapid hit identification, lead optimization and rational drug design. Moreover, CADD methods are classified as either ligand-based or structure-based methods (Sliwoski *et al.*, 2014). Ligand-based methods heavily rely on the knowledge of elucidated ligands of the target protein. In principle, the structure-based methods are analogous to high-throughput screening wherein accurate information about target and ligand structures is imperative. Structure-based approaches must be performed with available structural models of the target proteins, which are obtained either by X-ray diffraction, nuclear magnetic resonance (NMR) or molecular simulation (homology modelling) (Sliwoski *et al.*, 2014; Lin *et al.*, 2020). Virtual screening techniques aim to identify novel active small molecules from a large compound library that bind favourably to the target protein. Readily used tools for virtual screening techniques include molecular docking, pharmacophore modelling and Quantitative Structure-Activity Relationship (QSAR) (Sliwoski *et al.*, 2014; Lin *et al.*, 2020)

3.2.2 VIRTUAL SCREENING - MOLECULAR DOCKING

Molecular docking is a powerful screening technique that predicts the interaction patterns between proteins and their ligands, by modelling virtually a complex structure of the binding partners (Pantsar and Poso, 2018; Lin *et al.*, 2020). The theoretical basis for molecular docking is the “induced fit” model, in which ligand and receptor recognition depends on spatial shape complementarity and energy matching (Lin *et al.*, 2020). In structure-based docking, a small ligand molecule is aligned inside the binding cavity of the target protein with the intent to find the most favourable conformation or pose for complex formation. The docking process typically involves two independent stages which are: conformation generation; and the scoring of the resulting conformations (Pantsar and Poso, 2018).

3.2.3 CONFORMATIONAL SAMPLING AND DOCKING SIMULATION

The available conformations to both receptor and ligand present the sampling engine with a huge challenge of finding all optimal receptor-ligand conformations during docking (Klebe, 2006; Guedes *et al.*, 2014). Both structures are often flexible and dynamic in nature and

possess numerous translational and rotational positions in a 3D space. Therefore, an exhaustive conformational search on both the receptor and the ligand proves to be difficult, especially for large protein structures. Widely used conformational sampling algorithms overcome the computational cost of this process by limiting some flexibility in the structures, most often this is for the receptor (Leach *et al.*, 2006; Meng *et al.*, 2011; Pantsar and Poso, 2018; Salmaso and Moro, 2018). Rigid docking algorithms, much like the "lock-and-key" model, consider the receptor and ligands as rigid structures and only consider three translational and three rotational degrees of freedom during sampling. Semi-flexible algorithms treat ligands as flexible structures with rigid receptors and sample the conformational freedom of the ligands alongside the six translational and rotational ones. Flexible docking algorithms consider the ligand and receptor as flexible binding counterparts. Consequently, these algorithms present a great number of degrees of freedom to search, and as such computational resources are often augmented to optimise sampling and scoring to attain a balance between accuracy and speed (Salmaso and Moro, 2018).

In the context of semi-flexible docking, the selection of the docking algorithms, together with the setting of the search strategy and an appropriate level of conformational sampling, are crucial steps to a successful simulation of docking. Scoring functions assess performance which is affected by the conformational search and ligand placement. Search strategies may be systematic or random. Systematic searches incorporate a comprehensive sampling of the conformations and structural properties and thus, use significantly more time and resources to generate the poses and evaluate them individually (Prieto-Martínez *et al.*, 2018). As such, a systematic search is performed by constructing the ligand from different fragments, wherein one fragment serves as an anchor whilst the remaining fragments are sequentially added to avoid the generation of a combinatorial explosion (Prieto-Martínez *et al.*, 2018). Alternatively, a stochastic search is performed randomly using the Monte Carlo (MC), Tabu search, Swarm optimization (SO) or genetic algorithm (GA) methods. Each method develops different conformers based on bond rotations as degrees of freedom and these conformers are then evaluated by a scoring function for pose selection and filtering (Prieto-Martínez *et al.*, 2018; Salmaso and Moro, 2018). A stochastic search searches a broader range of conformations in a given timeframe, and this may be advantageous in terms of rapidly finding feasible solutions. However, the technique does not ensure the full search of the conformational space, meaning the true solution may be missed. Increasing the number of iterations of the algorithm thus mitigates this lack of convergence (Salmaso and Moro, 2018).

3.2.4 ENERGY SCORING FUNCTIONS

Docking simulations and conformational sampling may produce a great number of solutions. Energy scoring functions then evaluate the generated conformers of the ligands, separating the biologically relevant poses from the incorrect and inactive poses (Meng *et al*, 2011; Prieto-Martínez *et al.*, 2018). Scoring functions are fast, approximate mathematical methods used to assess the binding affinity between the binding partners after docking. Notably, scoring functions can be used within the search algorithm to accelerate the process of pose prediction (Du *et al.*, 2016). The favourable solutions are distinguished through the evaluation of a broad range of properties such as intermolecular interactions, desolvation, electrostatic, and entropic effects (Prieto-Martínez *et al.*, 2018). The scoring functions then estimate the binding affinity between the ligand and receptor by adopting various assumptions and simplifications (Meng *et al*, 2011). Scoring functions can be categorised as force-field-based, empirically-based or knowledge-based.

Force-field-based scoring functions estimate the energy of a system with regards to bonded (intramolecular) and non-bonded (intermolecular) components (Meng *et al.*, 2011; Salmaso and Moro, 2018). Binding affinity is assessed by calculating the sum of non-bonded interactions using a function that also accounts for bonded interactions (Meng *et al.*, 2011; Prieto-Martínez *et al.*, 2018). Intermolecular interactions include van der Waals and the electrostatic potential, which are described by the Lennard-Jones potential and the Coulomb function, respectively. Consequently, this means that the entropic contribution of solvation is not accounted for. Thus, a distance-dependent dielectric may be introduced to mimic the solvent effect (Salmaso and Moro, 2018). Moreover, force-field-based scoring functions have a slow computational speed and require cut-off distances to be introduced to handle intermolecular interactions. This reduces the accuracy of long-range effects involved in binding (Meng *et al.*, 2011).

Empirical scoring functions estimate the binding energy as a sum of several energy components such as hydrogen bonding, ionic interactions, hydrophobic effects and binding entropy. These empirical energy components are weighted by coefficients optimised from regression analysis fitted to a test set of ligand-protein complexes with known binding affinities (Du *et al.*, 2016; Salmaso and Moro, 2018; Meng *et al.*, 2011).

Knowledge-based scoring functions assume that more favourable interactions towards binding affinity have greater frequencies of occurrences between the binding partners. Thus, the functions use statistical analysis of ligand-protein complexes from a database of crystal

structures to obtain the interatomic contact frequencies and/or distances between the ligand and protein. The frequencies are computed and converted into an energy component. The score for a pose is calculated by summing up the tabulated energy components for all ligand-protein atom pairs (Salmaso and Moro, 2018; Meng *et al.*, 2011). A significant advantage of knowledge-based functions is computational simplicity that establishes a balance between performance and accuracy. Moreover, these functions also consider uncommon interactions like sulfur-aromatic or cation-π interactions (Meng *et al.*, 2011; Prieto-Martínez *et al.*, 2018).

3.3 AUTODOCK VINA

AutoDock Vina is a powerful open-source computational program for molecular docking and virtual screening (Trott and Olson, 2010). It borrows ideas and approaches from AutoDock 4 but is conceptually designed differently. The program is up to two orders of magnitude faster than AutoDock4 and features significant improvements such as an efficient optimization algorithm and a scoring-function-based search algorithm for estimating binding affinity and predicting reasonable poses, respectively (Jaghoori *et al.*, 2016; Vieira and Sousa, 2019).

AutoDock Vina uses the MC/BFGS search algorithm which comprises a Monte-Carlo (MC) iterated search partnered with the Broyden-Fletcher-Goldfarb-Shanno (BFGS) gradient-based optimizer. The MC search serves as a tool for stochastic global optimization, while the BFGS method is used for deterministic local optimization. The BFGS method is an efficient quasi-Newton method that takes the value and the derivatives (gradient) of the scoring function with respect to all the design variables, such as the position and orientation of the ligand, together with the values of the torsions for the active rotatable bonds in the ligand and flexible residues. After several iterations, the BFGS will converge to a point where the gradients vanish in all directions, within a negligibly small tolerance. This point of convergence marks the local minimum and is where the local search optimization is terminated. This powerful MC/BFGS algorithm infers search efficiency that leads to better docking results with fewer scoring function evaluations (Trott and Olson, 2010; Eberhardt *et al.*, 2021; Handoko *et al.*, 2012).

AutoDock Vina implements a hybrid scoring function that combines the empirical and knowledge-based functions. The scoring function is based on the pairwise interactions between atoms. These interactions are defined by five terms based on the surface distance between the atoms (Tanchuk *et al.*, 2016). The five terms that define interaction include a van der Waals-like potential (defined by a combination of a repulsion term and two attractive

Gaussians), a non-directional hydrogen-bond term, a hydrophobic term, and a conformational entropy penalty. Interestingly, AutoDock Vina lacks electrostatics and solvation which AutoDock 4 has. The binding energy is predicted as the sum of distance-dependent atom pair interactions (Quiroga and Villarreal, 2016; Eberhardt *et al.*, 2021).

The appeal of AutoDock Vina over other docking programs is its multi-core capability, high performance and enhanced accuracy, ease of use and free availability. The multi-core capability and high performance contribute to the characteristic fast speed of AutoDock Vina and make it a proficient choice for virtual screening (Vieira and Sousa, 2019). Moreover, calculations can be performed simultaneously in parallel over multiple Central Processing Unit (CPU) cores using multithreading. Additional features like exhaustiveness also contribute to the enhanced accuracy of AutoDock Vina in terms of the prediction of plausible poses. During a docking simulation, the program may repeat the conformational sampling several times with different randomizations. The exhaustiveness controls the number of times the conformational sampling can be repeated within the same randomization seed. Essentially, higher exhaustiveness exponentially raises the probability of finding a correct solution as more runs are performed, with the only drawback of increased computation time (Jaghoori *et al.*, 2016).

3.4 LIGAND EFFICIENCY

Ligand efficiency (LE) is defined as the binding energy per heavy atom (Hopkins *et al.*, 2004). LE was first proposed as a useful metric for the selection and optimization of favourable fragments, hits and leads with optimal physicochemical and pharmacological properties in drug discovery (Hopkins *et al.*, 2004; Abad-Zapatero and Metz, 2005; Orita *et al.*, 2011). Orita *et al.* (2011) reported that LE is useful in assessing the quality of hit compounds, as the metric represents a balance between potency and molecular weight, which relate to ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity) parameters. In simple terms, LE is the measurement of the goodness of interaction between a compound and its target protein (Orita *et al.*, 2011). Additionally, fragments, hits and leads with favourable LE values indicate a greater potential improvement in binding affinity in the process of structure-guided drug design (Chen *et al.*, 2017).

The application and scope of LE have since evolved in drug discovery to incorporate other properties of ligands that closely relate to ADMET parameters, such as lipophilicity, molecular mass, polar surface area, combinations of physicochemical properties, and

functional group contributions (Hopkins *et al.*, 2014). As a result, modified LE metrics have been proposed to undertake the extension of LE. These modified metrics include percentage efficiency index (PEI) and binding efficiency index (BEI) (which use molecular weight); ligand-lipophilicity efficiency (LLE) and logP/ligand efficiency (LELP) (which incorporate lipophilicity); and the surface-binding efficiency index (SEI) (which is based on PSA) (Orita *et al.*, 2011). With these extensions, LE and its modified metrics have provided practical means to estimate target druggability and to control the balance of molecular size and potency. This significantly improves the success rates of lead optimization and drug design (Hopkins *et al.*, 2014; Chen *et al.*, 2017).

3.5 METHODOLOGY

3.5.1 RECEPTOR PREPARATION

The crystallographic 3D structure of the SARS-CoV-2 M^{pro} was retrieved from the RCSB Protein Data Bank (PDB) (PDB id: 6XHM) with a resolution of 1.41 Å. The crystallographic waters were removed. The receptor consisted of three rotamers which were separated using a python script (*prepare_pdb_split_alt_confs.py*) provided by AutoDock tools (Morris *et al.*, 2009). The first rotamer was used (conf_A), and this was initially prepared by adding polar hydrogen atoms and merging all other hydrogen atoms. Thereafter, the Gasteiger charges were calculated, and atom types were assigned. In addition, non-standard residues were deleted from the receptor. The receptor preparation was performed using a python script (*prepare_receptor4.py*) provided by AutoDock tools and the rigid molecule output was saved in an AutoDock Protein Data Bank, Partial Charge (Q), & Atom Type (T) (PDBQT) file format.

3.5.2 LIGAND PREPARATION

The SDF files of the hexapeptides were converted into XMOL molecule model (XYZ) files using OpenBabel (O'Boyle *et al.*, 2011). The conversion resulted in the separation of the conformers into individual XYZ files presenting one conformation of the respective hexapeptide. The conformers were subjected to geometry optimization employing a Semi-empirical Quantum Mechanical Method (XTB Semi-empirical method) using xtb software (Werner Reckien, 2017). The optimised structures were converted into PDB files using OpenBabel and subsequently prepared by adding polar hydrogen atoms and merging all hydrogen atoms. The Gasteiger charges were calculated, and atom types were assigned using

an AutoDock tools python script (*prepare_ligand4.py*). The output was saved in PDBQT files. The ligand preparation process was automated through the use of a python script (that called these tools where appropriate) executed in a Linux-based high-performance cluster (Appendices B-D).

3.5.3 MOLECULAR DOCKING

Molecular docking was undertaken using AutoDock Vina (Trott and Olson, 2010). As there were no co-crystallized ligands in the binding pocket, the grid parameters were determined using AutoDock Tools. The AutoGrid setup in the Autodock Tools graphic user interface was used to determine the grid parameters for each chain of the receptor with respect to the positions of the catalytic dyad and other key substrate-binding residues (Goyal and Goyal, 2020). For chain A, the grid centre was set at 12.059, 8.933 and 29.021 in the x, y and z directions, whereas the grid centre for chain B was set at -18.444, -16.361 and 7.944 in the x, y and z directions, respectively. The grid box size was set at 20, 20 and 20 Å in the x, y and z directions. A Vina configuration file was created for each ligand and the receptor protein, wherein the energy range and exhaustiveness were set at 4 and 480, respectively. The cubic box size values and the coordinates of the central atom of the grid centre were specified in the configuration file (Appendix F). Each docking process was performed using 24 CPU cores. The docking simulations were performed in parallel on a high-performance cluster to compensate for the high computational costs and to speed up computations (Appendices G-I). The generation of the Vina configuration files was automated through the use of a python script (Appendix E).

3.5.4 MOLECULAR DOCKING ANALYSIS

Docking analysis was automated with the use of customised python scripts (Appendices J & K). The best conformational poses from output docking files were extracted based on low binding energies. The conformational poses were extracted alongside corresponding free energy of binding from log files. In each substrate, the pose with the lowest free energy of binding was used for the construction of peptide-enzyme complexes. Prior to the construction of peptide-enzyme complexes, the structures of the best poses (pdbqt) were converted into PDB format using a customised Perl script (Appendix M). Initially, the amino acid information was lost and all hexapeptide constituents were then labelled "LIG" in the process of terminal capping. The Perl script was thus used to restore the amino acid information.

Docking validation was assessed through the reproducibility of high-affinity binding poses across redocking of different conformers of the same substrate. High-affinity binding poses were superimposed to visually assess the reproducibility of the docking results. Molecular interactions at active sites were resolved using Biovia Discovery Studio 2020 Client. Adherence of docked substrates to binding subsites was visualised using PyMOL (DeLano 2002).

3.5.5 LIGAND EFFICIENCIES

The free energies of binding of best binding poses were exported onto a spreadsheet and sorted according to amino acid composition. The generic ligand efficiencies were determined using the equation:

$$\text{LE} = \frac{\mu(\Delta\text{G})}{N}$$

Equation 3.1: Ligand efficiency (Hopkins *et al.*, 2004).

where ΔG is the free energy of binding and N is the number of heavy atoms (non-hydrogen atoms) (Hopkins *et al.*, 2004). Additionally, other metrics of ligand efficiencies were determined namely binding efficiency index (BEI) and surface-binding efficiency index (SEI). BEI was calculated using the equation:

$$\text{BEI} = \frac{-\mu(\Delta\text{G})}{MW}$$

Equation 3.2: Binding Efficiency Index. Modified from Abad-Zapatero and Metz (2005).

where μ(ΔG) is the mean free energy of binding (the vina score was used for this value) and MW is molecular weight in kDa. SEI was calculated using the equation:

$$\text{SEI} = \frac{-\mu(\Delta\text{G})}{(PSA/100\ \text{Å})}$$

Equation 3.3: Surface-binding Efficiency Index. Modified from Abad-Zapatero and Metz (2005).

where ΔG is the free energy of binding and PSA is the polar surface area (Abad-Zapatero and Metz, 2005). PSA values were calculated using OpenBabel (O'Boyle *et al.*, 2011). All ligand efficiency calculations were performed on WPS Spreadsheets 2019.

3.6 RESULTS AND DISCUSSION

3.6.1 RECEPTOR RETRIEVAL

The purpose of the molecular docking studies was to construct quality 3D peptide-enzyme structures that best represent the SARS-CoV-2 M^{pro} complexed with its natural substrates. The identification and selection of a quality protein receptor were fundamental to the success of the docking studies. The crystallographic 3D structure of the SARS-CoV-2 M^{pro} was retrieved from the RCSB Protein Data Bank (PDB) under the PDB ID of 6XHM (Rose *et al.*, 2012). This crystal structure was resolved using X-ray Diffraction, with crystallization following an *Escherichia coli* expression system. The crystal structure had a resolution of 1.41 Å, and R-Value Free and R-Value Work values of 0.210 and 0.191, respectively. The observed R-Value was 0.192. There were no mutations or missing residues.

3.6.2 PRELIMINARY DOCKING STUDIES

The homodimeric M^{pro} demonstrates proteolytic activity in both monomers. This makes both monomers attractive targets in studies of protease characterisation and antiviral inhibition. The dimerization of M^{pro} is critical to the biological function of the protein, since the individual monomers do not exhibit enzymatic activity (Goyal and Goyal, 2020). Preliminary docking studies were carried out to identify the monomer with better substrate binding. The conformers of the randomly selected substrate, Arg-Leu-Gln-Ala-Ala-Asn (RLQAAN), were docked on both chains (figure 3.1), using the grid specifications mentioned in section 3.5.3.

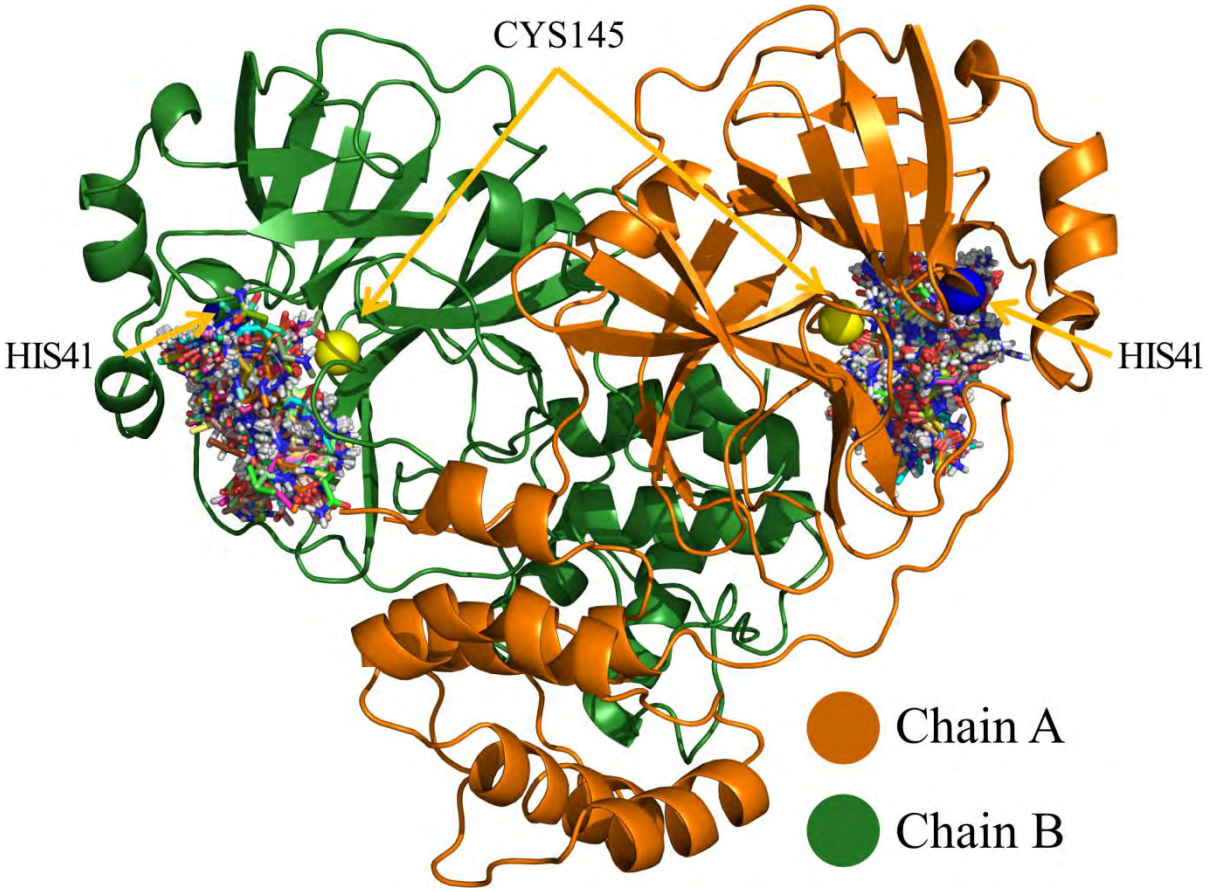


Figure 3.1. The identification of better binding monomer of SARS-CoV-2 M^{pro}. The SARS-CoV-2 M^{pro} (PDB ID:6XHM) was docked with conformers of randomly selected RLQAAN on both chains. Green chain represents chain A. Orange chain represents chain B. Catalytic residues are represented as spheres. Yellow spheres represent Cys145. Blue spheres represent His41. RLQAAN conformers are represented as sticks. The image was generated using PyMOL (DeLano 2002).

The docking results showed that chain B was producing much better results in terms of the lowest free energies of binding (supplementary figure 3.1, supplementary table 3.1). Chain B attained binding energies as low as -8.7 kcal.mol^{-1}. whereas chain A attained the lowest binding energy of -8.5 kcal.mol^{-1}. Moreover, more best poses were attaining binding affinities lower than -8.0 kcal.mol^{-1} in chain B than in chain A (supplementary figure 3.1). Hence, chain B was prioritised and used for the docking studies and subsequent calculations.

3.6.3 FREE ENERGIES OF BINDING

The assessment of the docking results is crucial in the evaluation of the overall performance of ligands at the active site. Docking analysis provides insight on the affinity of binding of a ligand, which in turn informs of its stability at the site of binding. The docking studies were carried out on a high-performance cluster given the computational cost of docking each substrate. Not all conformers were docked due to the extensive computational time required

for docking such systems with many degrees of conformational freedom. However, for each substrate at least ten conformations were redocked with exhaustiveness 480. Nevertheless, the docking results were arguably good; displaying high-affinity binding of the substrates onto the active site of the SARS-CoV-2 M^{pro}. The best binding poses registered affinities ranging between -8.7 and -7.0 kcal.mol^{-1} across all substrates (supplementary table 3.1).

Substrates RLQATF, RLQSGA and RLQSTF were the only binding poses that reached binding energies of -8.7 kcal.mol^{-1}; for these systems this involved docking of 24, 100 and 14 conformers, respectively. Successively, substrates RLQAAF, RLQAAN, RLQAGA, RLQALG, RLQAVN, TLQAGF, TLQAVA, VLQAAF and VLQAVF were the only binding poses that reached binding energies of -8.6 kcal.mol^{-1}; this involved the docking of 35, 100, 10, 40, 35, 35, 10, 35 and 40 conformers, respectively. These substrates represented the high-affinity binding poses of the resulting docking. Conversely, the binding pose of KLQSKM was the only substrate to only reach a binding energy of -7.0 kcal.mol^{-1}; this even after having redocked 10 conformers. The binding poses of substrates KLQAEM and TLQSLM registered binding energies of -7.1 kcal.mol^{-1}, followed by the binding poses of KLQSEM, KLQSKD, MLQAKM, MLQSKM and VLQAKD which registered binding energies of -7.2 kcal.mol^{-1}. The number of conformers redocked for these substrates was 10, 10, 24, 14, 24, 14 and 10 (supplementary table 3.1). These binding poses represented the lowest-affinity binding in the full set of docking experiments. The docking results are summarised in supplementary table 3.1; detailing the number of conformers redocked per substrate, alongside the best binding conformer and its respective binding energy.

Overall, there were a total of 408 binding poses that registered energies of free binding ≤ -8.0 kcal.mol^{-1}. The remaining 402 binding poses registered energies of free binding >-8.0 kcal.mol^{-1}. At present, there are no similar studies on SARS-CoV-2 M^{pro} in the literature that provide for comparison. However, there have been several studies investigating peptide-based inhibition of the SARS-CoV-2 M^{pro} (Porto, 2021; Gahlawat *et al.*, 2020; Micco *et al.*, 2020). Micco *et al.* (2020) and Ansari *et al.* (2020) reported docking scores within the same range as these findings, whilst other studies reported much higher affinities of binding; however, these binding affinities were not Vina scores and as such comparison is difficult (Çakır *et al.*, 2021; Gahlawat *et al.*, 2020). However, these findings collectively support the use of peptide molecules as a basis for drug discovery and improved drug design. Moreover, they highlight the necessity to elucidate the proteolytic mechanism employed by the M^{pro} at the atomic level. The resulting knowledge will in turn instruct the approach for designing potent antiviral inhibitors. In as much as the binding affinity was not quantified, the

docking scores of the best poses are very promising. The quantification of the binding affinity could benefit the selection of the best substrates for lead optimization and rational drug design. In computational studies, high affinity strengthens the stability of the ligands in the receptor binding interface. Ligand stability enables the creation of stable complexes and reduces the chances of unbinding events of substrates during molecular dynamics studies. Unbinding events during such dynamics studies are indicative of probable ejection of the ligand from the site of binding *in vitro*.

3.6.4 DOCKING REPRODUCIBILITY

The substrates for whom 100 conformers were redocked (supplementary table 3.1) were used to assess the reproducibility of the best docking poses. These substrates included KLQAAA, KLQAAD, KLQAAE, KLQAAF, KLQAAV, KLQSAV, KLQSTD, MLQSLN, MLQSVM, RLQAAN, RLQATE, RLQSGA, RLQSGF, RLQSSA and VLQSGD as shown in figure 3.2. Most of the substrates attained a similar pose in the best binding geometries across docking experiments. The backbone (α-carbons) of the peptide substrates were overlapped when superimposed. However, the KLQSTD, RLQAAN and RLQATE substrates attained two or more main poses in their best binding conformers that did not overlap, yet with the same binding energy (figure 3.2). Interestingly, these three substrates represent some of the best docking results obtained in this study. The exploration of these alternative poses would prove advantageous in the characterisation of the mechanism of the SARS-CoV-2 M^{pro}, especially the profiling of substrate specificity and affinity.

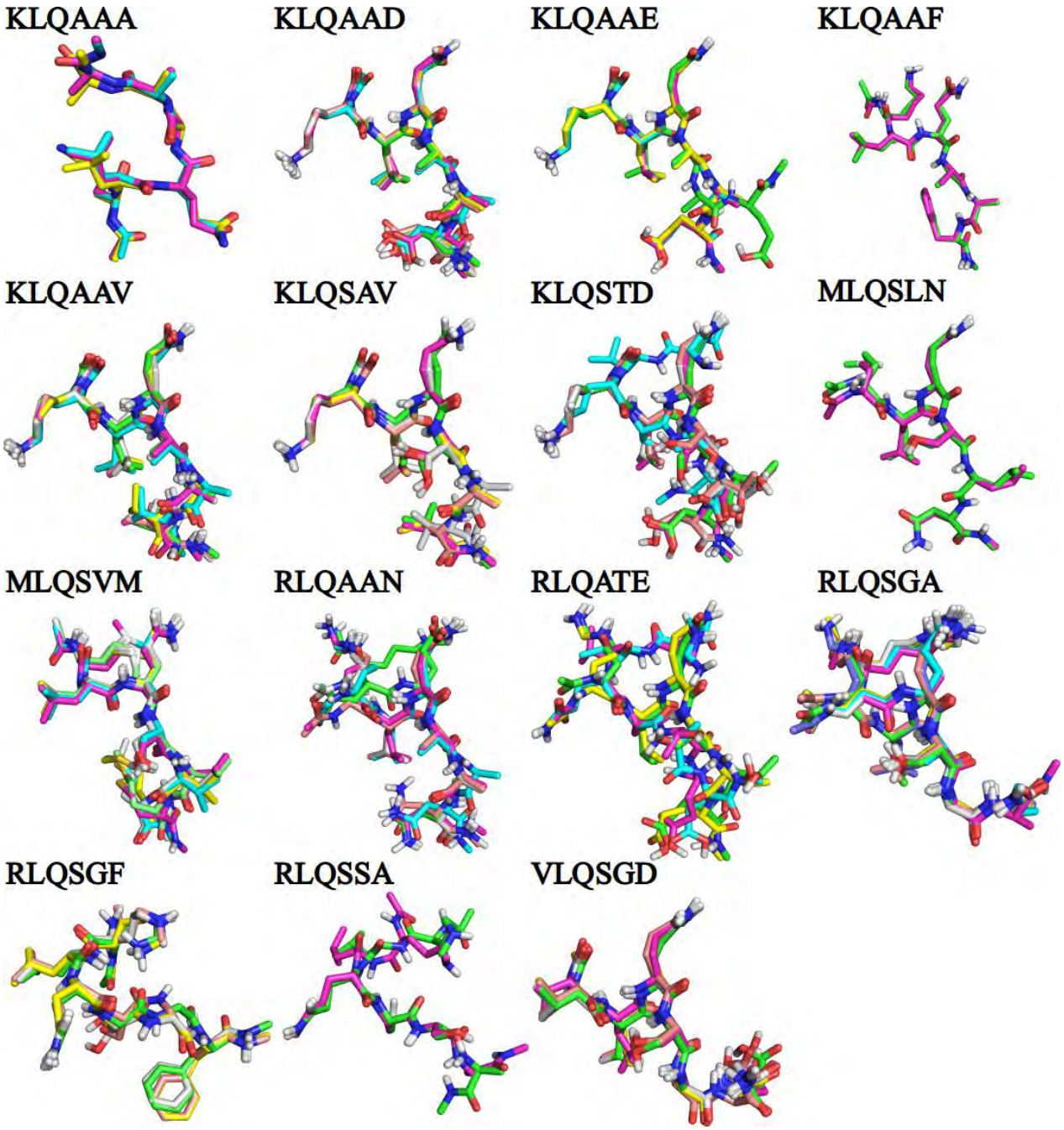


Figure 3.2 Validation of reproducibility of the docking results. The visualisation of the best poses of substrates with all 100 conformers docked. Image was generated using PyMOL.

High degrees of docking reproducibility was obtained in KLQAAA, KLQAAD, KLQAAV, KLQSAV, MLQSVM, RLQSGA, RLQSGF and VLQSGD substrates. These represent the substrates that attained more than two identical high-affinity docking poses across experiments, with overlap, and excluding the presence of alternative poses. Substrates KLQAAE, KLQAAF and MLQSLN displayed reproducibility, albeit with only two docking pose overlaps (figure 3.2).

3.6.5 LE OF VARIABLE PEPTIDES

The determination of LE for the range of hexapeptide substrates, bearing in mind the respective positions of residues on the peptides, was carried out to identify the amino acids that have a high (ΔG/HA of amino acid) contribution to the efficiency of substrate binding to the M^{pro}. These amino acids represented the fragments and/or constituents of the substrates that displayed high potential in the improvement of binding affinity. The underlying objective of this identification was to obtain optimal combinations of "efficient" amino acids that could serve as a basis for efficiency-driven drug design.

Before the calculation of LE, substrates were sorted according to the amino acids occupying the peptide position of interest, and the mean free energy of binding was calculated across all substrates with a particular amino acid residue at a specified position. The mean free energy of binding was divided by the number of heavy atoms in the specific amino acid to determine the LE of that amino acid in the full set of substrates. The values of the LE of peptides in their respective positions are reflected in table 3.1.

In the P3 position, amino acids threonine (T) and valine (V) attained the highest LE of -1.1, both having seven heavy atoms (HA). Following these amino acids was methionine (M) with a LE of -1.0, having eight HA (Table 3.1). All three amino acids showed promising potential for lead optimization as they represented the best results in the P3 position. However, valine proved to be supreme as it attained the lowest mean free energy of binding, indicating that several substrates containing valine at P3 attained lower binding energies compared to those containing threonine and methionine. Despite attaining the worst LE value in the group, arginine (R) had the best mean binding energy and the LE value was only reduced due to the high number of HA in the residue. It could still be important to consider arginine in lead optimization, given the observed favourable mean binding energy, and this could prove beneficial to the discovery of antiviral drugs against SARS-CoV-2.

In the P1′ (cleavage site) position, alanine (A) and serine (S) attained LE of -1.6 and -1.3, respectively. Both of these LE show promising potential for lead optimization. However, the alanine residue was shown to be more attractive as a P1′ candidate since it attained both the best mean binding energy and LE. The mean binding energies of the two residues were much closer in range in comparison to LE, owing to the differences in the number of HA in both of them.

Table 3.1: The ligand efficiencies of the hexapeptide substrates docked onto SARS-CoV-2 M^{pro} on basis of variable amino acid residues.

Position	Residue	mean ΔG (kcal.mol^{-1})	SD	HA	mean LE (kcal.mol^{-1}. per atom)	SD
P3	K	-8.0	0.26	9	-0.9	0.03
	M	-7.9	0.30	8	-1.0	0.04
	R	-7.9	0.33	11	-0.7	0.03
	T	-7.9	0.26	7	-1.1	0.04
	V	-8.0	0.23	7	-1.1	0.03
P1′	A	-8.0	0.27	5	-1.6	0.05
	S	-7.9	0.30	6	-1.3	0.05
P2′	A	-8.0	0.26	5	-1.6	0.05
	E	-7.9	0.24	9	-0.9	0.03
	G	-8.0	0.29	4	-2.0	0.07
	K	-7.7	0.27	9	-0.9	0.03
	L	-7.9	0.27	8	-1.0	0.03
	N	-8.0	0.25	8	-1.0	0.03
	S	-8.0	0.24	6	-1.3	0.04
	T	-8.0	0.26	7	-1.1	0.04
	V	-8.0	0.27	7	-1.1	0.04
P3′	A	-8.1	0.24	5	-1.6	0.05
	D	-7.9	0.26	8	-1.0	0.03
	E	-7.8	0.26	9	-0.9	0.03
	F	-8.1	0.30	11	-0.7	0.03
	G	-8.0	0.25	4	-2.0	0.06
	M	-7.7	0.23	8	-1.0	0.03
	N	-8.0	0.24	8	-1.0	0.03
	Q	-7.9	0.25	9	-0.9	0.03
	V	-8.0	0.22	7	-1.1	0.03

ΔG: free energy of binding; SD: standard deviation; HA: heavy atoms (non-hydrogen atoms); LE: ligand efficiency.

The examination of the LEs of the recognition sequences (P2-P1′) also showed results consistent with the above observations, showing that alanine was the efficient residue at P1′ in these results (Table 3.2). Since alanine was the most efficient residue of P1′, it is thus a favourable amino acid to consider in lead optimization. Nonetheless, the prioritisation of

serine as an alternative could be beneficial in drug design since both residues registered similar mean binding energies and also displayed similar occurrence frequencies (Ullrich and Nitsche, 2020).

Table 3.2: The ligand efficiencies of the hexapeptide substrates docked onto SARS-CoV-2 M^{pro} on basis of recognition sequence.

Sequence	mean ΔG ($kcal.mol^{-1}$)	SD	HA	mean LE ($kcal.mol^{-1}$per atom)	SD
LQ↓A	-8.0	0.28	22	-0.4	0.01
LQ↓S	-7.9	0.28	23	-0.3	0.01

ΔG: free energy of binding; SD: standard deviation; HA: heavy atoms (non-hydrogen atoms); LE: ligand efficiency.

In P2′, alanine, glycine (G) and serine were the highly efficient residues. The three amino acids were also among the best in terms of mean binding energy alongside asparagine (N), threonine and valine. Relating to LE, glycine attained the best value of -2.0, followed by arginine with -1.6 and serine with -1.3. The same explanation used for P1′ was also applied here. The reason for the differences in the LE values was owed to the differences in the number of HA atoms constituting the respective amino acids. Moreover, these three amino acids represented the smallest residues of the group; which explained why asparagine, threonine and valine attained poorer LE values in comparison. Glycine was the most favourable residue for lead optimization as it was the most efficient and the smallest of P2′. Binding efficiency and small size are desirable physicochemical and pharmacological properties in drug discovery as they relate to ADMET parameters. Similarly to the serine residue in P1′, alanine and serine (of P2′) could be used as alternatives.

P3′ residues showed similar patterns as seen in P1′ and P2′. The smallest amino acids of the group attained the best LE values. Alanine and glycine attained LE values of -1.6 and -2.0, respectively. The best binding was displayed by alanine and phenylalanine (F). Phenylalanine did not attain a favourable LE value due to its number of HA (Table 3.1). The prioritisation of alanine, glycine and phenylalanine could benefit the design of peptide-based drug candidates.

3.6.6 BEI VS SEI

BEI and SEI are some of the modified LE metrics that are widely used in drug discovery. Abad-Zapatero and Metz (2005) proposed the BEI and SEI to address the limitations of LE by developing easy to calculate indices that take account for differences between elements in different rows of the periodic table in compounds. BEI incorporates the ratio of potency and

molecular weight, whereas SEI incorporates the ratio of potency and PSA. Together, BEI and SEI combine three critical variables, namely potency, molecular weight and PSA. The combined use of BEI and SEI during optimization reduces the three variables to two and provides useful and comparable numerical scales for examining both indices simultaneously (Abad-Zapatero and Metz, 2005). In an optimization plane, the chemical series are placed on an SEI-BEI plot where both axis scales are similar, allowing for the simultaneous optimization of BEI and SEI. The general rule is to optimize the compounds towards the diagonal of the SEI-BEI plane, where both BEI and SEI are optimal (Abad-Zapatero and Metz, 2005).

The BEI and SEI of the substrates were calculated to account for the physicochemical properties that were excluded in LE (section 3.5.5) and plotted on an SEI-BEI plane per Abad-Zapatero and Metz (2005). The substrates were categorised by their P3-P1 residues. There were no substrates placed along the diagonal and all of the substrates were located within the lower portion of the diagonal on the SEI-BEI plane (figure 3.3). The placement of substrates on the plot showed that the compounds had high affinity per unit of PSA (Abad-Zapatero and Metz, 2005). This placement indicated that the compounds exhibited small PSA (high SEI), and low binding efficiency due to relatively large molecular mass. Promising compounds exhibit high BEI and SEI (lower molecular weight and PSA) as they relate to desirable pharmacokinetic properties (Abad-Zapatero and Metz, 2005).

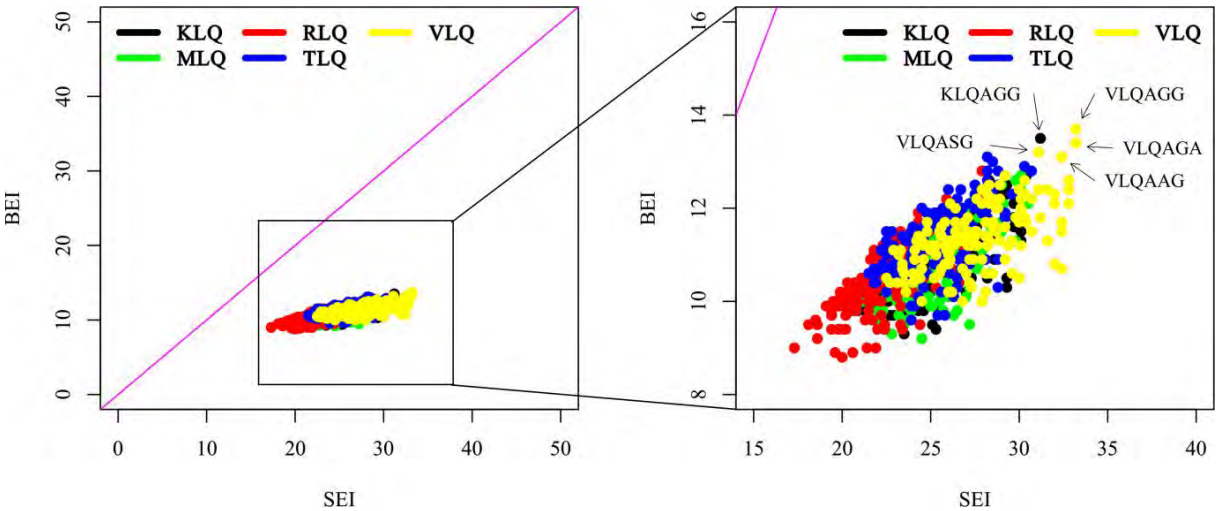


Figure 3.3. Mapping of surface-binding and binding efficiency indices in the SEI-BEI optimization plane for hexapeptide substrates of the SARS-CoV-2 M^{pro}. Substrates were categorised according to P1-P3 residues. The figure was generated using WPS Spreadsheets 2019 and RStudio.

The VLQ substrates were the most favourable compounds for lead selection, validation and optimization despite being the farthest from the optimization diagonal (figure 3.3). These

substrates attained the highest values of BEI and SEI in the entire results. The best overall substrates included VLQAAG, VLQAGA, VLQAGG and VLQASG, as well as KLQAGG. These substrates attained BEI values of 13.1, 12.5, 13.7, 13.2 and 13.5; and SEI values of 32.4, 28.9, 33.2, 31.1 and 31.2, respectively. After VLQ substrates, more promising BEI and SEI values were attained by a few hexapeptides belonging to KLQ, MLQ and TLQ substrates (figure 3.3).

The BEI-SEI results followed a similar trend with the LE results in section 3.5.5. The substrates containing Val or Thr at P3 displayed better ligand efficiency in comparison with other residues in the position. In addition, positions P1′ to P3′ were dominated by small residues in terms of LE which is also a visible trend in the BEI-SEI plane (figure 3.3).

RLQ substrates recorded the poorest BEI-SEI results despite being the closest to the diagonal (figure 3.3). Substrates with Arg (R) at P3 also attained poor LE in section 3.6.5. The simplest explanation for these recurring patterns relates to the size of residues constituting the RLQ substrates, especially the Arg residue. Molecular size is inversely related to BEI, and the large size of Arg (and other residues) in RLQ substrates accounted for the poor BEI values.

Despite registering the poorest overall BEI and SEI values, a few RLQ compounds demonstrated potential for lead optimization (figures 3.3). Because of this, RLQ substrates could not be written off as unfavourable ligands to consider for drug design, even though the majority had unfavourable characteristics. In addition, some RLQ substrates were among the best binding ligands in the docking results (supplementary table 3.1), attained the highest mean binding energy compared to other residues at P1′ (table 3.1), and were closest to the diagonal indicating that the substrates exhibited the best balance between BEI and SEI in the results (Abad-Zapatero and Metz, 2005; figure 3.3). These points further support the consideration of RLQ substrates for lead selection and optimization. Thus, the mapping of RLQ substrates according to their constituting residues allowed for the visualisation and identification of the favourable residues in promising compounds (figures 3.4 and 3.5).

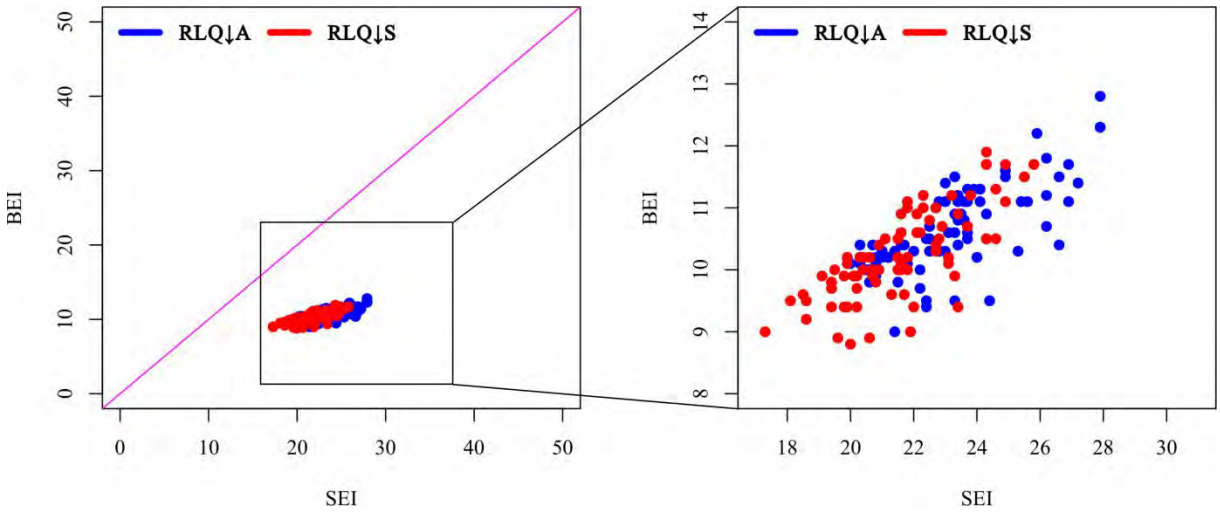


Figure 3.4. Mapping of surface-binding and binding efficiency indices in the SEI-BEI optimization plane for RLQ hexapeptide substrates of the SARS-CoV-2 M^{pro}. The RLQ substrates were grouped by P1′ residues. The figure was generated using WPS Spreadsheets 2019 and RStudio.

The mapping of the recognition sequence of RLQ substrates showed that Ala was the more efficient residue at P1′ over Ser. Substrates with Ala (P1′) attained the best BEI values and best SEI values in the group (figure 3.4). Nonetheless, there were select few substrates with Ser (P1′) that attained high values. Furthermore, the substrates closest to the diagonal consisted of Ser at P1′. Thus, further mapping of the RLQ substrates was carried out to identify the favourable residues in both Ala (P1′) and Ser (P1′) substrates.

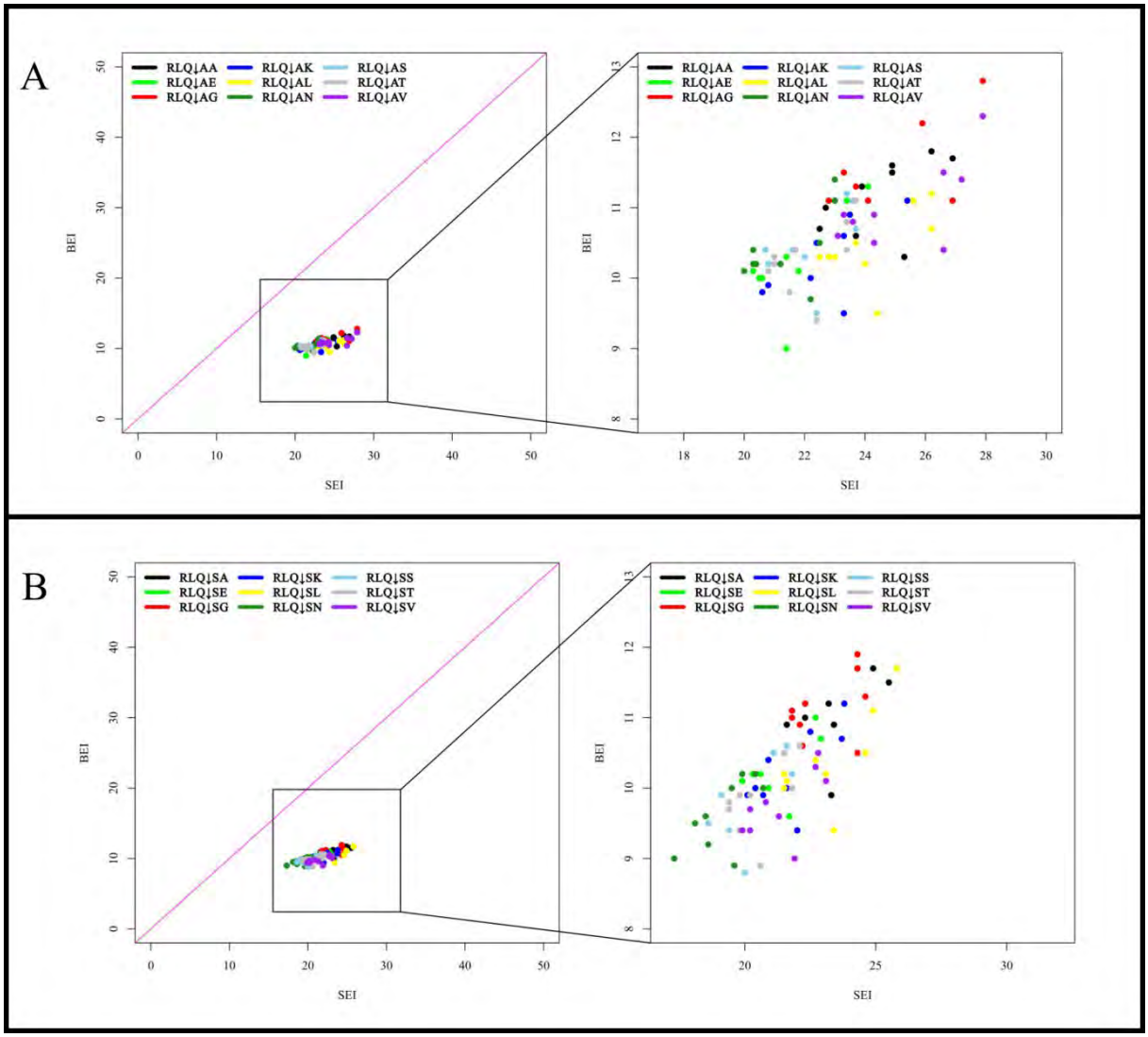


Figure 3.5. The identification of optimal residues constituting RLQ substrates hexapeptide substrates of the SARS-CoV-2 M^{pro}. The surface-binding and binding efficiency indices were mapped in the SEI-BEI optimization plane for RLQ hexapeptide substrates. The RLQ substrates were grouped by P2′ residues. **A)** shows the substrates with Ala at P1′. **B)** shows the substrates with Ser at P1′. The figure was generated using WPS Spreadsheets 2019 and RStudio.

For substrates with Ala (P1′), the highest SEI and BEI were mostly registered by substrates consisting of Ala, Gly or Val at P2′ (figure 3.5A). Ala and Gly have very small volumes/sizes, whereas Val is considered medium-sized (IMGT, 2020). For substrates with Ser (P1′), the highest SEI and BEI values were attained by substrates consisting of Ala, Gly or Leu at P2′ (figure 3.5B). Curiously, Leu is a large amino acid and this molecular weight decreases BEI (IMGT, 2020). Thus, the high placement of Leu (P2′) substrates was owing to the presence of very small residues occupying P3′, namely Ala and Gly. Nevertheless, both groups of RLQ substrates showed similar trends in terms of substrates closest to the diagonal (figure 3.5A and B). Substrates consisting of Asn at P2′ were most proximal to the diagonal, followed by substrates consisting of Ser and Thr at P2′.

3.6.6.1 BEI VS SEI - RECOGNITION SEQUENCE

The mapping of substrates on the SEI-BEI plot, according to the recognition sequence (now grouped by residues in P2-P1′), showed that the best values were attained by those consisting of Ala at P1′, as opposed to Ser (figure 3.6). The efficiency of Ala at P1′ was visible in LE results (tables 3.1 & 3.2) and RLQ BEI-SEI mapping (figure 3.4). This further supports the prioritisation of Ala at P1' during drug design. However, as emphasised in section 3.6.5, the use of Ser at P1′ (as a substitute at the very least) in drug design could prove beneficial as there were select promising substrates with Ser at the cleavage site (figures 3.4 & 3.6).

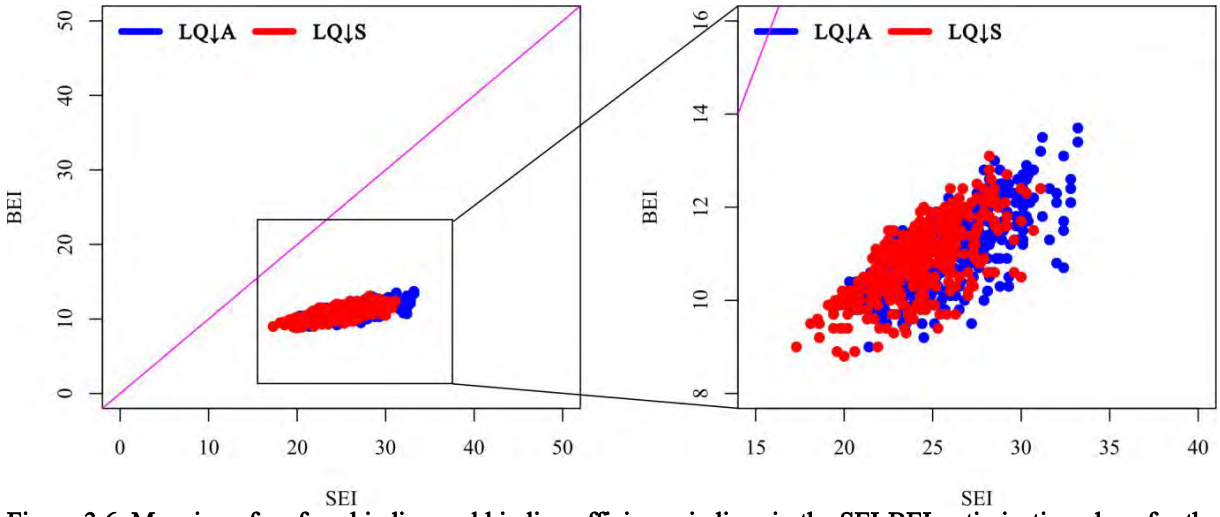


Figure 3.6. Mapping of surface-binding and binding efficiency indices in the SEI-BEI optimization plane for the recognition sequences of hexapeptide substrates of the SARS-CoV-2 M^{pro}. The substrates were grouped by P2-P1′ residues. The figure was generated using WPS Spreadsheets 2019 and RStudio.

3.6.7 SUBSTRATE RECOGNITION AND SPECIFICITY OF M^{PRO}

The SARS-CoV-2 M^{pro} mainly recognises substrate residue ranging from P4 to P1′ (Ullrich and Nitsche, 2020). However, the substrate specificity is determined by residues P2-P1′ as they display the highest degree of conservation among the pp1a/ab cleavage sites (Ullrich and Nitsche, 2020). These substrate residues are recognised and anchored onto the binding pocket by specific active site residues that comprise subsites. For binding P1 and P2, respectively, subsite 1 (S1) consists of His163, Glu166, Cys145, Gly143, His172, and Phe140, while S2 consists of Met49, His41, Cys145 and Thr25 (Khan *et al.*, 2020). Both S1 and S2 deeply bury the P1 and P2 residues and are involved in hydrophobic and electrostatic interactions (Mengist *et al.*, 2021; Khan *et al.*, 2020). S3 is comprised of residues Met165, Met49, and His41 (Lu *et al.*, 2006). Residues Thr25, Thr26, Leu27 and Cys145 constitute part of S1′ which generally form polar contact interactions with substrates (Mengist *et al.*, 2021). These substrates surround the catalytic dyad which consists of His41 and Cys145, that perform the

digestion of the polyproteins (Ullrich and Nitsche, 2020). Overall, the key active site residues that support substrate binding and processing include His41, Met49, Gly143, Ser144, His163, His164, Met165, Glu166, Leu167, Asp187, Arg188, Gln189, Thr190, Ala191 and Gln192 (Goyal and Goyal, 2020).

3.6.7.1 MAPPING OF SUBSITES

Subsites are crucial to the binding, anchoring and stabilization of substrates in the active site. Thus, the mapping of the subsites was carried out to assess whether the binding of the hexapeptides was following the nomenclature of Schechter and Berger (1967) and that each peptide corresponded to its respective subsite. Further, subsite mapping was performed to visualise the binding mode of the substrates in the active site, which in turn informed of the substrate recognition. This was performed using both the best binding poses which attained the lowest (figures 3.7 and 3.8), and the poorest performers with the highest (figures 3.9, 3.10 and 3.11) binding energies in the docking studies.

In as much as producing the best docking score (-8.7 kcal.mol^{-1}), the binding modes of RLQATF and RLQSTF were not in accordance with Schechter and Berger (1967). In both substrates, the side chain of P3 was anchored in S1, whereas the side chains of P2 and P1 were anchored in S3 and S2, respectively (figure 3.7). Nevertheless, RLQSGA demonstrated the desired binding mode as the side chains of P3-P1 residues interacted with corresponding subsites (figure 3.7).

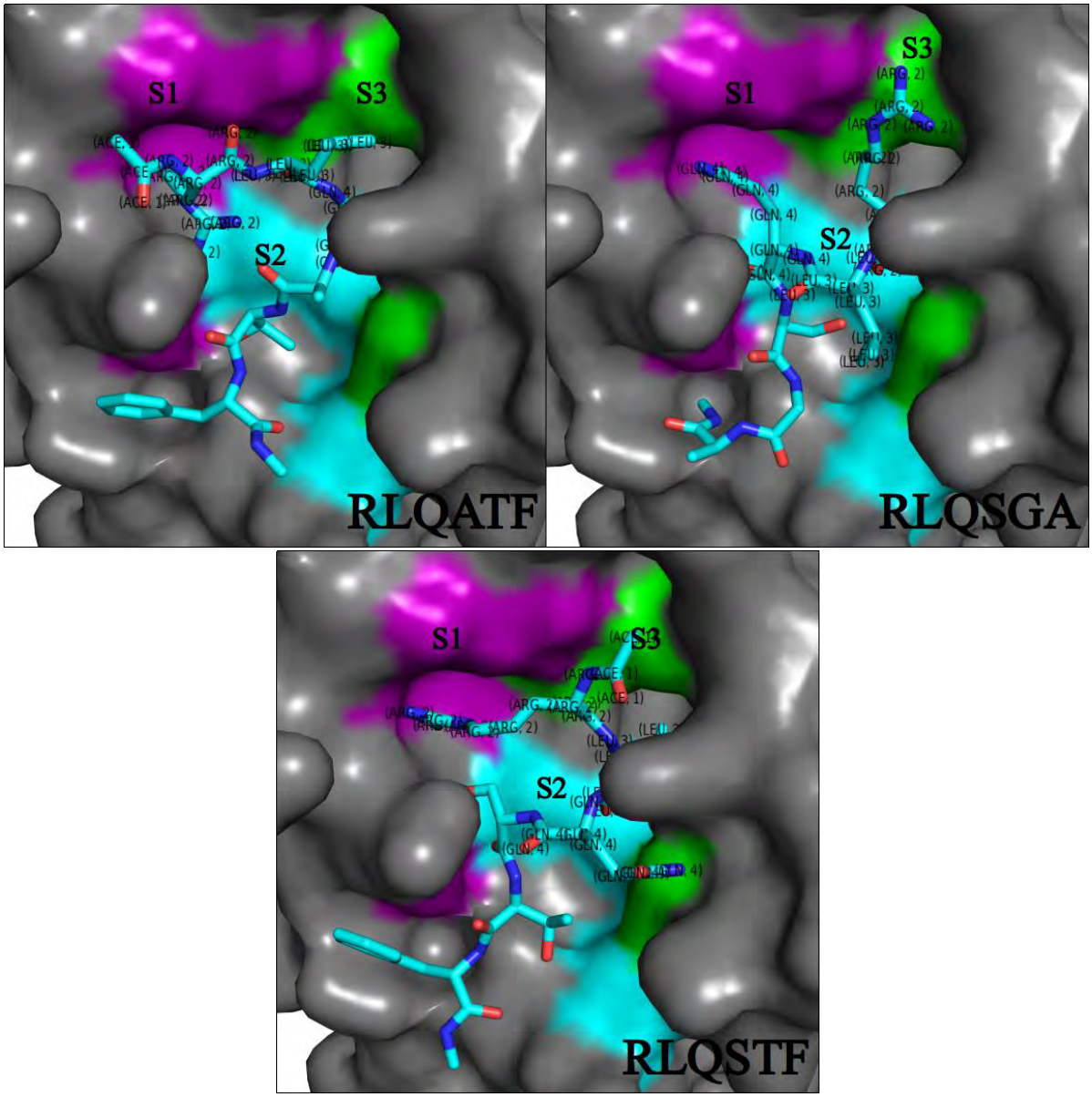
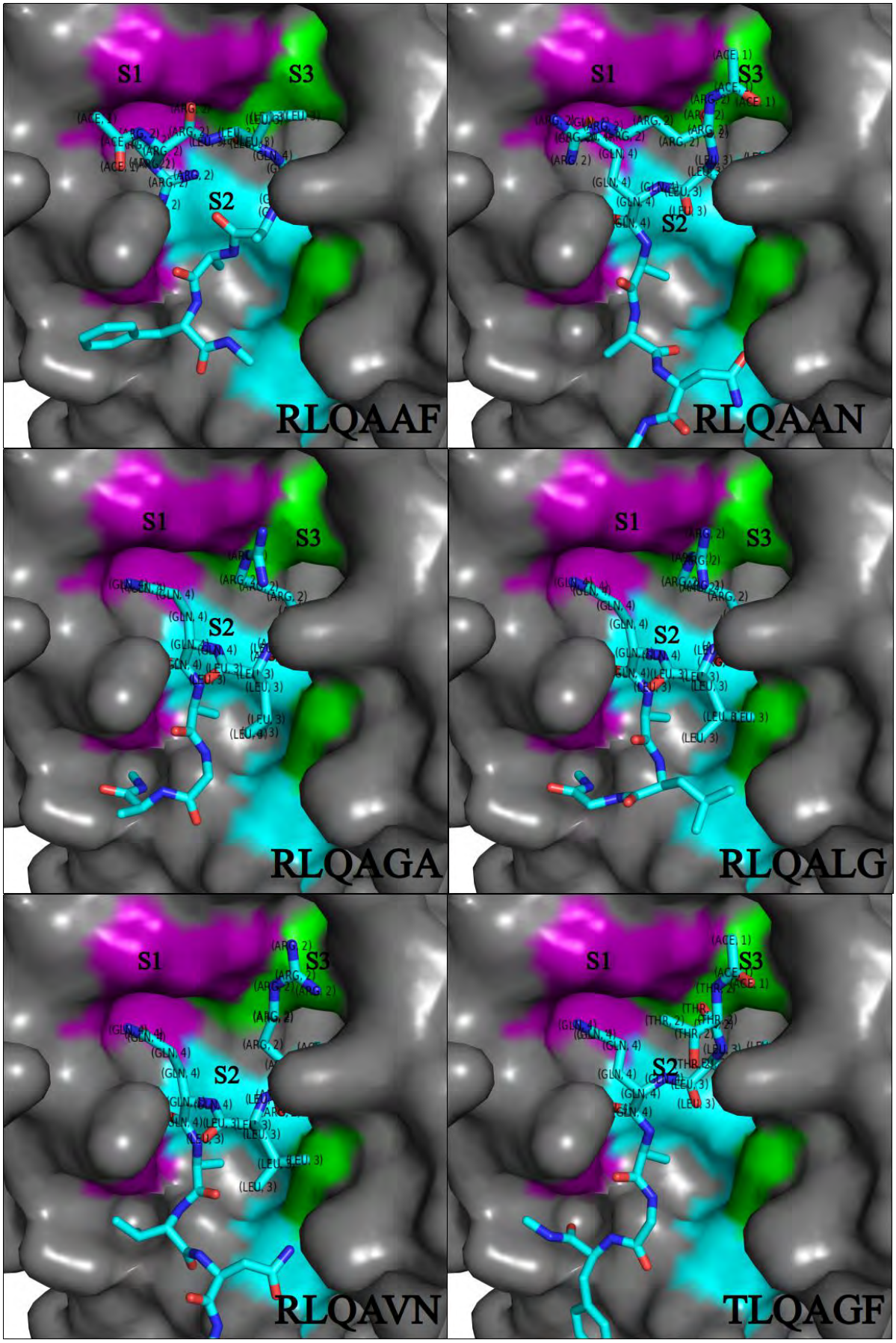


Figure 3.7 Confirmation of SARS-CoV-2 M^{pro} substrate recognition in binding poses for substrates RLQATF, RLQSGA and RLQSTF. The surface of SARS-CoV-2 M^{pro} (PDB ID:6XHM) showing docked substrates and substrate binding subsites colour-coded as follows: Purple: S1, Cyan: S2; Green: S3. The substrates attained a docking score of -8.6 kcal.mol^{-1}. The image was generated using PyMOL.

Desired binding modes were majorly evident in the second group of highest binding poses (-8.7 kcal.mol^{-1}; figure 3.8). While substrates RLQSGA, RLQALG, RLQAVN, TLQAGF, TLQAVA, VLQAAF and VLQAVF were all bound as expected in terms of protease-peptide binding modes, RLQAAF and RLQAAN violated the nomenclature of Schechter and Berger (1967). The binding mode of RLQAAF (figure 3.8) was similar to that of RLQATF and RLQSTF (figure 3.7). Curiously, the binding mode of RLQAAN showed S1 anchoring the side chains of P1 and P3, whilst S2 rightfully anchored P2 (figure 3.8). Interestingly, the ACE cap and P3 backbone were anchored in S3.


S1
S3
S2
RLQAAF
S1
S3
S2
RLQAAN
S1
S3
S2
RLQAGA
S1
S3
S2
RLQALG
S1
S3
S2
RLQAVN
S1
S3
S2
TLQAGF

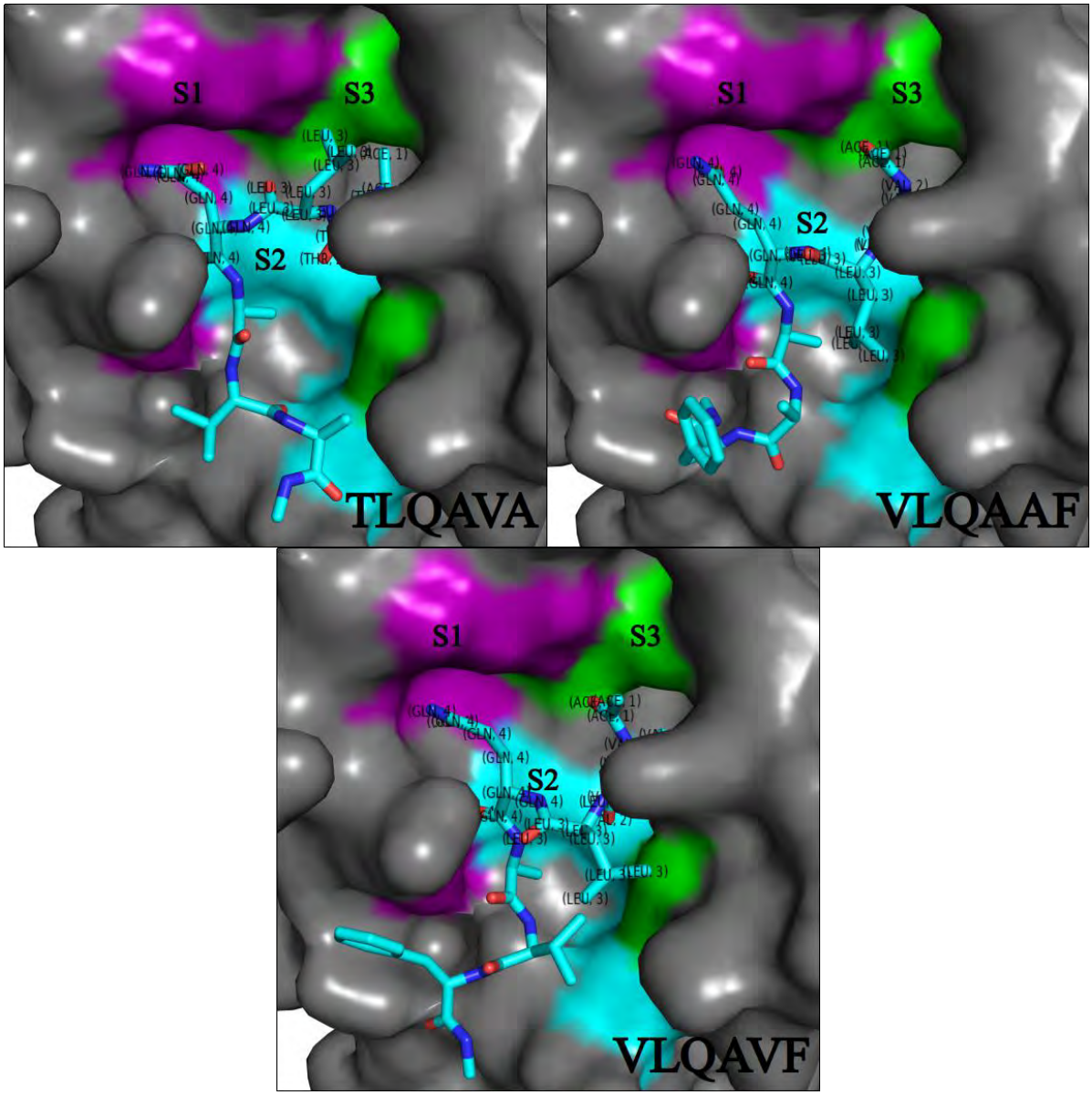


Figure 3.8 Confirmation of SARS-CoV-2 M^{pro} substrate recognition in binding poses for substrates RLQAAN, RLQAAF, RLQAGA, RLQALG, RLQAVN, TLQAGF, TLQAVA, VLQAAF and VLQAVF. The surface of SARS-CoV-2 M^{pro} (PDB ID:6XHM) showing docked substrates and substrate binding subsites colour-coded as follows: Purple: S1, Cyan: S2; Green: S3. The substrates attained a docking score of -8.6 kcal.mol^{-1}. The image was generated using PyMOL.

The violation of the nomenclature of Schechter and Berger (1967), was less prevalent in the protease-peptide complexes which attained the poorest docking scores (figures 3.9-3.11), with the exception of the KLQAEM complex (figure 3.10). Unlike previous violations of the binding rule in figures 3.7 and 3.8, KLQAEM displayed interactions between P3 and S3; P2 and S1; and P1 and S2 (figure 3.10).

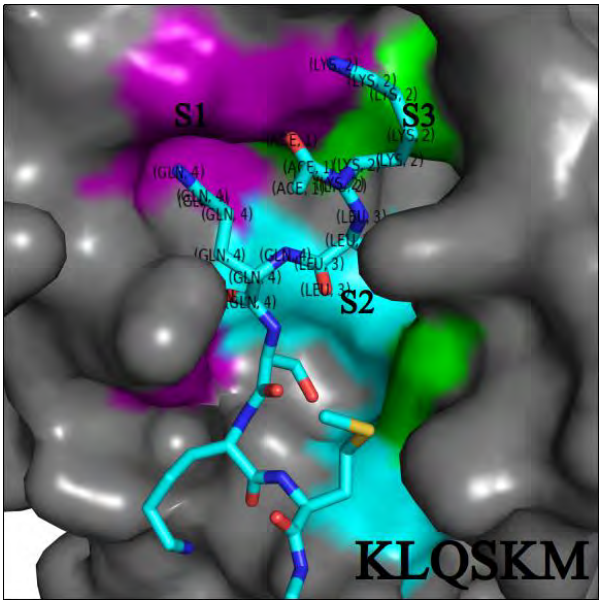


Figure 3.9. Confirmation of SARS-CoV-2 M^{pro} substrate recognition in binding poses for substrate KLQSKM. The surface of SARS-CoV-2 M^{pro} (PDB ID:6XHM) showing docked substrates and substrate binding subsites colour-coded as follows: Purple: S1, Cyan: S2; Green: S3. The substrates attained a docking score of -7.0 kcal.mol^{-1}. The image was generated using PyMOL.

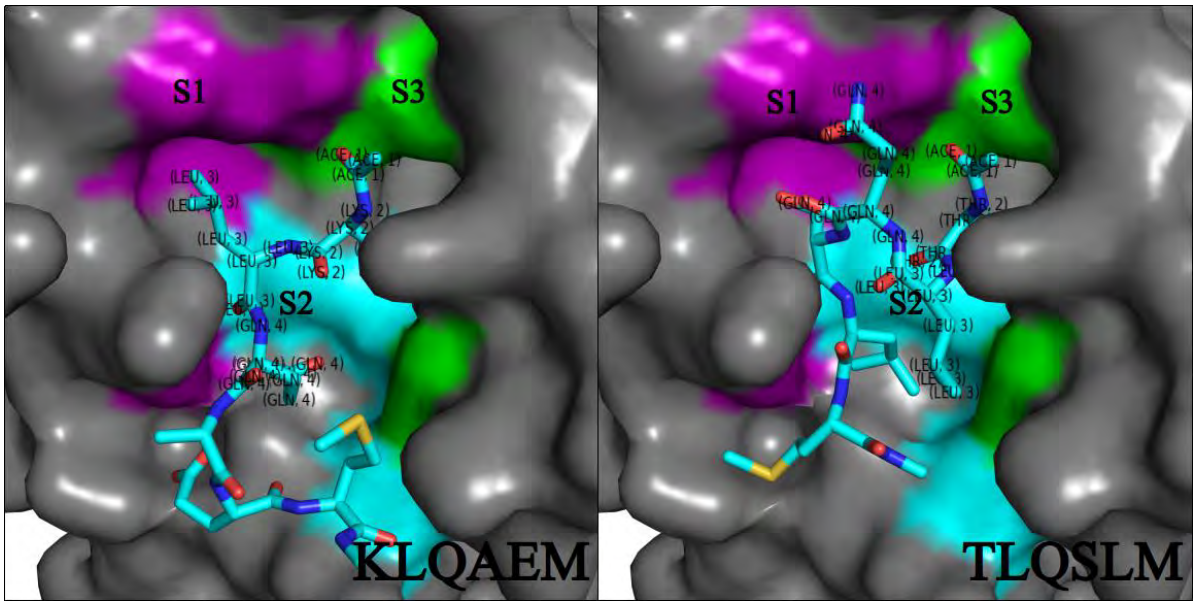
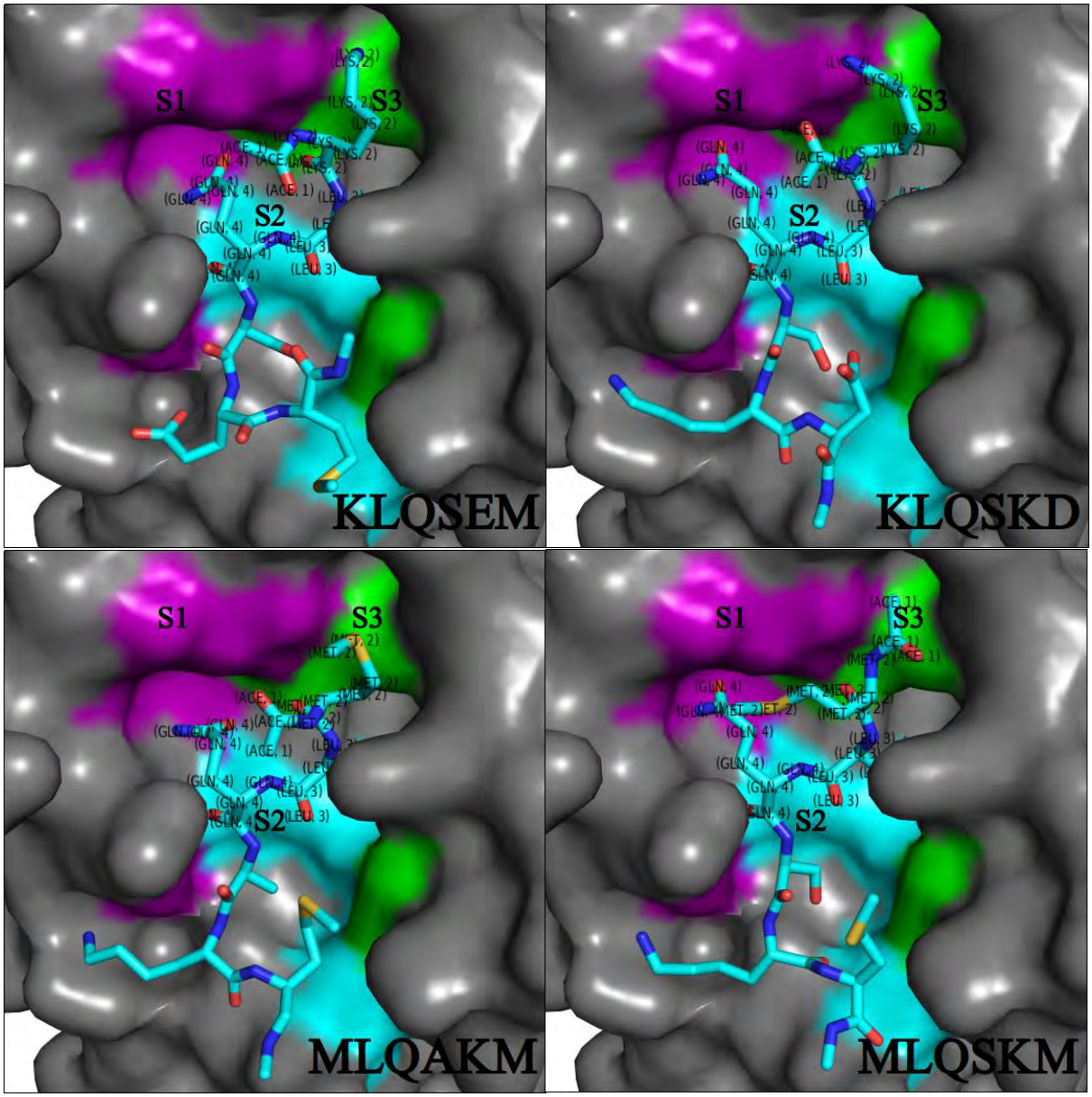


Figure 3.10 Confirmation of SARS-CoV-2 M^{pro} substrate recognition in binding poses for substrates KLQAEM and TLQSLM. The surface of SARS-CoV-2 M^{pro} (PDB ID:6XHM) showing docked substrates and substrate binding subsites color-coded as follows: Purple: S1, Cyan: S2; Green: S3. The substrates attained docking score of -7.1 kcal.mol^{-1}. Image was generated using PyMOL.


S1
S3
S2
KLQSEM
S1
S3
S2
KLQSKD
S1
S3
S2
MLQAKM
S1
S3
S2
MLQSKM

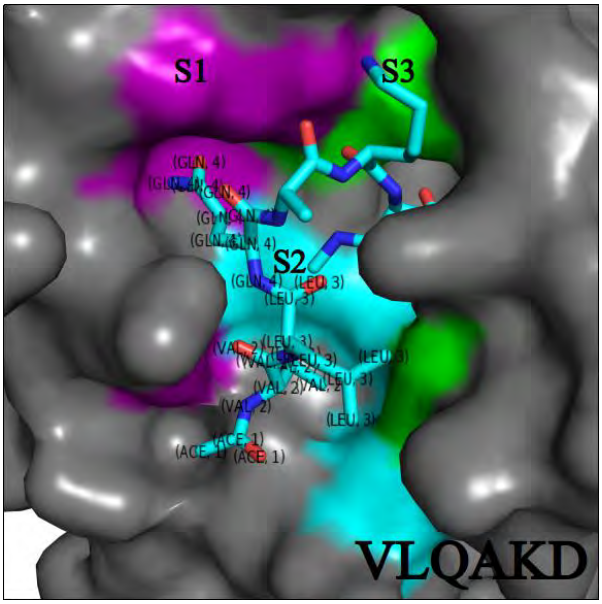


Figure 3.11 Confirmation of SARS-CoV-2 M^{pro} substrate recognition in binding poses for substrates KLQSEM, KLQSKD, MLQAKM, MLQSKM and VLQAKD. The surface of SARS-CoV-2 M^{pro} (PDB ID:6XHM) showing docked substrates and substrate binding subsites colour-coded as follows: Purple: S1, Cyan: S2; Green: S3. The substrates attained a docking score of -7.2 kcal.mol^{-1}. The image was generated using PyMOL.

Collectively, these binding modes indicate that the crystal structure of the M^{pro} displayed the substrate recognition of SARS-CoV-2 M^{pro}. The side chains of the hexapeptides were appropriately accommodated in the subsites of the active site, despite having large differences in free energies of binding. Furthermore, these results also inform that the generation of the hexapeptides was a success since the resultant structures were recognisable to the receptor protein. Substrate recognition of SARS-CoV-2 M^{pro} towards the hexapeptide was thus, confirmed. Therefore, hexapeptide-based drug design could prove beneficial to the pursuit of antiviral agents since recognition is well established.

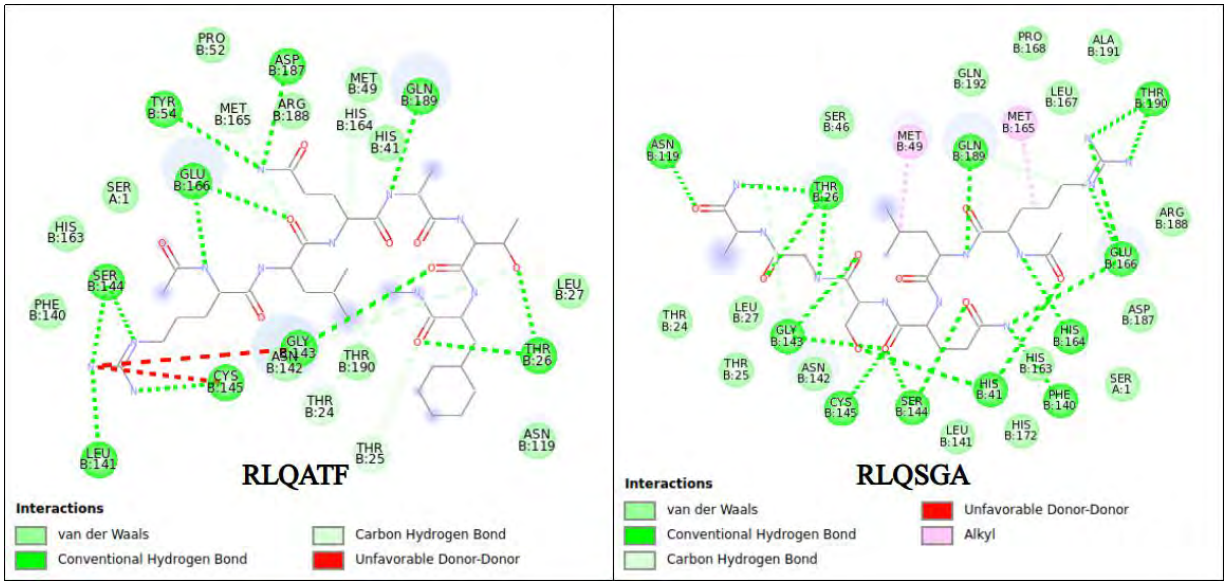
3.6.7.2 M^{PRO}-SUBSTRATE MOLECULAR INTERACTIONS

Protein-ligand molecular interactions at the binding site are a crucial part of characterising the binding affinity of the protein for promising fragments, hits and leads. Molecular interactions at the interface of protease and substrate in the protease active site determine substrate specificity (Schauperl *et al.*, 2015). Thus, Discovery Studio was used to resolve the protease-peptide molecular interactions and to generate 2D representations of these interactions shown in figures 3.12 to 3.16.

Hydrogen bonds were the most prevalent interactions throughout the complexed structures as shown in figures 3.12 to 3.16 and supplementary tables 3.2 to 3.4. These hydrogen bonds included conventional hydrogen bonds, carbon-hydrogen bonds and Pi-donor hydrogen bonds. Hydrogen bonds, especially conventional hydrogen bonds, are the fundamental stabilising

forces in biomolecular structures, which also underpin structure, function and conformational dynamics (Horowitz and Trievel, 2012; Herschlag and Pinney, 2018). In addition, complementarity and stability in protein-ligand complexes are largely owed to the formation of hydrogen bonds in the binding interface (Lippert and Rarey, 2009; Norel *et al.*, 1999). Therefore, the prevalence of hydrogen bond formation at the active site indicate shape and electrostatic complementarities between M^{pro} and hexapeptides and points towards high affinity of the protein towards the peptides. This is evident in binding energies in supplementary table 3.1. Furthermore, the stabilising effect of hydrogen bonds would be integral in maintaining the structures in dynamic processes.

The key functional residues, His41 and Cys145, formed various interactions with the substrates. Cys145 typically formed conventional hydrogen bonds with the oxygen atoms of the carboxyl group of P1 (RLQSGA, RLQAAN, RLQAGA, RLQALG, RLQAVN, TLQAVA, VLQAAF and VLQAVF) or P1′ (RLQSTF, TLQSLM KLQSEM and MLQAKM); placing the catalytic residue in close proximity to the cleavage site and the scissile peptide bond. Remarkably, most of these substrates demonstrated appropriate binding modes in compliance with the nomenclature of Schechter and Berger (1967). Other interactions with Cys145 included hydrogen bonds with the nitrogen atoms of P3 Arg (RLQATF and RLQAAF); Pi-Alkyl/Alkyl interaction with the β-carbon of P1′ Ala (TLQAGF); a hydrogen bond with the oxygen atoms of the carboxyl group of P2 Leu (KLQAEM); carbon-hydrogen bond with the β-carbon of P1′ Ser (MLQSKM); and Pi-Alkyl/Alkyl interaction with the β-carbon of P3 Val (VLQAKD).



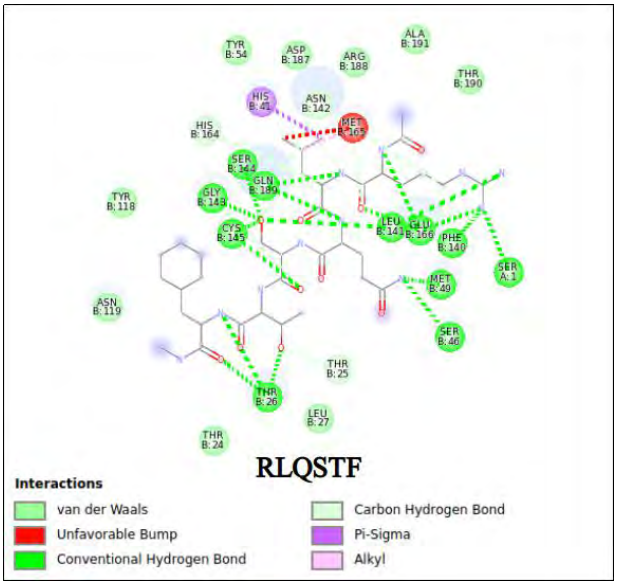
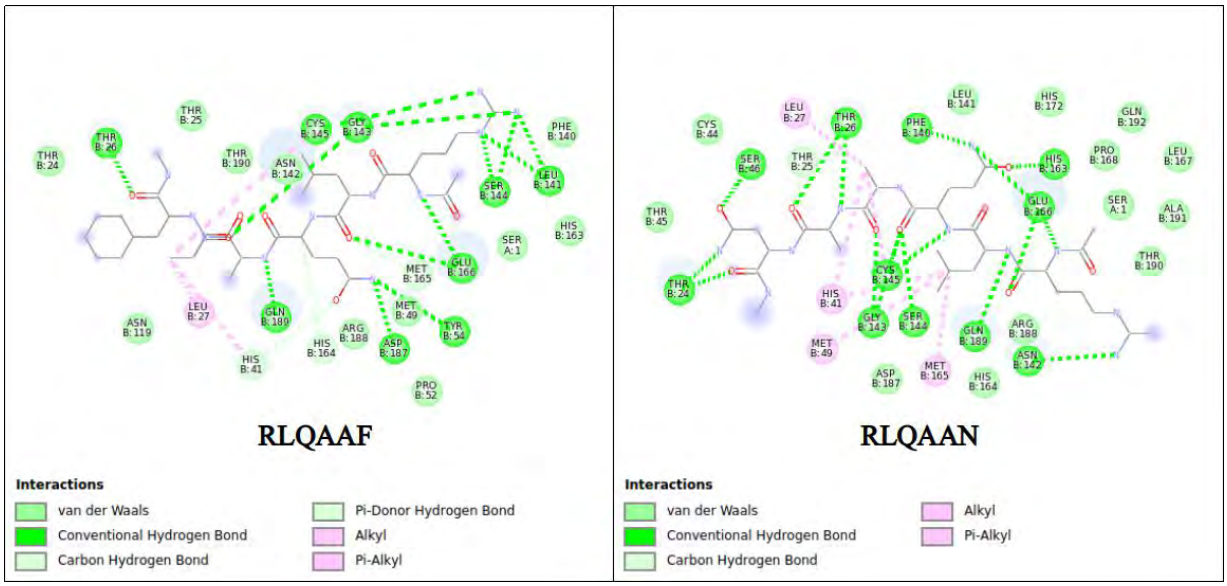
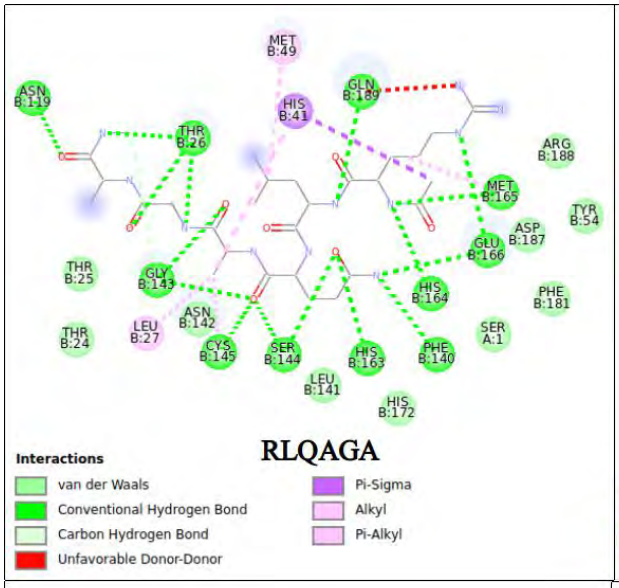
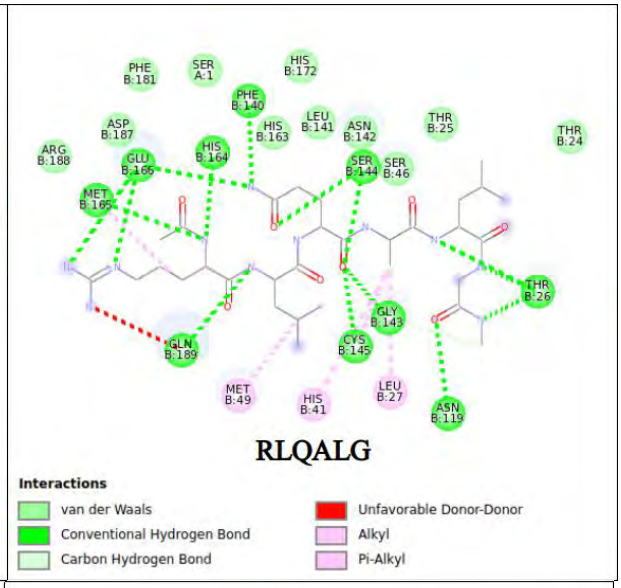


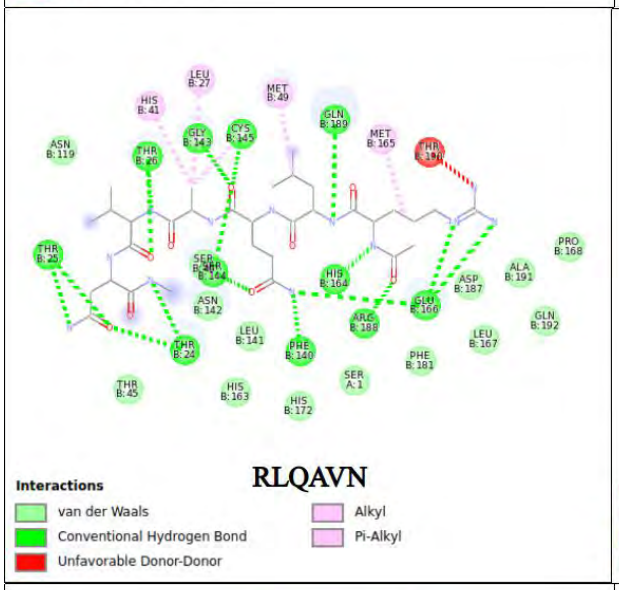
Figure 3.12. Resolution of intermolecular interactions between M^{pro} and substrates at the active site. 2D representation of the protein-ligand interactions at active sites for M^{pro} complexed with RLQATF, RLQSGA and RLQSTF. The images were generated on BIOVIA Discovery Studio 2020 Client.

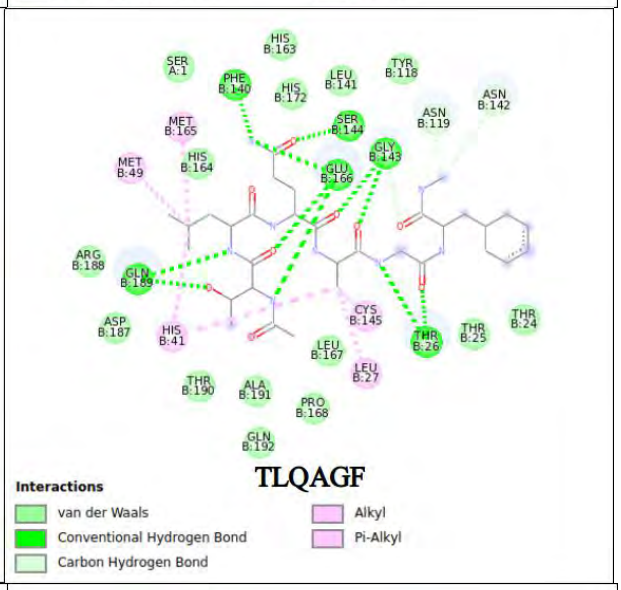
.

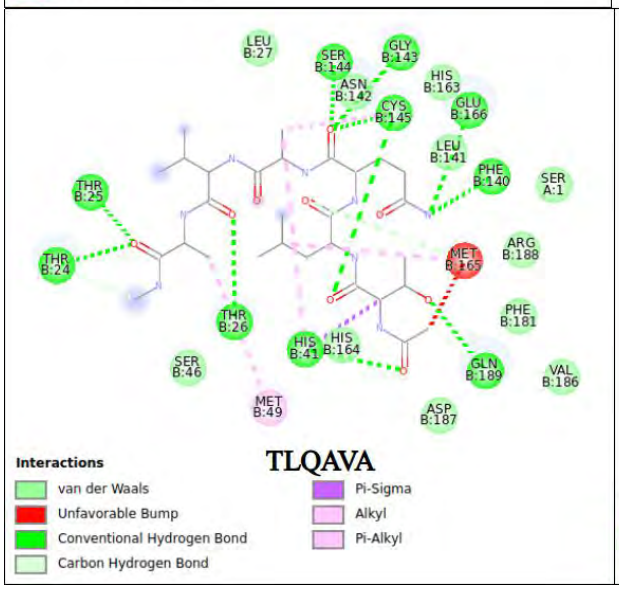


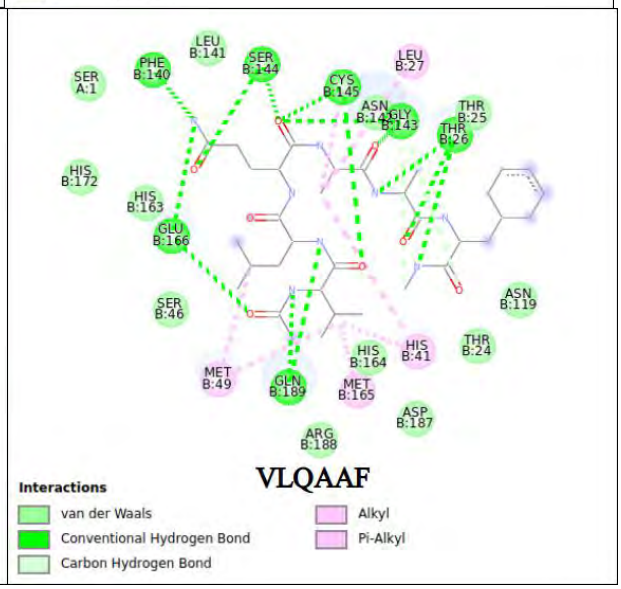

RLQAGA
Interactions
van der Waals
Conventional Hydrogen Bond
Carbon Hydrogen Bond
Unfavorable Donor-Donor
Pi-Sigma
Alkyl
Pi-Alkyl


RLQALG
Interactions
van der Waals
Conventional Hydrogen Bond
Carbon Hydrogen Bond
Unfavorable Donor-Donor
Alkyl
Pi-Alkyl


RLQAVN
Interactions
van der Waals
Conventional Hydrogen Bond
Unfavorable Donor-Donor
Alkyl
Pi-Alkyl


TLQAGF
Interactions
van der Waals
Conventional Hydrogen Bond
Carbon Hydrogen Bond
Alkyl
Pi-Alkyl


TLQAVA
Interactions
van der Waals
Unfavorable Bump
Conventional Hydrogen Bond
Carbon Hydrogen Bond
Pi-Sigma
Alkyl
Pi-Alkyl


VLQAAF
Interactions
van der Waals
Conventional Hydrogen Bond
Carbon Hydrogen Bond
Alkyl
Pi-Alkyl

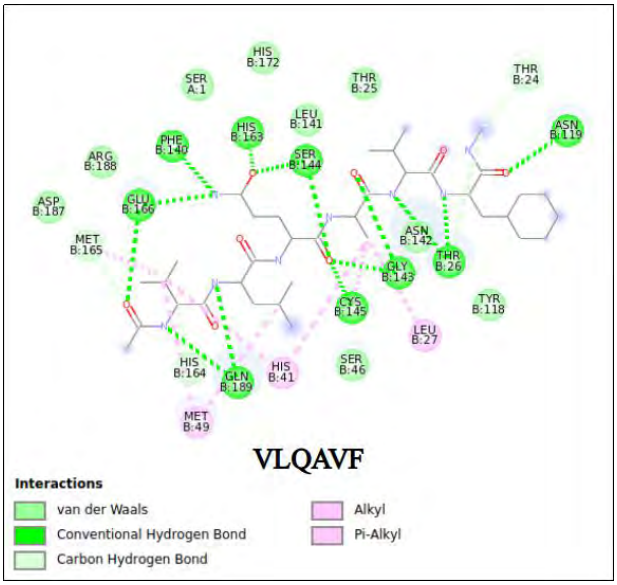


Figure 3.13. Resolution of intermolecular interactions between M^{pro} and substrates at the active site. 2D representation of the protein-ligand interactions at active sites for M^{pro} complexed with RLQAAF, RLQAAN, RLQAGA, RLQALG, RLQAVN, TLQAGF, TLQAVA, VLQAAF and VLQAVF. The images were generated on BIOVIA Discovery Studio 2020 Client.

The intermolecular interactions involving His41 varied to a much greater extent across the complexed systems as shown in figures 3.12 to 3.16. Across the systems, the typical interactions with His41 were Alkyl/Pi-Alkyl interactions with the β-carbon of P1′ Ala (RLQAAN; RLQAGA; RLQALG; RLQAVN; TLQAVA; VLQAAF; VLQAVF; MLQAKM) which were mostly among the top binding substrates; and Pi-Sigma interactions with a δ-carbon of P2 Leu (RLQSTF; KLQSKM; KLQSKD; MLQAKM). Much like with conventional hydrogen bonds involving Cys145, the Alkyl/Pi-Alkyl interactions with the β-carbon of P1′ Ala placed His41 in proximity with the cleavage site. Notably, these interactions were mainly found in substrates that demonstrated appropriate binding modes.

In figure 3.12, the top binding substrates were involved in additional interactions which included a cleavage site proximal interaction (RLQATF); hydrogen bonds with oxygen atoms of the ACE cap and the side chain of P1′ Ser (RLQSGA). For substrates with the binding energy of -8.6 $kcal.mol^{-1}$, additional interactions involving His41 are shown in figure 3.13. These included a Pi-donor hydrogen bond with the side chain oxygen atom of P1 Gln (RLQAAF); Pi-Sigma with the carbon atom of the ACE cap (RLQAGA); an Alkyl/Pi-Alkyl interaction with the γ-carbon of P2 Leu (TLQAGF); a hydrogen bond with oxygen atoms of the ACE cap together with Pi-Sigma with the α-carbon of P3 Thr (TLQAVA); and an Alkyl/Pi-Alkyl interaction with β-carbon P1 Val (VLQAVF).

For the poorest binding substrates, more intermolecular interactions with His41 were shown (figure 3.14-3.16). However, a frequent interaction in the group was the Alkyl/Pi-Alkyl

interaction with the γ-carbon of P2 Leu (KLQSEM and MLQSKM). Other interactions included a hydrogen bond with the oxygen atom of P1 Gln, and an Alkyl/Pi-Alkyl interaction with the γ-carbon of P3 Lys (KLQAEM); a Pi-donor hydrogen bond with the oxygen atom on the side chain of P3 Thr, and (TLQSLM); a carbon-hydrogen bond with the β-atom of P1′ Ser (MLQSKM); and lastly, a Pi-donor hydrogen bond with P3′ Asp alongside an Alkyl/Pi-Alkyl interaction with the β-carbon of P3 Val (VLQSKD). The catalytic residues engaged in varying interactions with the respective substrates. However, the common trend across complexes was Cys145 forming interactions (mostly hydrogen bonds) with the backbone atoms around the scissile bond, whereas His41 was forming interactions with the side chains of peptides around the scissile bond. The explanation for this pattern could owe to the flexible nature of the peptide chains, which resulted in a torsional rotation that orientated the substrates in a way that the substrates could only interact with His41 via side chains.

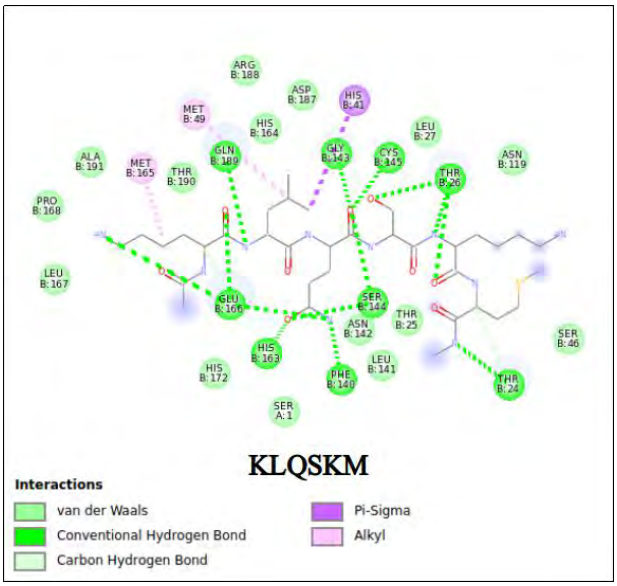


Figure 3.14. Resolution of intermolecular interactions between M^{pro} and substrates at the active site. 2D representation of the protein-ligand interactions at the active site for M^{pro} complexed with KLQSKM. The images were generated on BIOVIA Discovery Studio 2020 Client.

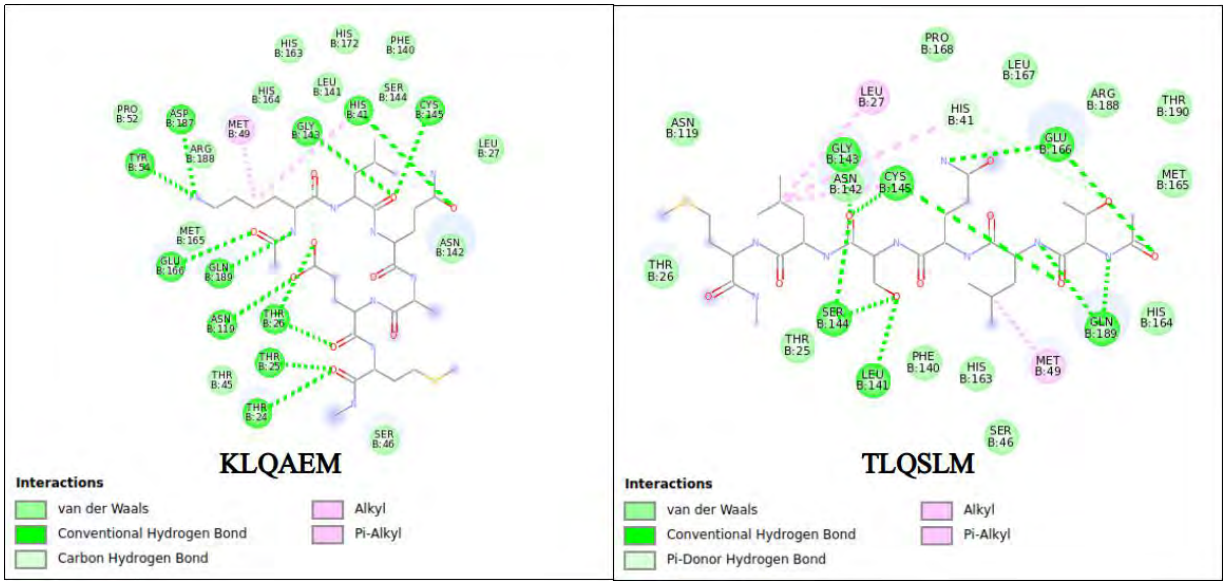
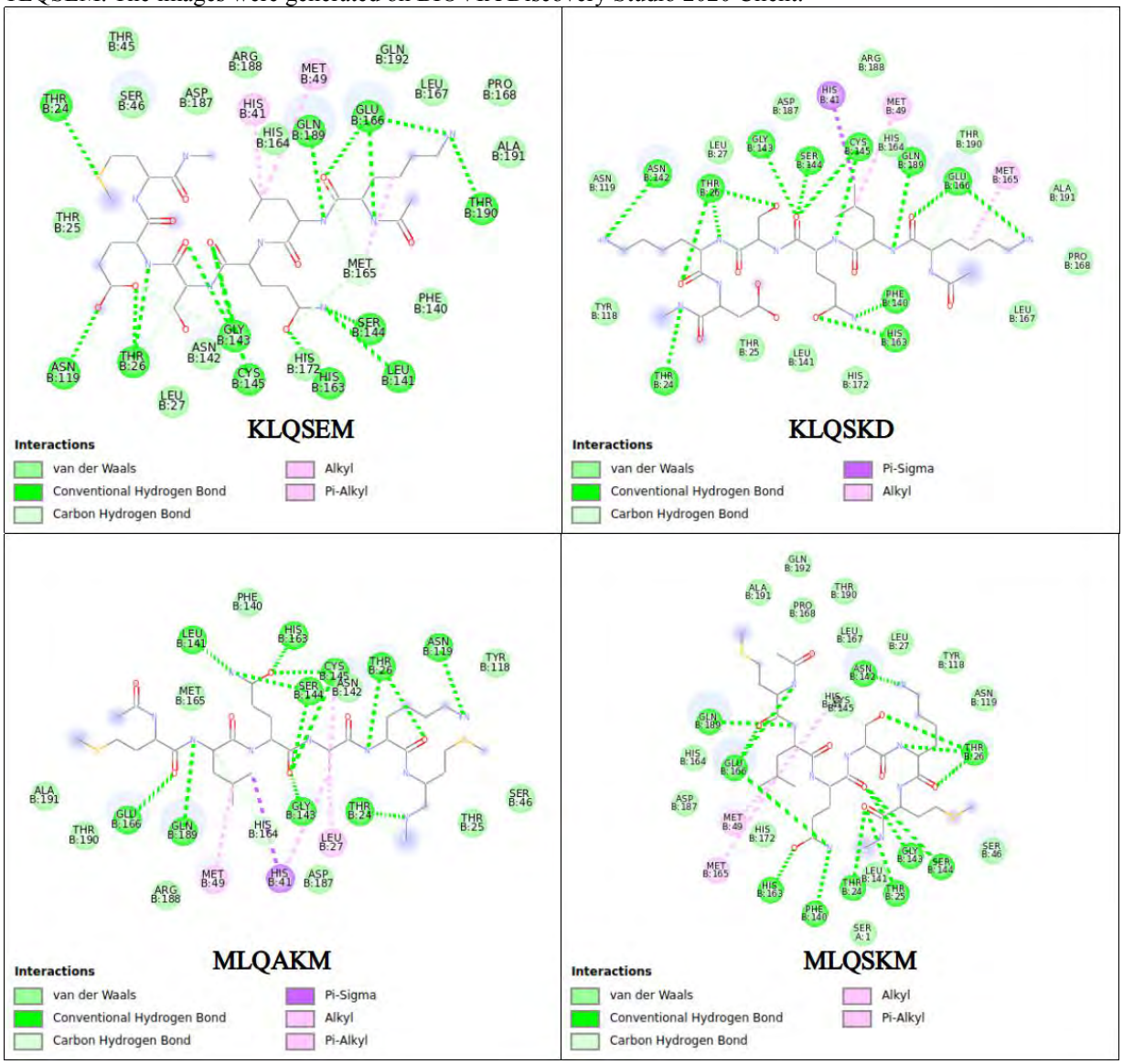


Figure 3.15. Resolution of intermolecular interactions between M^{pro} and substrates at the active site. 2D representation of the protein-ligand interactions at active sites for M^{pro} complexed with KLQAEM and TLQSLM. The images were generated on BIOVIA Discovery Studio 2020 Client.



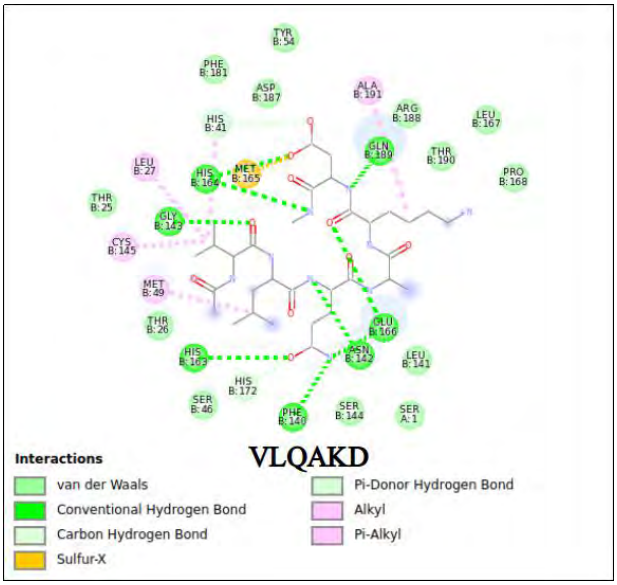


Figure 3.16. Resolution of intermolecular interactions between M^{pro} and substrates at the active site. 2D representation of the protein-ligand interactions at active sites for M^{pro} complexed with KLQSEM, KLQSKD, MLQAKM, MLQSKM and VLQAKD. The images were generated on BIOVIA Discovery Studio 2020 Client.

Key residues that play an important role in substrate binding and substrate stabilization in the binding pocket also participated in non-covalent interactions with the substrates. These included the residues 138 to 146 which constitute the so-called oxyanion loop that confers substrate stability during the proteolytic process. The backbones of Gly143 and Cys145 form the oxyanion hole which in turn stabilises the partial negative charge that arises at the P1 carbonyl group of the substrate during the hydrolysis of the scissile bond (Suárez and Díaz, 2020). In supplementary tables 3.2 to 3.4, these residues are (indicated in purple) participated in the key stabilising forces (hydrogen bonds and van der Waals forces of attraction) that promoted the formation of stable complexes.

Other key residues include Met49, His163, His164, Met165, Glu166, Leu167, Asp187, Arg188, Gln189, Thr190, Ala191 and Gln192, which underpin the subsites and accommodate the appropriate binding of substrate residues (Goyal and Goyal, 2020; Muramatsu *et al.*, 2016; Hsu *et al.*, 2005). These residues mostly mediate the binding of substrates onto their respective subsites via side-chain rearrangement and also contribute to anchoring the substrates (Lee *et al.*, 2020; Muramatsu *et al.*, 2016; Hsu *et al.*, 2005). Similar to the oxyanion loop residues, these residues also formed hydrogen bonds with substrates and inferred stability to the bound substrates (supplementary tables 3.2-3.4). Thus, stable complexes were formed, with many stabilising interactions. The prevalent formation of hydrogen bonds at the binding interface pointed towards substrate specificity.

The mapping of sites and non-covalent interactions were also resolved for KLQ***

complexes (supplementary figures 3.2-3.3). For the most part, the aforementioned trends in substrate binding modes with relation to subsites and intermolecular interactions were consistent with the patterns discussed in this chapter. The mapping of subsites revealed high recognition of substrate residues by corresponding subsites in most systems, excluding the KLQASV, KLQAVQ, KLQSAF, KLQSAQ, KLQSSF, KLQSTE, and KLQSVF complexes. Nevertheless, substrate recognition was suggested by the binding modes of the KLQ*** substrates. In a similar manner to the substrates discussed above, KLQ*** substrates also showed a prevalence of hydrogen bond formation between the substrates and the active site residues of M^{pro}. In terms of the catalytic residues, the typical interactions discussed above were also prevalent in KLQ complexes. Cys145 was seen frequently forming hydrogen bonds with oxygen atoms of the carboxyl group of residues P1 or P1′, whereas His41 interacted with various side chains of the substrates. Subsequently, the KLQ*** complexes were used in molecular dynamic studies to assess the behaviour of the complexed systems and examine the stability between M^{pro} and substrates conferred by the intermolecular interactions. The selection of only KLQ*** complexes for further study was a practical one, given the extensive computational resources required, and to essentially conduct dynamic studies on a complete subset of systems.

3.7 CHAPTER SUMMARY

In this chapter, the conformers of the hexapeptides were docked onto a suitable crystal structure of SARS-CoV-2 M^{pro}. Before the docking studies, a preliminary study was carried out to determine which chain would produce better docking results and the results favoured the use of chain B. Docking studies were thus conducted on chain B using the exhaustiveness of 480, and repeated with 10 or more starting substrate conformations. The resulting free energies of binding ranged between -8.6 and -7.0 kcal/mol. The reproducibility of docking results was visually assessed in select substrates where 100 conformers were redocked per substrate. Reproducibility was achieved in high-affinity binding poses. LE was performed to determine the binding efficiencies of the constituent residues. In P3, Val was shown to be the most efficient residue in the position despite Arg registering the best binding energies. In P1′, Ala was a more efficient residue in comparison to Ser, although Ser produced better docking results in terms of binding energy. Gly and Ala were consistently more efficient residues at P2′ and P3′, with Gly attaining the most desirable LE score through the docking results. The favourable LE scores were a result of the small sizes of residues since the majority of the residues did not attain the best docking scores in their respective positions. The most efficient

recognition sequence or cleavage site was shown to be LQA, over LQS. A BEI-SEI plane was constructed for all the substrates, using the BEI and SEI indices which accounted for the physicochemical properties that LE does not account for. VLQ substrates were shown to be most desirable for lead optimization, followed by KLQ and TLQ substrates. Despite being the least desirable, RLQ substrates were curiously the closest to the diagonal of the SEI-BEI plane and were further analysed which in turn informed that RLQAG, RLQAV, RLQAA, RLQAA, RLQSG and RLQSL substrates displayed high potential for lead optimization. Subsite mapping showed the binding modes of the top (RLQATF, RLQSGA, TLQSTF, RLQAAF, RLQAAN, RLQAGA, RLQALG, RLQAVN, TLQAGF, TLQAVA, VLQAAF and VLQAVF), and poorest (KLQSKM, KLQAEM, TLQSLM, KLQSEM, KLQSKD, MLQAKM, MLQSKM and VLQAKD) binding poses. Interestingly, appropriate substrate binding was more prevalent in the poorest binding substrates as the hexapeptide residues were anchored by corresponding subsites at the binding interface. Across the top and bottom-most binding poses, the hexapeptides bound favourably to the active site of M^{pro}; hydrogen bonds were the prominent intermolecular interaction and these were formed with binding pocket residues. The catalytic dyad, alongside the oxyanion loop and other key residues, also formed favourable interactions with the substrates. Cys145 typically interacted with the atoms of the backbone of P1 or P1′, whilst His41 formed interactions with side chains of residues proximal to the cleavage site and various others. Key substrate binding residues and the oxyanion loop residues mostly formed stability conferring hydrogen bonds and van der Waals forces of attraction with the substrates. Subsite mapping alluded to substrate recognition and the resolution of non-covalent interactions pointed towards substrate specificity of M^{pro} for the hexapeptides. KLQ*** substrates were also assessed for subsite binding modes and intermolecular interaction. Overall trends were consistent with the typical patterns seen in both the top-performing and poorest binding substrates.

CHAPTER FOUR

MOLECULAR DYNAMICS SIMULATIONS AND TRAJECTORY ANALYSIS

COVID-19 has dealt a devastating blow to the world. The widespread distribution of the disease is propagated by the exponential transmission of the pathogenic SARS-CoV-2. Variants of the coronavirus have conferred even greater transmission and infection rates and enriched the virus with advanced immunological evasive mechanisms (Harvey *et al.*, 2021). The increased virulence is attributed to mutations in the spike protein that mediates cell entry. Thus, the investigation of antiviral agents which specifically target highly conserved non-structural proteins has become even more imperative to combat SARS-CoV-2-related illnesses and deaths and will allow the development of highly effective broad-spectrum treatments against the SARS-CoV-2 variants and other coronaviruses. As such we explore in this study sequences that are recognized by the SARS-CoV-2 M^{pro}. This part of the study is a complete study on molecular dynamics of substrate sequences KLQ*** (P3′-P1′) of length 6 in complex within the M^{pro}, detailing the stability and prominent motion/conformational changes of the protein-substrate systems.

4.1 INTRODUCTION

Proteins are dynamic entities in cellular solution with functions governed essentially by their dynamic personalities. Protein dynamics are manifested as changes in molecular structure, or conformation as a function of time. *In silico* and *ab initio* techniques assist in the resolution of protein structures which serve as a solid basis for structure-function studies that contribute to the elucidation of many dynamic aspects of enzymatic mechanism such as substrate binding, orientation, catalysis, and product release. Molecular dynamics are the prominent technique used in protein dynamics to approximate the interactions and behaviour of proteins in protein-protein or protein-ligand complexes. The simulations treat both binding partners as flexible entities, allowing for motion and conformational changes that provide insight into protein function and mechanism of action (Yang *et al.*, 2014; David and Jacobs, 2014;

Salsbury, 2010).

4.2 MOLECULAR DYNAMIC SIMULATIONS

Molecular dynamics (MD) simulations are a powerful tool to study and understand the structure and behaviour of protein systems with extreme detail - in scales where the motion of individual particles can be tracked (Lindahl, 2008). MD simulations are widely used in the elucidation of protein structure-to-function relationships, as they are useful in studies of protein folding events, enzymatic catalytic mechanisms, protein conformational changes, and allostery (Hospital *et al.*, 2015). Rational drug design also relies on MD simulations in the elucidation of molecular recognition, the binding and unbinding of drugs and overall mechanisms of action for a drug and its target (Galeazzi, 2009; Do *et al.*, 2018).

MD simulations treat all particles in a system as flexible entities and simulate their movements and dynamic behaviour as a function of time (Salmaso and Moro, 2018). During a simulation, the trajectories of all atoms in a system are computed by the solution of Newton's laws of motion (also known as the Classical mechanics), and an empirical force field (Berendsen *et al.*, 1995; Salmaso and Moro, 2018; Binder *et al.*, 2004). In principle, the classical equations of motion (Newton's equations of motion) are solved by using the forces between atoms to compute successive atomic configurations and to assess the movement of these atoms based on their interactions (Adcock, and McCammon, 2006). To solve Newton's equations of motion, velocities are calculated using the Maxwellian distribution centred on the desired temperature, using the atom positions are obtained from the coordinate information in the structure PDB file (van der Spoel *et al.*, 2005). In these computations, the molecule is described as a series of charged spheres or radii (atoms) linked by springs (bonds) based on molecular mechanics, to decrease the computational cost associated with macromolecular simulations (Vanommeslaeghe and Guvench, 2014). In addition, the movement of atoms is calculated in small steps, based on the Cartesian coordinates of the particle, allowing the sampling of molecular motion on the nanosecond and microsecond scale, and consequently enables the study of millisecond scale processes such as protein folding (Abraham *et al.*, 2015). The interactions (inter- or intramolecular, and forces) that mediate movement are described and evaluated by the force field of choice (Berendsen *et al.*, 1995; Binder *et al.*, 2004; Heinz *et al.*, 2013). In essence, MD simulations involve the solution of equations of motions detailing the forces acting on all atoms in a system. The most efficient and common way for the forces to be calculated in large biological systems is through molecular mechanics and the use of force fields.

4.2.1 FORCE FIELDS

A force field is a mathematical expression that describes the dependence of the energy of a system on the coordinates of its constituent particles (González, 2011). A force field consists of a functional form of the inter-atomic potential energy of a system, together with the set of molecular mechanical parameters for different types of atoms, chemical bonds, dihedral angles, out-of-plane interactions, non-bonded interactions, and other possible terms that fit into this form (González, 2011; Heinz *et al.*, 2013). These parameters are typically obtained from *ab initio* or semi-empirical quantum mechanical calculations, or by fitting to experimental data (such as neutron, X-ray and electron diffraction, NMR, infrared, Raman and neutron spectroscopy, etc.) (González, 2011).

The components of a force field include bonded terms for interactions in atoms linked by covalent bonds computed using Hooke's law, and non-bonded terms that describe the long-range electrostatic and van der Waals forces which are computed using Coulomb's law and a Lennard-Jones potential, respectively (Heinz *et al.*, 2013). Notably, the terms described in a force field are very specific and may vary from terms in other force fields, but the general expression of the total energy in a force field is written as follows:

$$E_{\text{total}} = E_{\text{bonded}} + E_{\text{nonbonded}}$$

Equation 4.1: Summation of the bonded and non-bonded components of the total energy in a force field.

where the terms of the bonded and non-bonded contributions are generally expressed by the following summations:

$$E_{\text{bonded}} = E_{\text{bond}} + E_{\text{angle}} + E_{\text{dihedral}}$$
$$E_{\text{nonbonded}} = E_{\text{electrostatic}} + E_{\text{van der Waals}}$$

Equation 4.2: Summations of the bonded and non-bonded terms constituting the bonded and non-bonded components of a force field

Commonly used force fields in biological simulations include CHARMM (Brooks *et al.*, 2009), NAMD (Phillips *et al.*, 2005), AMBER (Case *et al.*, 2005), GROMACS (van der Spoel *et al.*, 2005) and GROMOS (Christen *et al.*, 2005). In terms of protein simulations, these force fields typically provide the parameters for standard and non-standard amino acid residues, as well as few post-translational modifications (Margreitter *et al.*, 2013).

In practical terms, molecular dynamics will involve steps of: 1) topology generation; 2)

defining periodic boundary conditions; 3) solvation; 4) neutralization; 5) minimization; 6) equilibration (2 steps for temperature and pressure); 7) production dynamics; 8) trajectory analysis. In this present study, the protein dynamics of the SARS-CoV-2 M^{pro}-hexapeptide complexes were explored using GROMACS software.

4.2.2 GROMACS

GROningen MAchine for Chemical Simulation or GROMACS is an open-source software widely used in the dynamical simulation studies of biomolecules, in aqueous and membrane environments. The software features a powerful set of calculation types, preparation and analysis tools and supports several advanced techniques for free-energy calculations. GROMACS exhibits much more flexibility as it supports the accurate use of different force fields which are useful in studying biomolecular dynamics. GROMACS tools are implemented by the use of *gmx* commands to perform the preparation, running and analysis of MD simulations and trajectory analysis. GROMACS is optimized for complex calculations through the use of multi-CPU and even multi-GPU, which contribute to the acceleration of the performance (van der Spoel *et al.*, 2005; Abraham *et al.*, 2015). GROMACS offers Message Passage Interface (MPI) which allows the splitting of a computation into independent units of work that are handled in parallel (Abraham *et al.*, 2015). Numerous MPI schemes are implemented through enhanced parallelization algorithms. Moreover, optimal performance of the toolkit is attributed to the combined use of multi-GPUs and MPI schemes, which achieve optimization through acceleration and parallelization. Similarly, MPI tools, which compile preparation and simulation calculations with parallelization algorithms, are implemented with the use of *gmx_mpi* commands (GROMACS, 2015). In this study GROMACS was used at the Center for High Performance Computing (CHPC), running on 24 cores across 8 nodes (total of 192 cores per simulation) with a wall time of 24h on the PBS queue management system.

4.2.3 TRAJECTORY ANALYSIS

MD simulations yield a wealth of data about the structure, dynamics, and function of biomolecules by modelling the physical interactions between their atomic constituents. In order to deduce meaningful conclusions from the simulations, MD trajectories need to be analysed in terms of the positions (and possibly velocities and forces) of individual or selected subsets of atoms for each time frame of the trajectory (Michaud-Agrawal *et al.*, 2011). The trajectory files obtained contain the position of all atoms at frames during the

simulation, and these may be used to provide structural information on the course of dynamics.

4.2.3.1 ROOT MEAN SQUARE DEVIATION (RMSD)

The RMSD of certain atoms in a molecule is a measure of distance, or dissimilarity, between molecular conformations, with respect to a reference structure. In terms of MD simulations, RMSD is used as a primary benchmark to measure how structures or parts of structures change over time in comparison to the reference structure - which is frequently the first frame of the trajectory. Relatively constant RMSD values signify proper convergence and stabilization with the RMSD having a narrow range after equilibration. This indicates stabilization of the backbone (Gowers *et al.*, 2019).

4.2.3.2 RADIUS OF GYRATION (Rg)

The Rg is defined as the distribution of atoms of a protein around its axis. In MD simulation, the Rg measures the degree of compactness and folding of protein systems, wherein a constant Rg value in the simulation period signifies that there is protein folding stability; conversely, protein unfolding is indicated by changes in Rg values over time (Lobanov *et al.*, 2008; Sneha and Doss, 2016).

4.2.3.3 ROOT MEAN SQUARE FLUCTUATION (RMSF)

The RMSF is a measure of the deviation between the position of a particle with respect to its reference position. Unlike RMSD which is averaged over the particles to give time specific values, RMSF is averaged over time to give values for each particle. RMSD calculates the overall deviation of the structure from its reference structure, whereas RMSF determines individual residue flexibility and thus allows the identification of the most mobile/flexible regions during a simulation (BioChemCoRe 2018, 2021; Dong *et al.*, 2018).

4.2.3.4 PRINCIPAL COMPONENT ANALYSIS (PCA)

PCA, or essential dynamics, is an advanced analysis tool for identifying essential protein motions in trajectories from MD simulations. In principle, PCA extracts concerted protein motion in different frames during simulations to identify the most prominent motions of the protein backbone. PCA is conducted in two steps, involving the construction of a variance/covariance matrix using α-carbons and the diagonalization of the covariance matrix. The covariance matrix is routinely constructed from the atomic fluctuations after the removal

of the translational and rotational movement. The diagonalization of the covariance matrix yields a set of eigenvectors and eigenvalues. The eigenvectors show the directions in a 3 N-dimensional (the N is the number of atoms used for constructing the covariance matrix) conformational space and describe the majority of the protein motion along those directions. On the other hand, the eigenvalues are measurements of the mean square fluctuations of the system along the corresponding eigenvectors. The underlying assumption for PCA is that only a few eigenvectors with large corresponding eigenvalues are important for describing the overall motions of a protein. Therefore, the motion of the protein is identified by projecting the original data onto the first two eigenvectors (ev1 vs ev2), to create the first two principal components (PC1 and PC2) which contain the maximum motions (Srikumar *et al.*, 2014; Yang *et al.*, 2014).

4.3 METHODOLOGY

4.3.1 TOPOLOGY GENERATION

Following molecular docking, the hexapeptide (KLQ***) PDBQT output files were converted to PDB format using OpenBabel. The custom Perl script (Appendix M) was used to restore the amino acid information of the constituent residues of each best binding pose (Appendices L & N). The hexapeptide PDB information was then added onto the receptor PDB as a third subunit, thus constructing complexes as a single PDB file with three protein chains (the first two from the SARS-CoV-2 M^{pro} homodimer, the third the hexapeptide KLQ***) (Appendix O). The topology and coordinate files for the *apo*-M^{pro} and M^{pro}-Hexapeptide systems were created using the GROMACS version 2018.1 *pdb2gmx* command, employing the AMBER03 protein, nucleic AMBER94 force field (Duan *et al.*, 2003). The topology file (.top) contains all the necessary information to define the molecule within a simulation, including non-bonded (atom types and charges) and bonded (bonds, angles and dihedrals) parameters. The configuration file (.gro) contains all the coordinates of the molecule in the system, together with its corresponding parameter files (.itp) which contain the connection of atoms in each subunit.

4.3.2 BOX DEFINITION, SOLVATION AND ADDITION OF IONS

To establish an aqueous and neutral system that mimics cellular conditions *in vitro*, the M^{pro} systems were subjected to solvation and neutralisation. The *apo*-M^{pro} and M^{pro}-hexapeptide systems were solvated in a cubic box of dimension 10 nm using a TIP3P water model. The

structures were centred in the cubic box and placed under periodic boundary conditions (PBC). Subsequently, the systems were neutralised to a net charge of zero by the addition of the Na^+ and Cl^- counter ions.

4.3.3 ENERGY MINIMIZATION

To avoid steric clashes and unfavourable geometries in the systems arising from solvation and neutralisation, the M^{pro} systems were subjected to energy minimization. The minimization was prepared using the GROMACS *grompp* command and minimization was initiated with the GROMACS *mdrun* command. Energy minimization was performed using the steepest descent minimization algorithm for 50000 steps and minimization was set to stop when the maximum force of <10.0 kJ/mol was achieved to avoid high-energy interactions.

4.3.4 SYSTEM EQUILIBRATION

In order to optimise the solvent and the ions surrounding the protein structures, equilibration was carried out to bring the systems to desirable simulation temperature and pressure. The temperature of the system was equilibrated with an NVT ensemble (constant Number of atoms, Volume and Temperature) at 300K for 100 ps. Sequentially, the pressure was equilibrated with an NPT ensemble (constant Number of atoms, constant Pressure and constant Temperature) at 1.0 bar for 100 ps. A modified Berendsen thermostat was employed in both equilibration ensembles. In both steps position restraints for heavy atoms were included.

4.3.5 MD PRODUCTION

Following the temperature and pressure equilibration, the system position restraints were released and 20 ns MD production runs were executed using the GROMACS *mdrun*. The time steps were set at 2 fs and the trajectory and coordinate information were saved every 10 ps, resulting in 2000 frames saved for every system simulated. Energy minimization, equilibration and production runs were automated on the CHPC cluster, employing multiple nodes and CPU cores to compensate for the computational expense.

4.3.6 TRAJECTORY ANALYSIS

Upon the completion of the production runs for the 20 ns simulations, the structures were removed from the PBC simulation box and centred within the box using the *trjconv* GROMACS command. The trajectory files were analysed by calculating the RMSD, RMSF,

and Rg using the GROMACS commands *rms*, *rmsf* and *gyrate*, respectively. Conformational changes and structural motions over the course of the trajectory for the protein backbone were monitored using PCA. The dynamics of the structures over the simulation time were visually inspected using Visual Molecular Dynamics (VMD) (Humphrey *et al.*, 1996).

4.4 RESULTS AND DISCUSSION

The strength and stability of a protein-ligand complex are related to the intermolecular interactions between these binding partners (Pantsar and Poso, 2018). Hence, MD simulations for the *apo*- (unbound) and complexed proteins were carried out to assess the binding strength of protein and substrates in the complexed systems, and to gain insight into changes in the structure and stability of M^{pro} as a result of binding the KLQ*** hexapeptides. Thus, 20 ns trajectories of the various *apo* and KLQ-complexed M^{pro} systems were analysed and plotted to assess such changes, globally (figures 4.1-4.2), locally (figure 4.3) and based on their prominent protein motions during the trajectory (figures 4.4-4.10). 131 KLQ*** substrates were used in the final data set for MD and were subsequently analysed.

4.4.1 GLOBAL STRUCTURAL STABILITY OF THE SARS-CoV-2 M^{PRO}

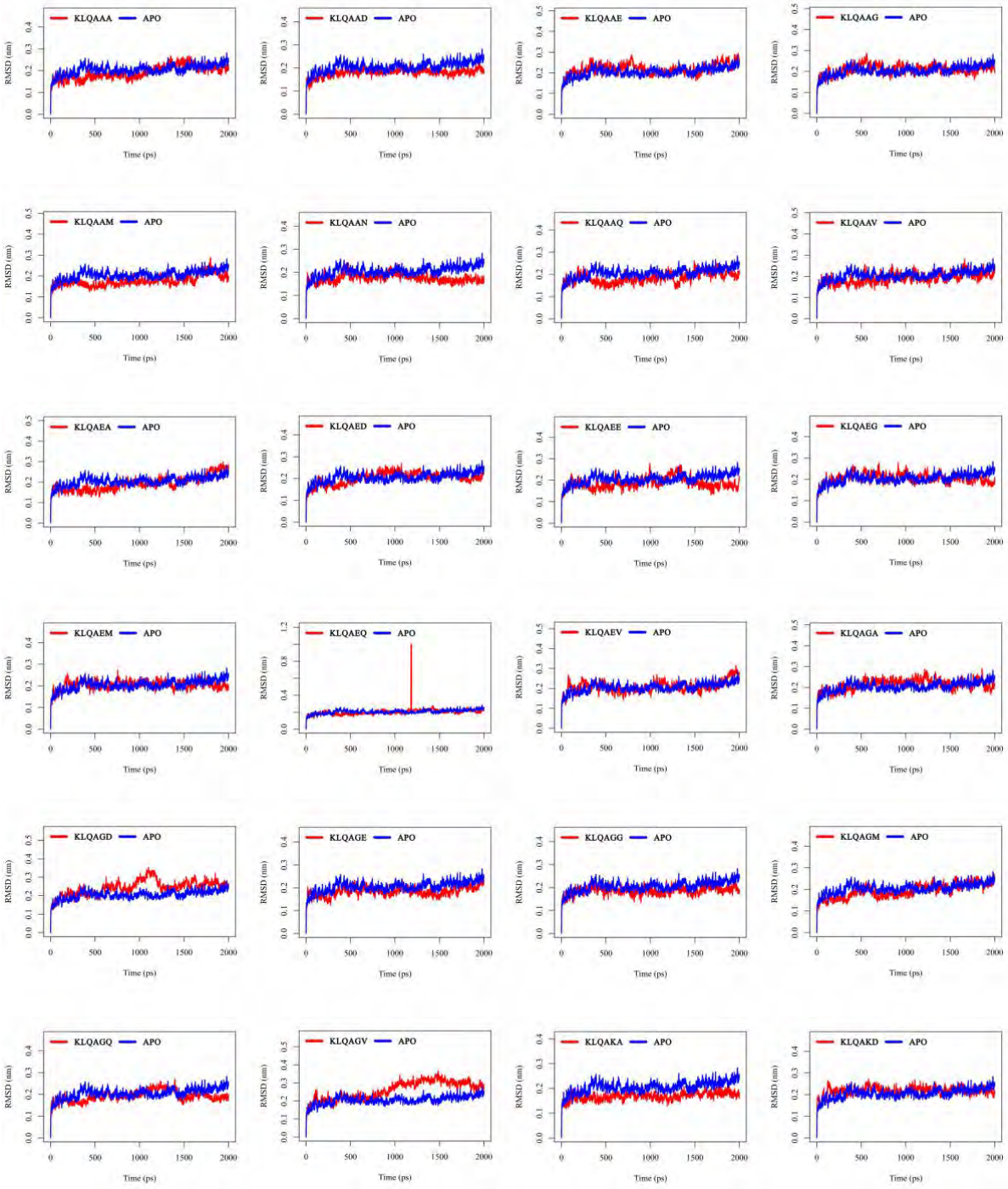
4.4.1.1 RMSD

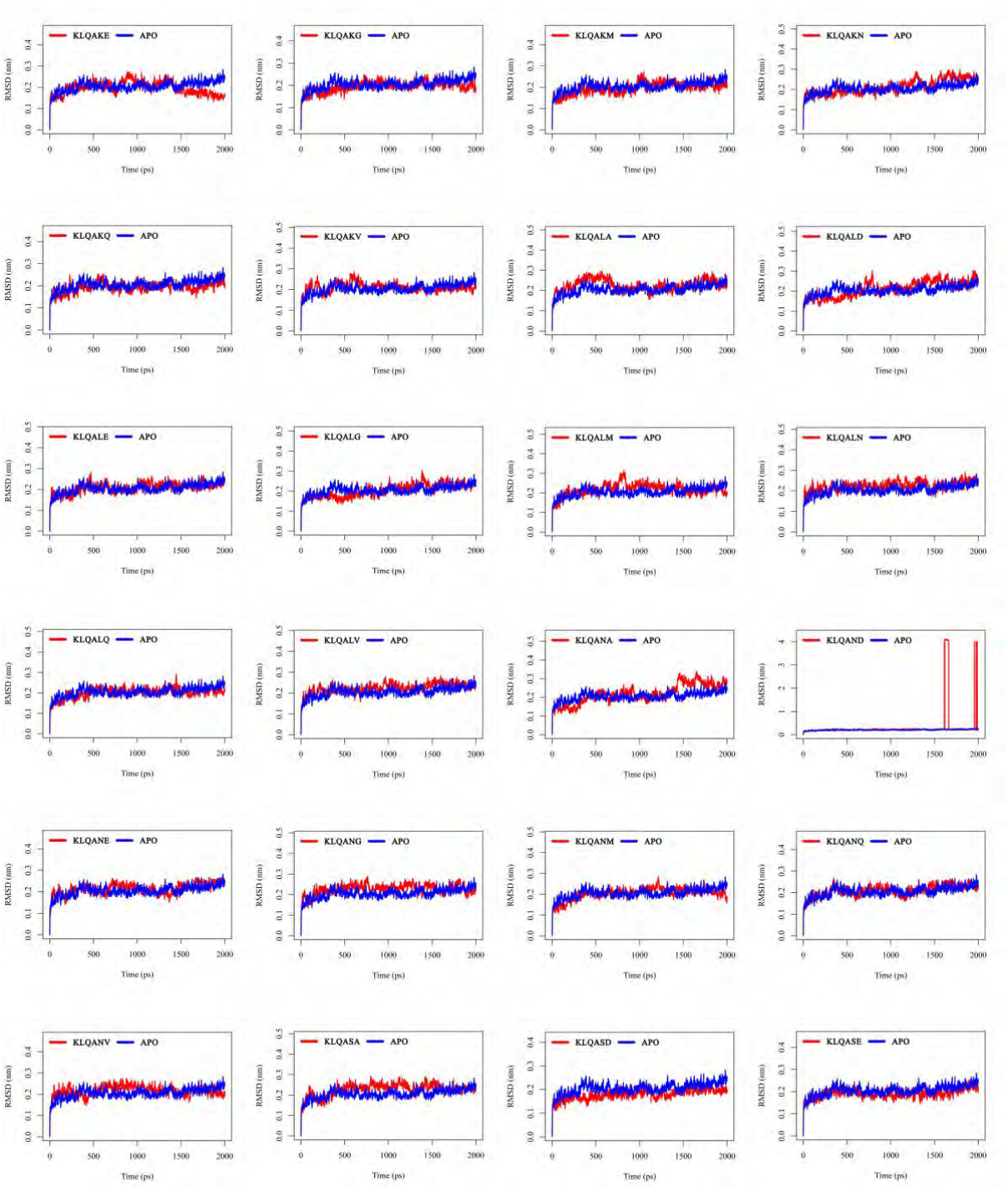
The stability of the M^{pro} systems was assessed through the calculation of RMSD for the protein backbone (α-carbons) from the 20 ns trajectories and plotted against time as shown in figure 4.1. The *apo*-protein retained RMSD values around 0.2 nm, after equilibration, after 400 ps of simulation time. The majority of the complexed systems also equilibrated to an RMSD of around 0.2 nm, showing that the binding of the substrates did not introduce structural changes to the protein, and also that the complexes were stable. However, there were a few complexed systems that equilibrated at a lower RMSD in comparison to the *apo*-protein, with values ranging roughly between 0.10 and 0.22 nm like KLQAAM, KLQAKA, KLQASD, KLQASG, KLQSND and KLQSSN. These complexes indicate the subtle reduction of backbone fluctuation in the protein as a consequence of substrate binding. These lower RMSD values are quite interesting since they show the nuances of binding of a set of substrates that are in essence quite similar in composition.

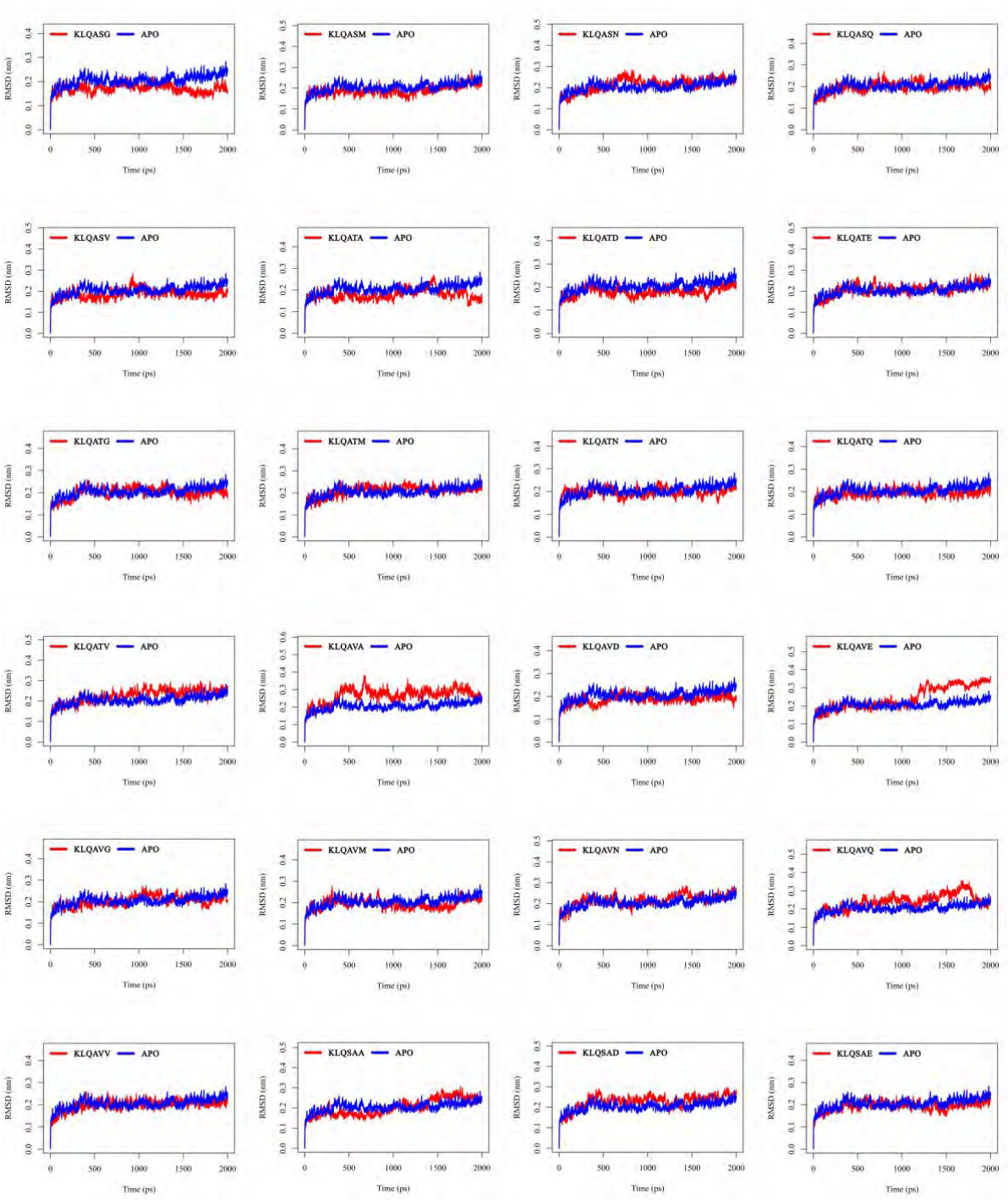
A few complexed systems did attain higher fluctuations, accompanied by steady increases in the RMSD as the simulation progressed (for example, KLQAGD, KLQALA, KLQALD, KLQANA and KLQSEA). Backbone fluctuation was common among these systems, as

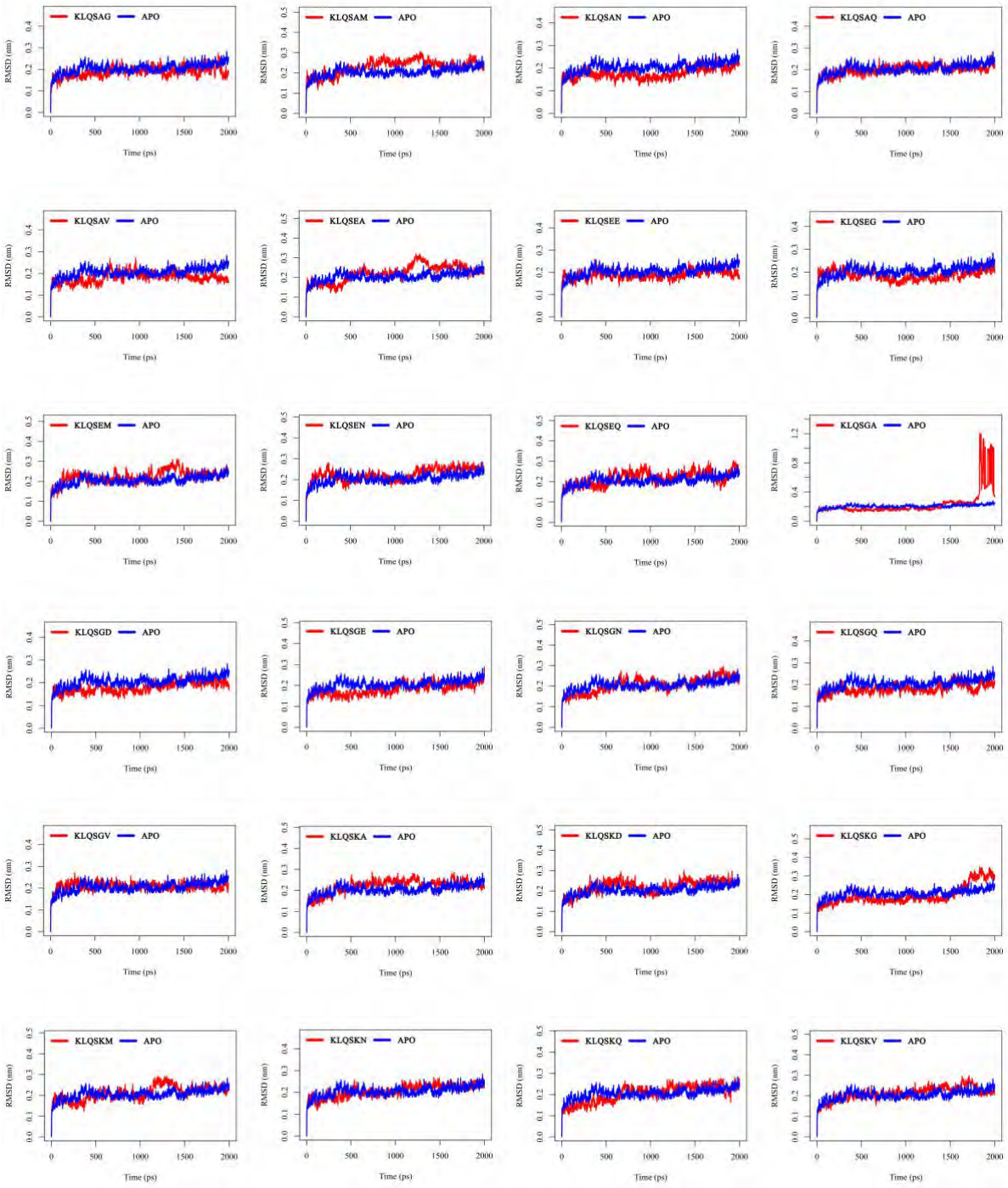
periodic drops and instability in the RMSD were evident. The values of RMSD in some cases would reach as high as 0.36 nm. Some of these systems would retain this higher value for a longer period as when compared to others (indicating a variation in stability at the 0.36 nm RMSD). For instance, systems KLQAGV, KLQAVA, KLQAVE, KLQSKG, KLQSLA and KLQSTN retained RMSD > 0.3 nm for: the final 10 ns; final 15 ns; final 8 ns; final 3 ns; final 5 ns; and between 10 and 18 ns, of the simulations, respectively

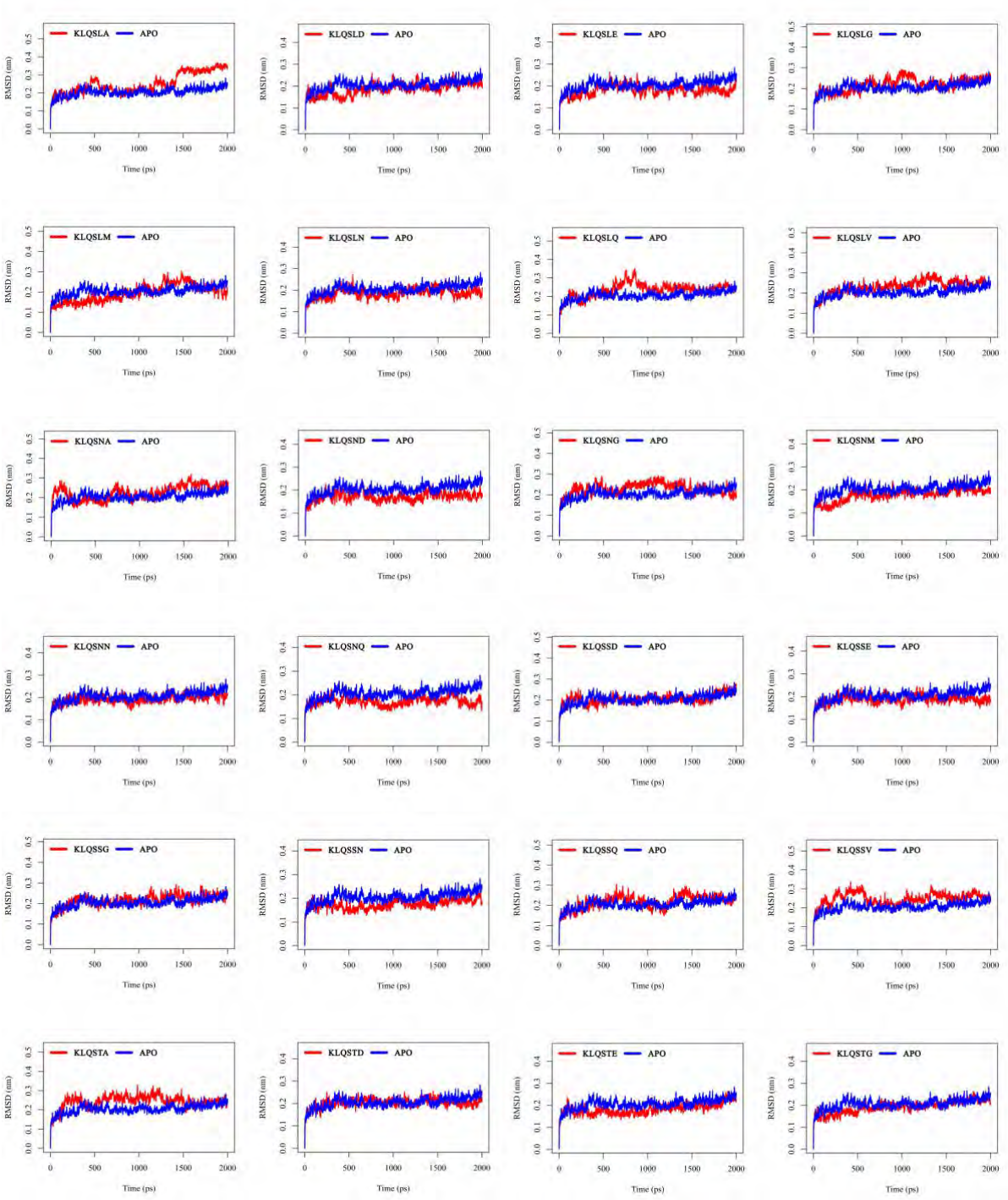
The last group of systems registered significantly steep increases in RMSD at various points of the simulation (KLQAEQ, KLQAND, KLQSGA and KLQSVQ). These spikes were indicative of significant changes in the protein dynamics during the simulations. Bear in mind that the peptide substrate is chain C of the PDB file, and cannot be separated from this RMSD calculation for the protein. Of these, the complex KLQSVQ displayed the highest increase in RMSD, with values reaching 4.20 nm, followed by KLQAND with the highest RMSD reaching 4.13 nm. The increases shown in KLQAEQ and KLQSGA were less drastic when compared to KLQAND and KLQSVQ, reaching peaks of 1.00 and 1.20 nm, respectively. The details of the dynamic events that postulate these steep increases in RMSD for KLQAEQ, KLQSGA and KLQSVQ are discussed below. To summarize, most systems consistently achieved RMSD equilibration around 0.2 nm. With the exception of KLQAEQ, KLQAND, KLQSGA and KLQSVQ, the RMSD values for the systems ranged between 0.1 and 0.38 nm. Similar RMSD values for the M^{pro} *apo*-dimer that fall within this range were reported by Suárez and Díaz (2020), where they sought to inhibit the protein using peptide-based inhibitors.


KLQAAA APO
KLQAAD APO
KLQAAE APO
KLQAAG APO
KLQAAM APO
KLQAAN APO
KLQAAQ APO
KLQAAV APO
KLQAEA APO
KLQAED APO
KLQAEE APO
KLQAEG APO
KLQAEM APO
KLQAEQ APO
KLQAEV APO
KLQAGA APO
KLQAGD APO
KLQAGE APO
KLQAGG APO
KLQAGM APO
KLQAGQ APO
KLQAGV APO
KLQAKA APO
KLQAKD APO
RMSD (nm)
Time (ps)


KLQAKE APO
KLQAKG APO
KLQAKM APO
KLQAKN APO
KLQAKQ APO
KLQAKV APO
KLQALA APO
KLQALD APO
KLQALE APO
KLQALG APO
KLQALM APO
KLQALN APO
KLQALQ APO
KLQALV APO
KLQANA APO
KLQAND APO
KLQANE APO
KLQANG APO
KLQANM APO
KLQANQ APO
KLQANV APO
KLQASA APO
KLQASD APO
KLQASE APO
RMSD (nm)
Time (ps)


KLQASG APO
KLQASM APO
KLQASN APO
KLQASQ APO
KLQASV APO
KLQATA APO
KLQATD APO
KLQATE APO
KLQATG APO
KLQATM APO
KLQATN APO
KLQATQ APO
KLQATV APO
KLQAVA APO
KLQAVD APO
KLQAVE APO
KLQAVG APO
KLQAVM APO
KLQAVN APO
KLQAVQ APO
KLQAVV APO
KLQSAA APO
KLQSAD APO
KLQSAE APO
RMSD (nm)
Time (ps)


KLQSAG APO
KLQSAM APO
KLQSAN APO
KLQSAQ APO
KLQSAV APO
KLQSEA APO
KLQSEE APO
KLQSEG APO
KLQSEM APO
KLQSEN APO
KLQSEQ APO
KLQSGA APO
KLQSGD APO
KLQSGE APO
KLQSGN APO
KLQSGQ APO
KLQSGV APO
KLQSKA APO
KLQSKD APO
KLQSKG APO
KLQSKM APO
KLQSKN APO
KLQSKQ APO
KLQSKV APO
RMSD (nm)
Time (ps)


KLQSLA APO
KLQSLD APO
KLQSLE APO
KLQSLG APO
KLQSLM APO
KLQSLN APO
KLQSLQ APO
KLQSLV APO
KLQSNA APO
KLQSND APO
KLQSNG APO
KLQSNM APO
KLQSNN APO
KLQSNQ APO
KLQSSD APO
KLQSSE APO
KLQSSG APO
KLQSSN APO
KLQSSQ APO
KLQSSV APO
KLQSTA APO
KLQSTD APO
KLQSTE APO
KLQSTG APO
RMSD (nm)
Time (ps)

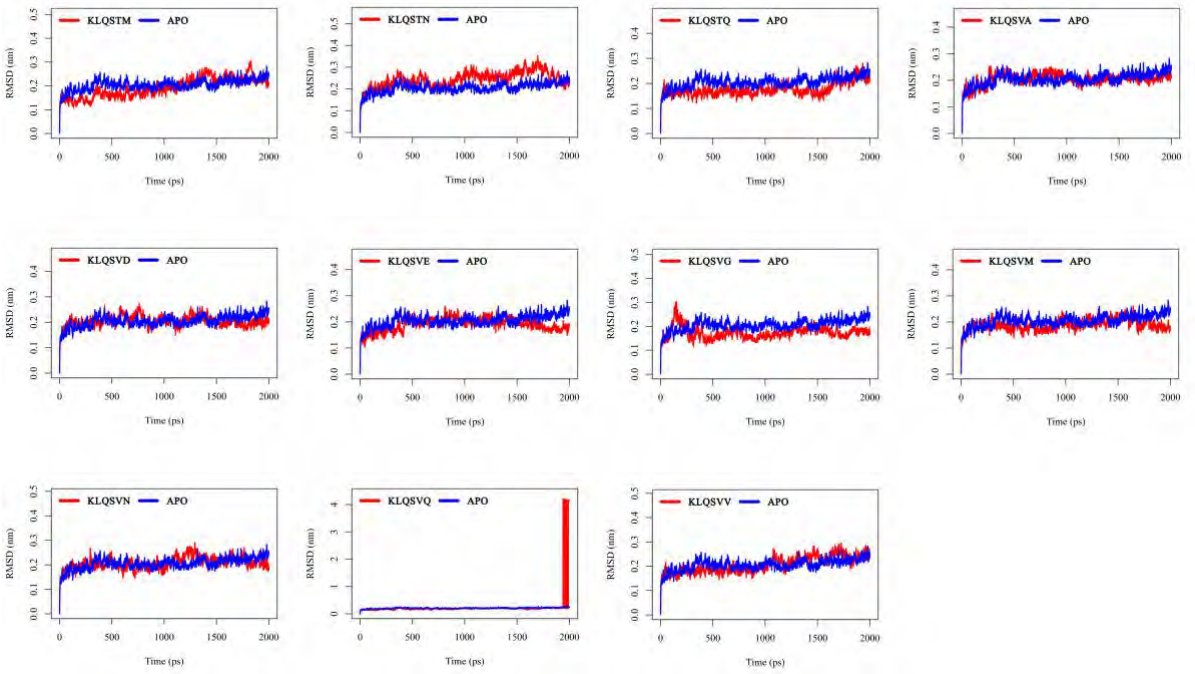


Figure 4.1. The global stability of the M^{pro} and M^{pro}-Hexapeptide complexes. RMSD of the backbone α-carbon atoms for the *apo*-protein and KLQ*** hexapeptide bound M^{pro} systems during the 20 ns MD simulation. Plots were created using RStudio.

4.4.1.2 Rg

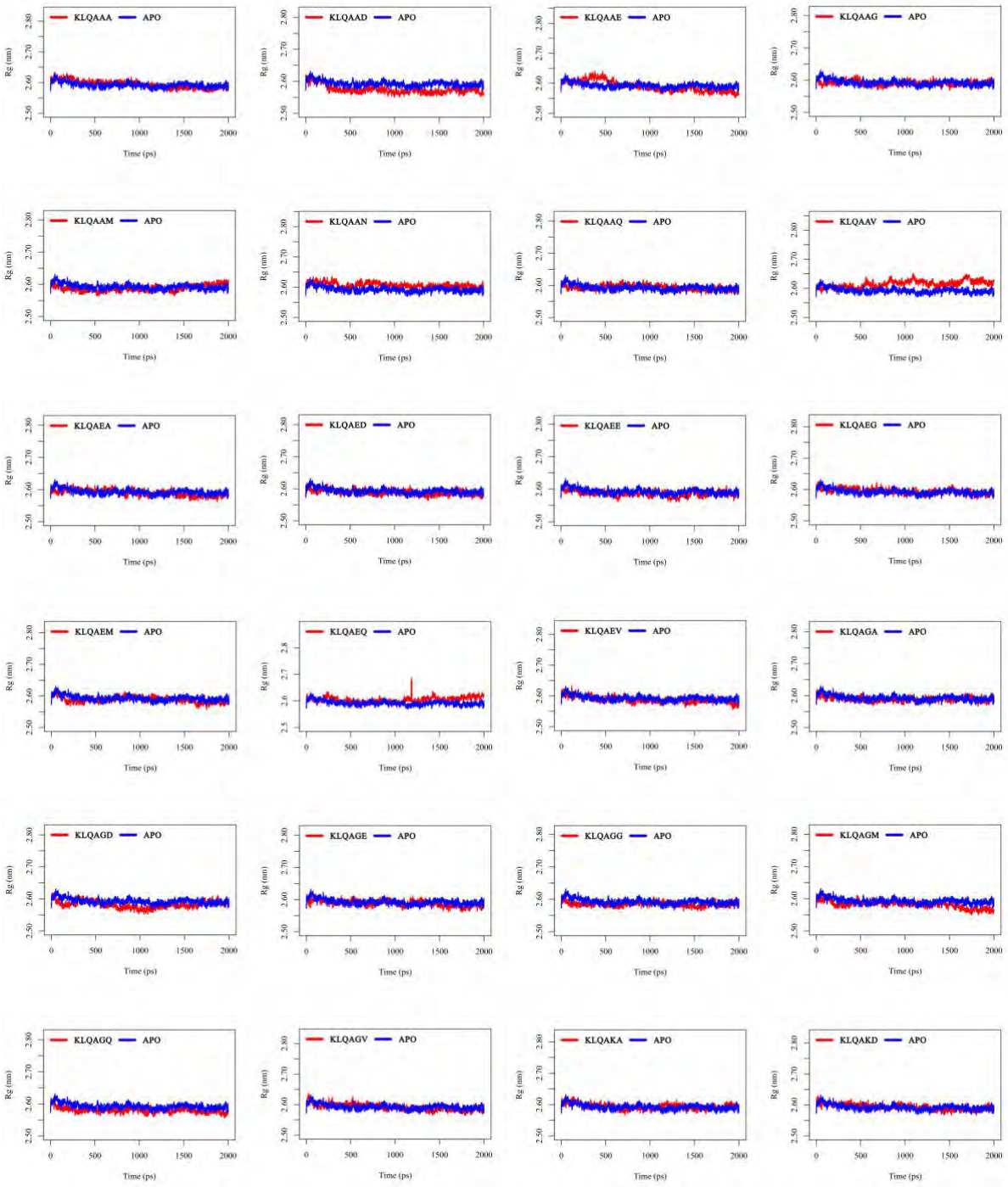
The degree of compactness and folding of the M^{pro} systems was monitored through the Rg plotted against time as shown in figure 4.2. Rg was used to monitor changes in protein structure with respect to its native state, thus relaying information about the folding and unfolding of the M^{pro} structure during the 20 ns simulations.

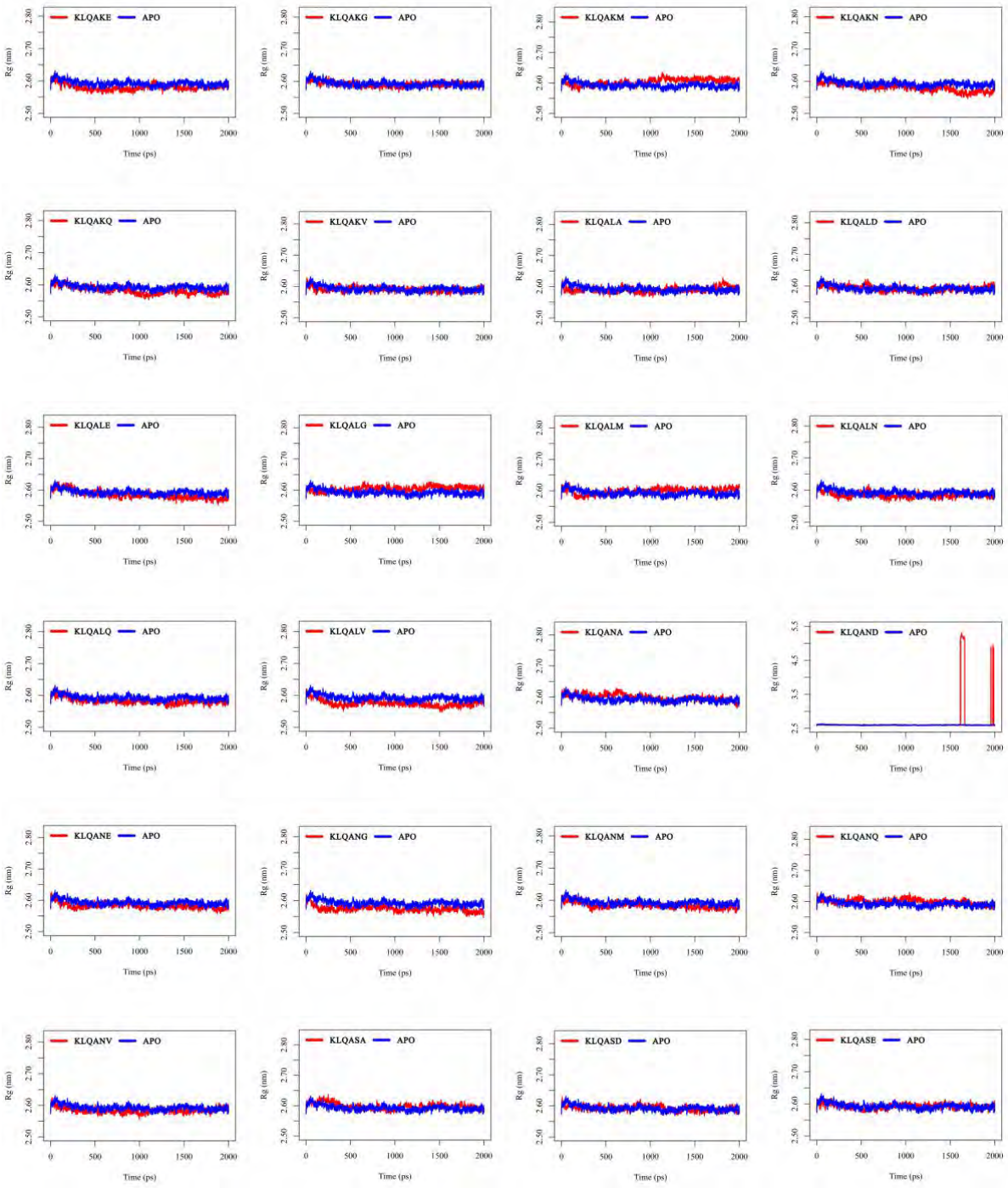
The majority of the systems retained similar degrees of compaction, as shown by the equilibration of the Rg around 2.60 nm throughout the simulation. The *apo*-M^{pro} was also among these systems. In the context of the complexed systems, this trend indicates that the effect of substrate binding does not confer instability in protein structure; further pointing to the substrate specificity of M^{pro} for this set of KLQ*** hexapeptides. This is also indicated in the observed proper substrate binding (chapter 3, section 3.6.7.1), and the prevalence of stabilising intermolecular interactions (chapter 3, section 3.6.7.2) known to confer strength and stability in complexes.

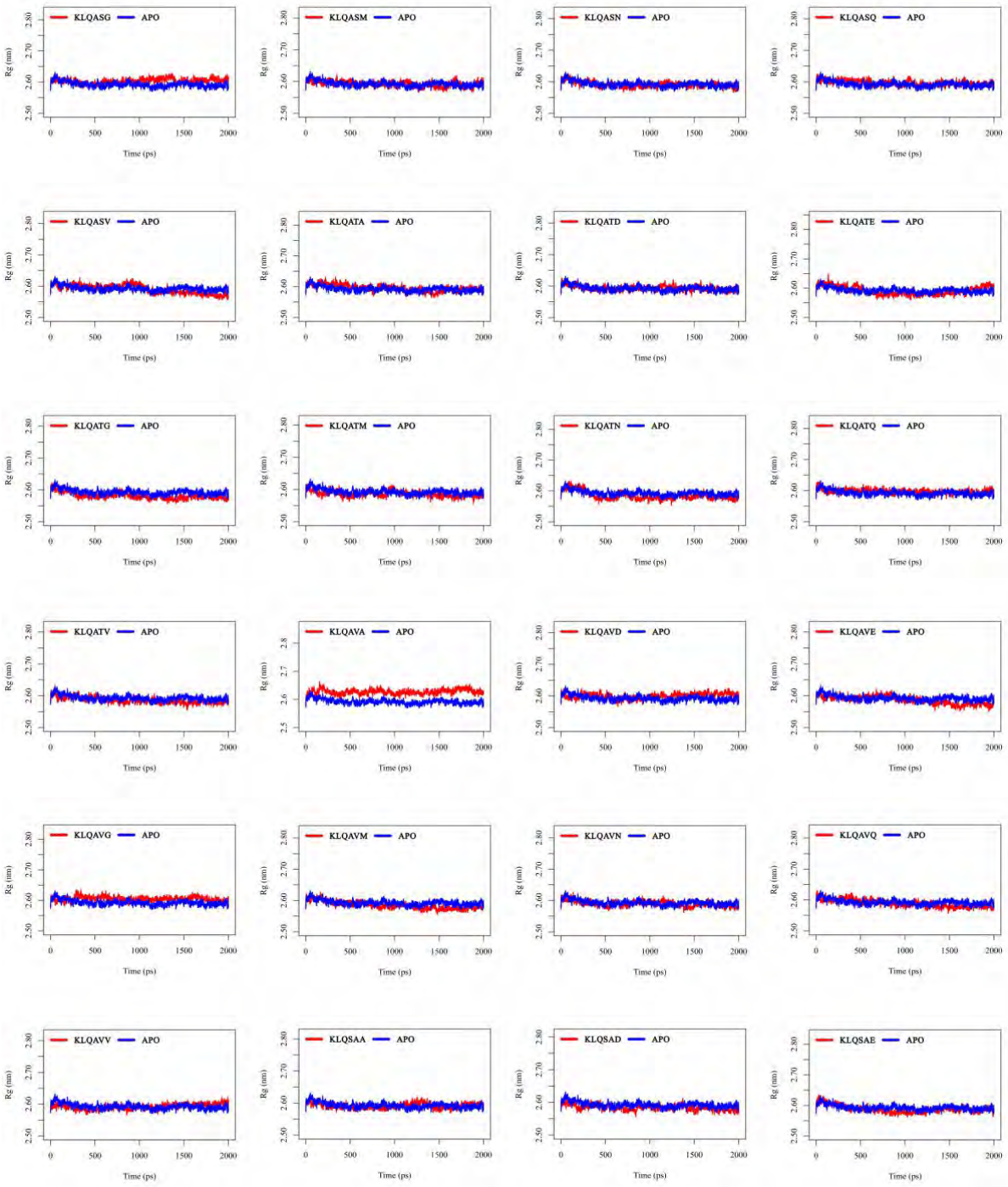
Interestingly, some complexed systems displayed steady drops in Rg as the simulation progressed (KLQAAA, KLQAEV, KLQAGM, KLQAGQ, KLQALE, KLQAKN, KLQATG, KLQATN, KLQATV, KLQAVN, KLQSAE, KLQSAM, KLQSEA, KLQSEM, KLQSEN, KLQSLA, and KLQSNM). The decreases in Rg values never fell below 2.55 nm. In these systems, the binding of substrates introduced, increased the compaction of the protein

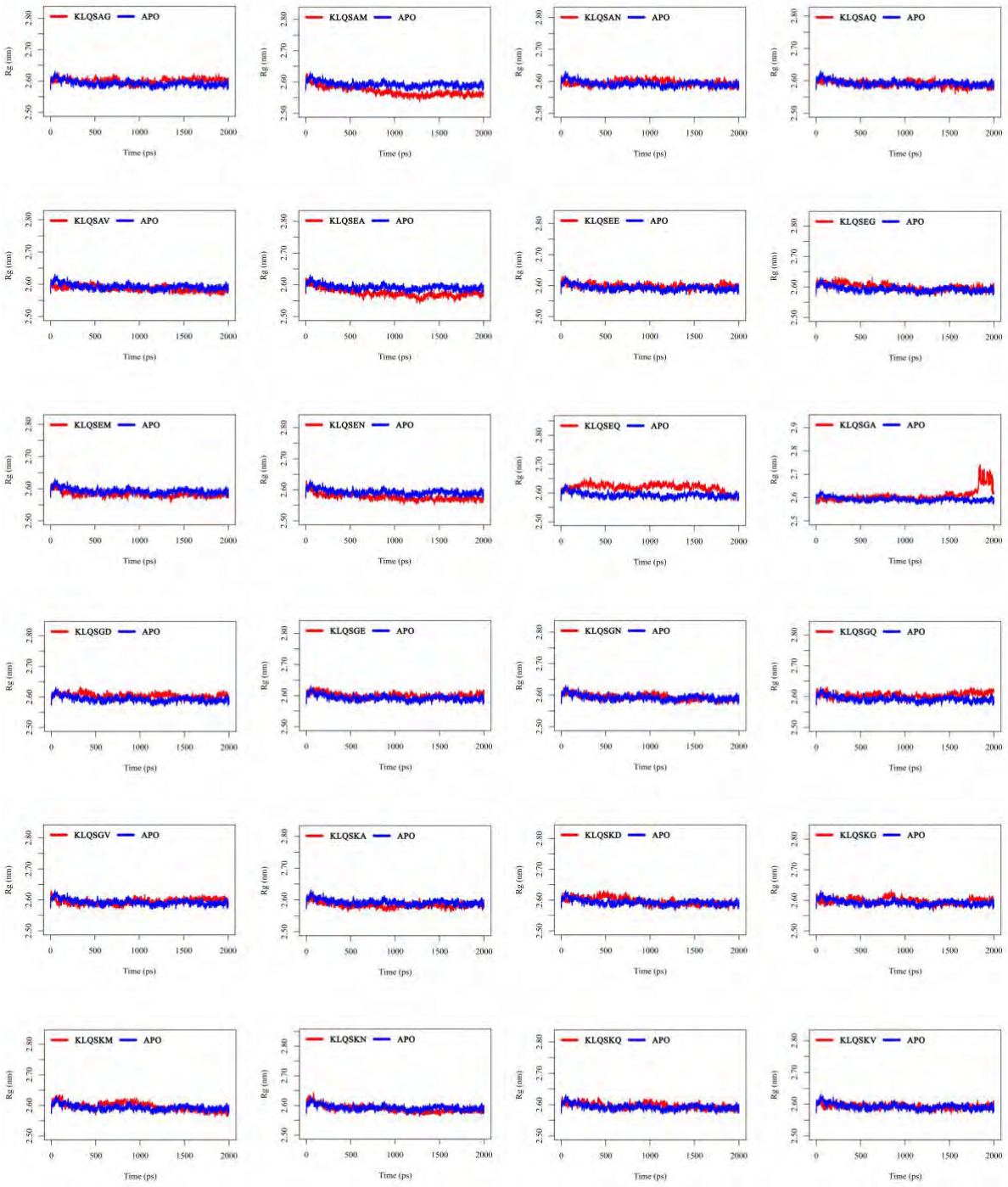
structure, and resulted in a slightly greater contraction of the M^{pro}-hexapeptide structure. Similarly, some of the systems displayed steady increases in Rg (from 2.6 nm) as the simulation progressed, without exceeding 2.65 nm (KLQAAV, KLQAKM, KLQSGV and KLQSLV). Structural compaction was reduced in the systems, showing minor flexibility in the M^{pro}-hexapeptide structure as the simulation progressed. There were also a few systems that consistently attained and retained Rg equilibration above (KLQAAN, KLQANA, KLQAVA and KLQSEQ), and below (KLQAAD, KLQAAE, KLQAEM, KLQSLE and KLQSLG) the typical 2.6 nm point of equilibration. The overall Rg values of these aforementioned systems roughly ranged between 2.57 and 2.62 nm. Notably, the reference state for these systems had an Rg of 2.57 nm. Thus, M^{pro}-hexapeptide and *apo*-M^{pro} systems retained close levels of flexibility to their native states.

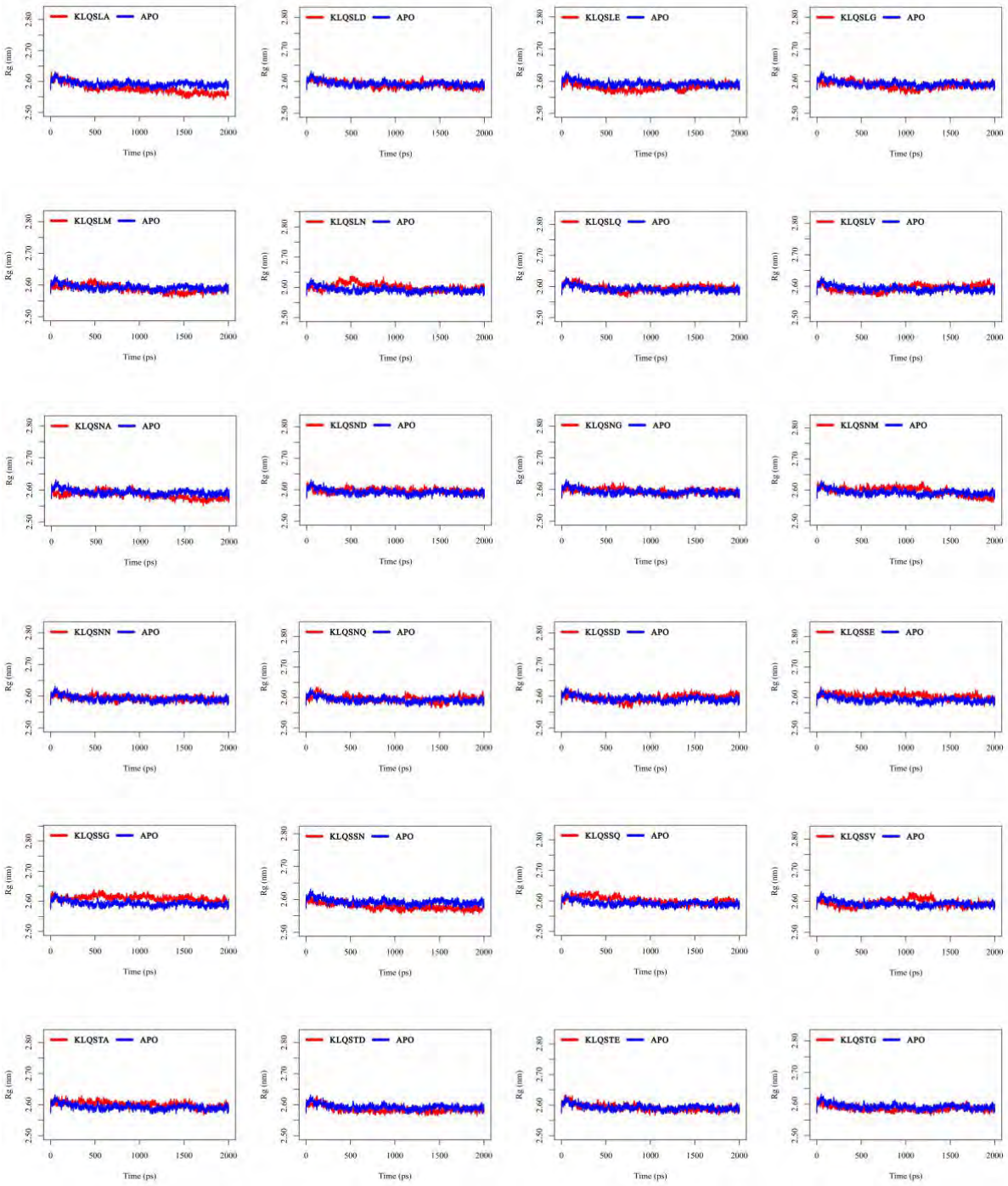
As also seen in the RMSD, the systems KLQAEQ, KLQAND, KLQSGA and KLQSVQ displayed steep hikes in Rg at various times of their simulations. KLQSVQ and KLQAND achieved the steepest hikes, reaching maximum values of approximately 5.40 and 5.31 nm, respectively. As with RSMD, KLQAEQ and KLQSGA attained less drastic increases in Rg than KLQSVQ or KLQAND, with approximate peaks of 2.67 and 2.74 nm, respectively. Fascinatingly, these peaks in Rg corresponded to the peaks of RMSD with regards to scale and the timestamps of the simulation period, suggesting that the sudden increase in backbone fluctuations correlated with the sudden increases in structural flexibility. Outside these timestamps, these systems display the typical hexapeptide-protein behaviour demonstrated by the majority of the systems simulated. Curiously, a look into the binding modes of substrates reveals that all were bound onto the M^{pro} active site appropriately and in accordance with the nomenclature of Schechter and Berger (1967).


KLQAAA APO
KLQAAD APO
KLQAAE APO
KLQAAG APO
KLQAAM APO
KLQAAN APO
KLQAAQ APO
KLQAAV APO
KLQAEA APO
KLQAED APO
KLQAEE APO
KLQAEG APO
KLQAEM APO
KLQAEQ APO
KLQAEV APO
KLQAGA APO
KLQAGD APO
KLQAGE APO
KLQAGG APO
KLQAGM APO
KLQAGQ APO
KLQAGV APO
KLQAKA APO
KLQAKD APO
Rg (nm)
Time (ps)


KLQAKE APO
KLQAKG APO
KLQAKM APO
KLQAKN APO
KLQAKQ APO
KLQAKV APO
KLQALA APO
KLQALD APO
KLQALE APO
KLQALG APO
KLQALM APO
KLQALN APO
KLQALQ APO
KLQALV APO
KLQANA APO
KLQAND APO
KLQANE APO
KLQANG APO
KLQANM APO
KLQANQ APO
KLQANV APO
KLQASA APO
KLQASD APO
KLQASE APO
Rg (nm)
Time (ps)


KLQASG APO
KLQASM APO
KLQASN APO
KLQASQ APO
KLQASV APO
KLQATA APO
KLQATD APO
KLQATE APO
KLQATG APO
KLQATM APO
KLQATN APO
KLQATQ APO
KLQATV APO
KLQAVA APO
KLQAVD APO
KLQAVE APO
KLQAVG APO
KLQAVM APO
KLQAVN APO
KLQAVQ APO
KLQAVV APO
KLQSAA APO
KLQSAD APO
KLQSAE APO
Rg (nm)
Time (ps)


KLQSAG APO
KLQSAM APO
KLQSAN APO
KLQSAQ APO
KLQSAV APO
KLQSEA APO
KLQSEE APO
KLQSEG APO
KLQSEM APO
KLQSEN APO
KLQSEQ APO
KLQSGA APO
KLQSGD APO
KLQSGE APO
KLQSGN APO
KLQSGQ APO
KLQSGV APO
KLQSKA APO
KLQSKD APO
KLQSKG APO
KLQSKM APO
KLQSKN APO
KLQSKQ APO
KLQSKV APO
Rg (nm)
Time (ps)


KLQSLA APO
KLQSLD APO
KLQSLE APO
KLQSLG APO
KLQSLM APO
KLQSLN APO
KLQSLQ APO
KLQSLV APO
KLQSNA APO
KLQSND APO
KLQSNG APO
KLQSNM APO
KLQSNN APO
KLQSNQ APO
KLQSSD APO
KLQSSE APO
KLQSSG APO
KLQSSN APO
KLQSSQ APO
KLQSSV APO
KLQSTA APO
KLQSTD APO
KLQSTE APO
KLQSTG APO
Rg (nm)
Time (ps)

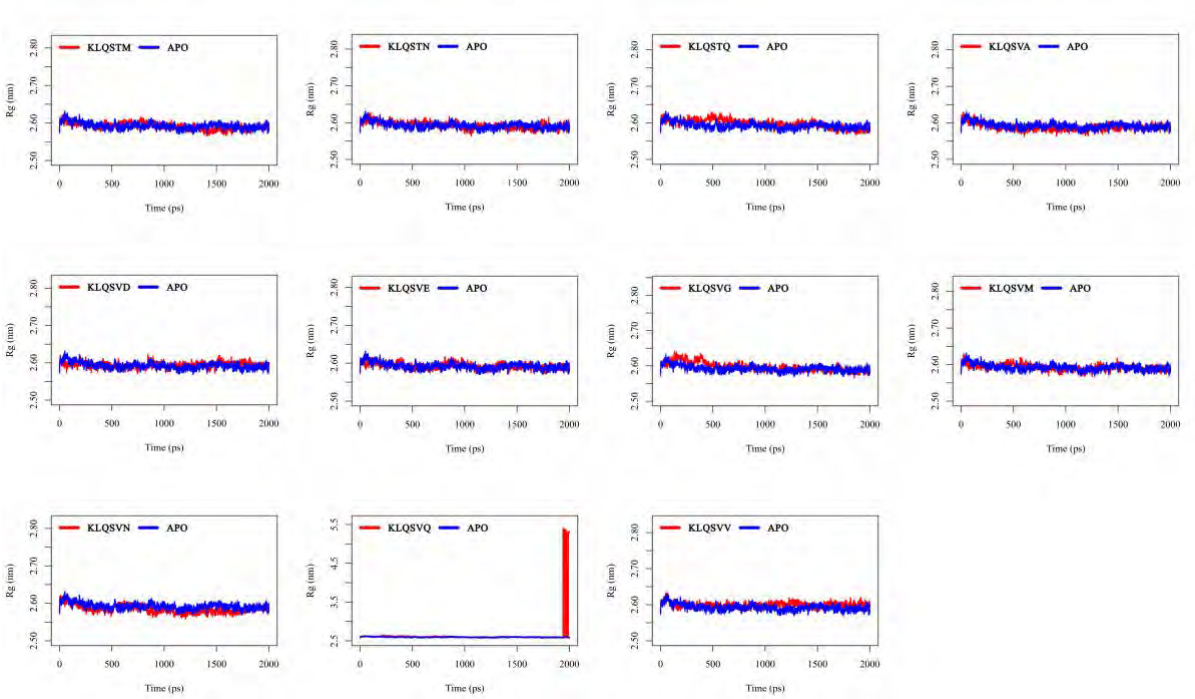


Figure 4.2. The global stability of the M^{pro} and M^{pro}-Hexapeptide complexes. Rg of the backbone α-carbon atoms for the *apo*-protein and KLQ*** hexapeptide bound M^{pro} systems during the 20 ns MD simulation. Plots were created using RStudio.

As a conclusion, the RMSD and Rg of these systems, except for KLQAEQ, KLQAND, KLQSGA and KLQSVQ, displayed no indication of destabilising effect of the KLQ*** hexapeptides on the M^{pro} structure and further confirmed the stabilising force and strength of the intermolecular interactions in the substrate binding interface. At present, there is little in the literature surrounding the behaviour during the dynamics of these and similar systems. However, interestingly, Gupta *et al.* (2020) reported RMSD and Rg values for SARS-CoV-2 M^{pro} within the same range for an inhibitory study using natural compounds. In contrast to this study, they performed slightly longer 30 ns MD simulations

4.4.2 LOCAL STRUCTURAL STABILITY OF THE SARS-COV-2 M^{PRO}

4.4.2.1 RMSF

Local chain fluctuations of the M^{pro} were measured by calculating the RMSF and assessed using heatmaps. Heatmaps allowed the identification of high-flexibility regions, which were subsequently mapped on the M^{pro} crystal structure to reveal the positions of these regions within the 3D protein structure.

Across all systems, the RMSF of chain A approximated the RMSF of chain B, despite the fact only chain B had a bound substrate in complex systems (figures 4.5, 4.7 and 4.9). Moreover,

slightly higher RMSF values were registered in chain A. Overall, the highest RMSF values were obtained in systems KLQAND and KLQSVQ (figure 4.9 & 4.10). Since these values were disproportionately higher than the rest of the systems, the data were separated to optimise visualisation as follows: i) systems with KLQ*** substrates with Ala at P1′ (figures 4.5 & 4.6); ii) systems KLQ*** substrates with Ser at P1′ (figure 4.7 & 4.8); iii) systems with KLQAND and KLQSVQ (figures 4.9 & 4.10).

Figure 4.5 shows the heatmaps for both chains of KLQA** systems alongside the *apo*-M^{pro}. While values vary from one system to the next, both chains of the M^{pro} demonstrated similar values and high-flexible regions in each system. Overall, the RMSF values for KLQA** systems ranged between 0.0384-0.5805 nm for chain A and 0.0394-0.5698 nm for chain B. High flexibility was observed in residues 21-26, 44-80, 92-97, 118-127, 141-144, 152-156, 167-171, 188-198, 215-288, and 298-302 in chain A (figure 4.5A); and residues 1-4, 22-24, 44-80, 92-96, 118-125, 153-156, 168-171, 188-197, 212-288, and 297-301 in chain B (figure 4.5B), respectively.

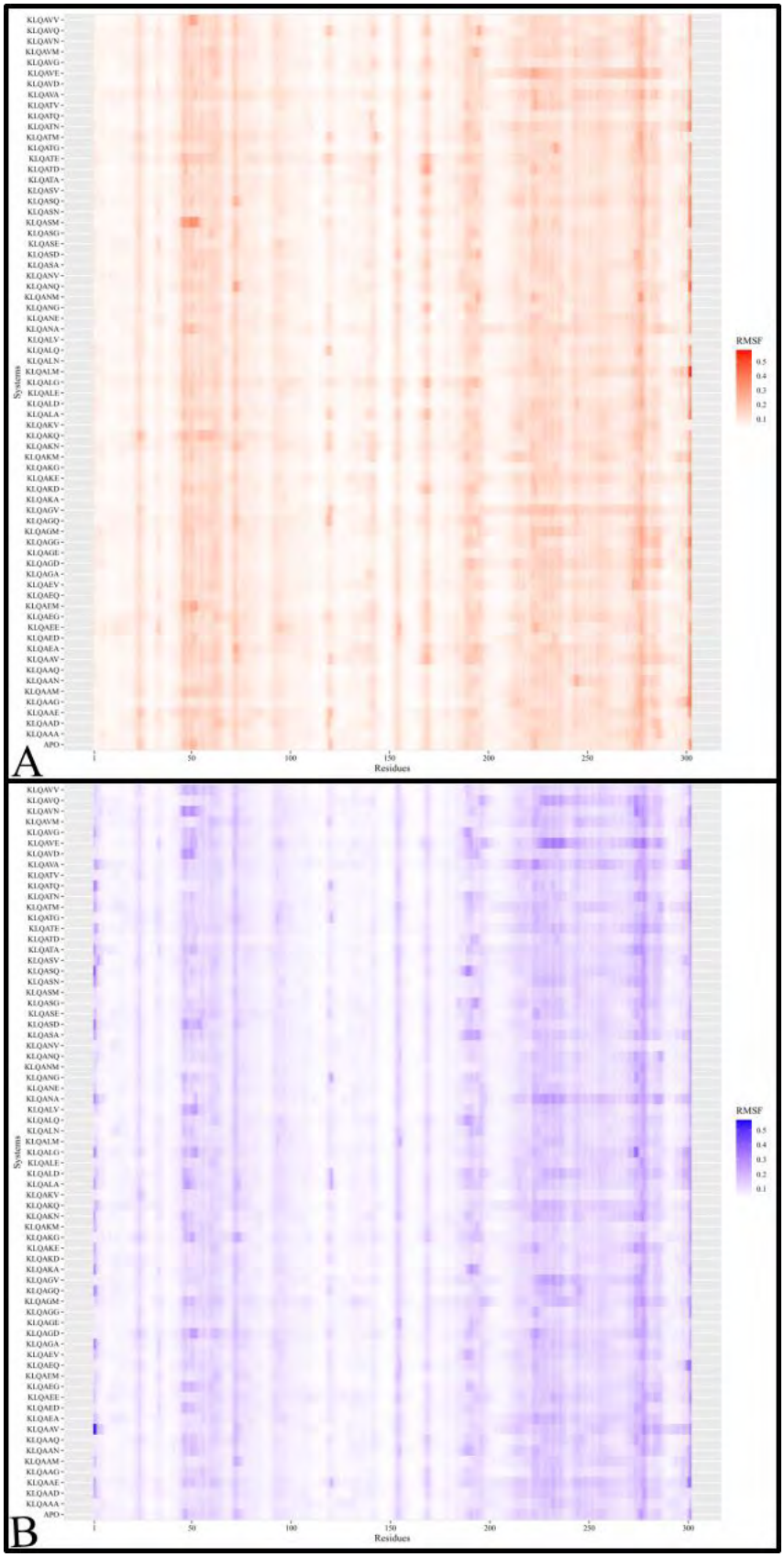


Figure 4.5. The local stability of the M^{pro} and M^{pro}-KLQA Hexapeptide complexes**. RMSF of the backbone α-carbon atoms for the *apo*-protein and KLQA** hexapeptide bound M^{pro} systems during the 20 ns MD simulation. **A** and **B** refer to the separate chains of the M^{pro} homodimer. Images were created using RStudio.

The majority of the highly fluctuating residues constituted loop regions in both monomers of

the M^{pro}, as shown (as red, and blue) in figure 4.6. Loop regions are highly flexible structures in their native state. Of all the flexible loop regions, the residues 1-4 displayed flexibility exclusively in chain B. Residues 1-9 are known to form the N-finger terminal region which plays a crucial role in dimer formation through interactions with Domain II of chain A (Sheik Amamuddy *et al.*, 2020). Semi-flexibility in β-sheets was displayed on either end of the structure in both monomers, connecting to or from loop regions. These β-sheets constituted Domains I and II. Additionally, there were α-helices displaying semi-flexibility, comprising of residues 44-80 (domain I) and 212-288 (domain III). Residues 44-80 are part of the catalytic domain which is responsible for catalysis and M^{pro} autocleavage (Mengist *et al.*, 2021). Sequentially, this α-helix comes after the catalytic His41 and is comprised of key residues like Met49 which contribute to substrate stabilisation. Considering the RMSD and Rg of these systems, this apparent semi-flexibility is not indicative of instability in the binding pocket, but instead shows functional flexibility that accommodates the bound substrate. The α-helices of Domain III (figure 4.6 lower region of protein) consistently displayed high flexibility in both monomers. Curiously, α-helices typically demonstrate restricted motion but can confer great flexibility that is essential to protein function (Skipper, 2005). However, since the helices are connected by long loop chains, fluctuation/deviation is highly likely to be present.

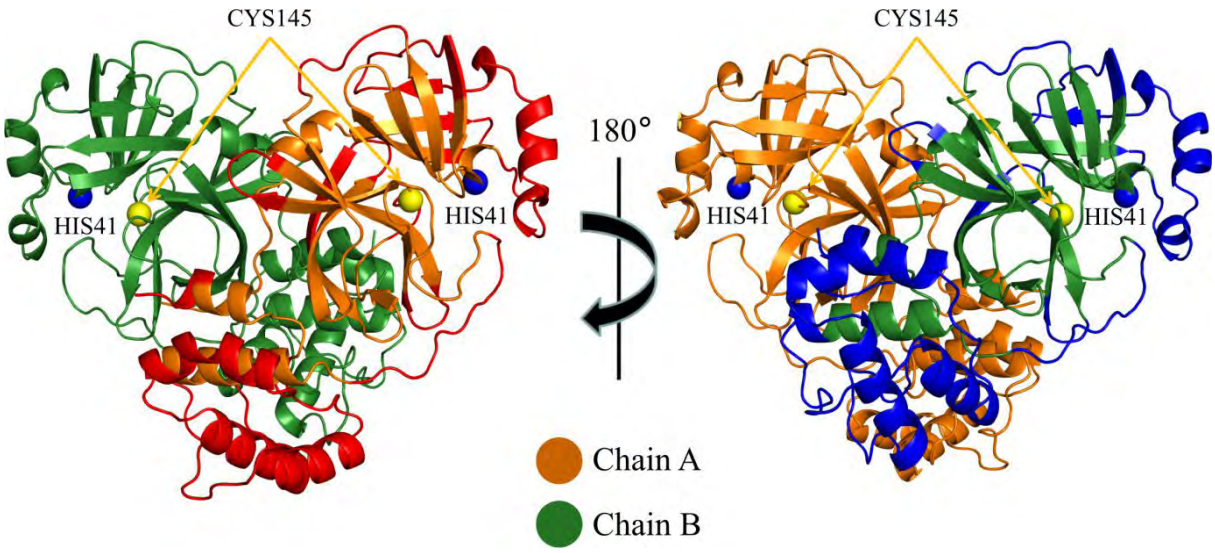


Figure 4.6. The localisation of high-fluctuation residues of the M^{pro} in *apo*- and KLQA Hexapeptide bound systems**. The 3D structure of the M^{pro} show chain A and chain B in orange and green, respectively. Residues displaying high fluctuations are shown in red in chain A, and in blue in chain B. Catalytic His41 is shown as blue spheres and catalytic Cys145 is shown as yellow spheres on each chain. Images were generated using PyMOL.

The KLQS** systems showed similar trends to the KLQA** systems in terms of RMSF and the localisation of flexible residues. Much like KLQA** systems, the RMSF of chain A were

similar to the RMSF of chain B (figure 4.7). The value for RMSF ranged between 0.0399 - 0.4223 nm for chain A and 0.0372-0.3726 nm for chain B. Similarly, flexibility was demonstrated in residues 21-26, 32-35, 44-80, 92-97, 119-123, 139-143, 153-156, 167-171, 187-197, 212-288, and 297-302 in chain A; and residues 1-5, 21-26, 33-35, 44-66, 70-80, 92-98, 152-156, 167-170, 186-197, 212-238, 241-286, and 297-301 for chain B. Flexibility in N-finger terminal residues was also exclusive to chain B (figures 4.7B & 4.8).

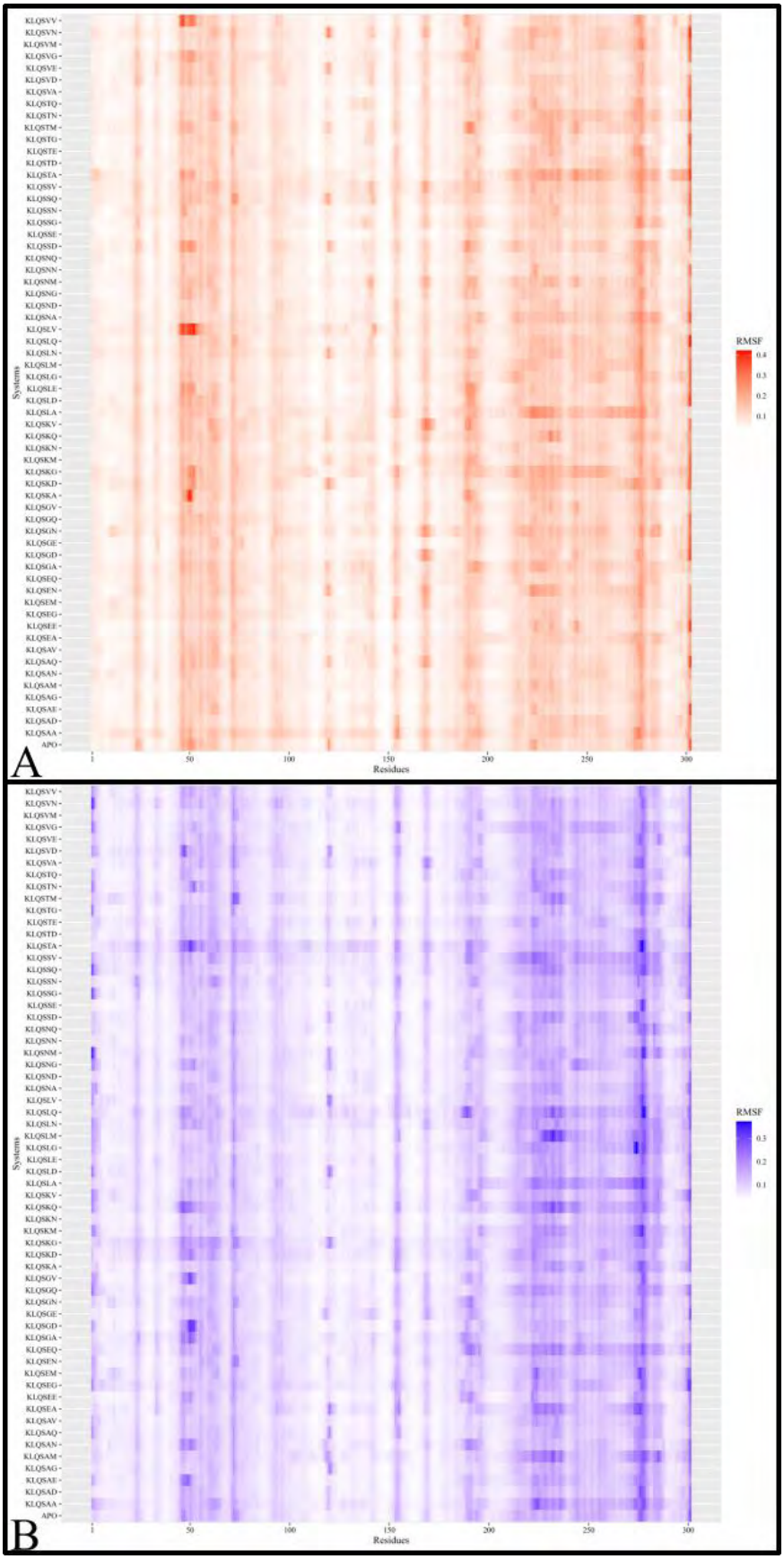


Figure 4.7. The local stability of the M^{pro} and M^{pro}-KLQS Hexapeptide complexes**. RMSF of the backbone α-carbon atoms for the *apo*-protein and KLQS** hexapeptide bound M^{pro} systems during the 20 ns MD simulation. **A** and **B** refer to the chains of the M^{pro} homodimer. Images were created using RStudio.

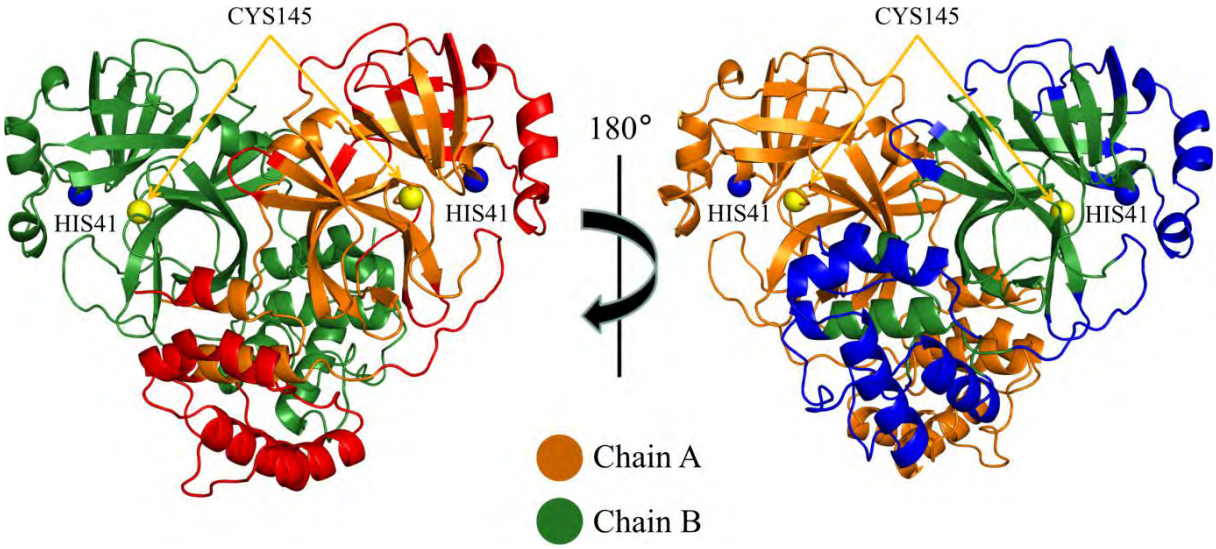


Figure 4.8. The localisation of high-fluctuation residues of the M^{pro} in *apo*- and KLQS Hexapeptide bound systems**. The 3D structure of the M^{pro} show chain A and chain B in orange and green, respectively. Residues displaying high fluctuations are shown in red in chain A, and in blue in chain B. Catalytic His41 is shown as blue spheres and catalytic Cys145 is shown as yellow spheres on each chain. Images were generated using PyMOL.

In the context of the two KLQAND and KLQSVQ systems, residue fluctuation was the highest across all systems with values ranging between 0.2261-1.1565 nm for chain A, and 0.2174-1.1333 nm for chain B. The flexible residues were similar to those in the KLQA** and KLQS** systems, and included residues 1-17, 69-73, 96-100, 111-127, 138-144, 152-157, 202-209, 210-223, 224-234, 236-237, 242-254, 255-259, 260-276, 277-285, 286-298, and 299-302 in chain A; and residues 1-19, 24-29, 69-74, 95-100, 111-128, 138-143, 151-157, 170-173, 199-206, 207-223, 224-227, 247-288, 291-299, and 300-301 in chain B. These were the only instances where the N-finger terminal residues displayed flexibility in both chains (figure 4.9 & 4.10).

Furthermore, the localisation of the flexible residues showed more residue fluctuation in Domain II involving β-sheets; this was not evident in all other systems. The α-helix semi-flexibility (residues 44-63) around the catalytic dyad was also not shown in these systems, pointing towards potential inactivity of the active site residues. Further, α-helices residues constituting Domain III demonstrated greater flexibility than any of the aforementioned systems.

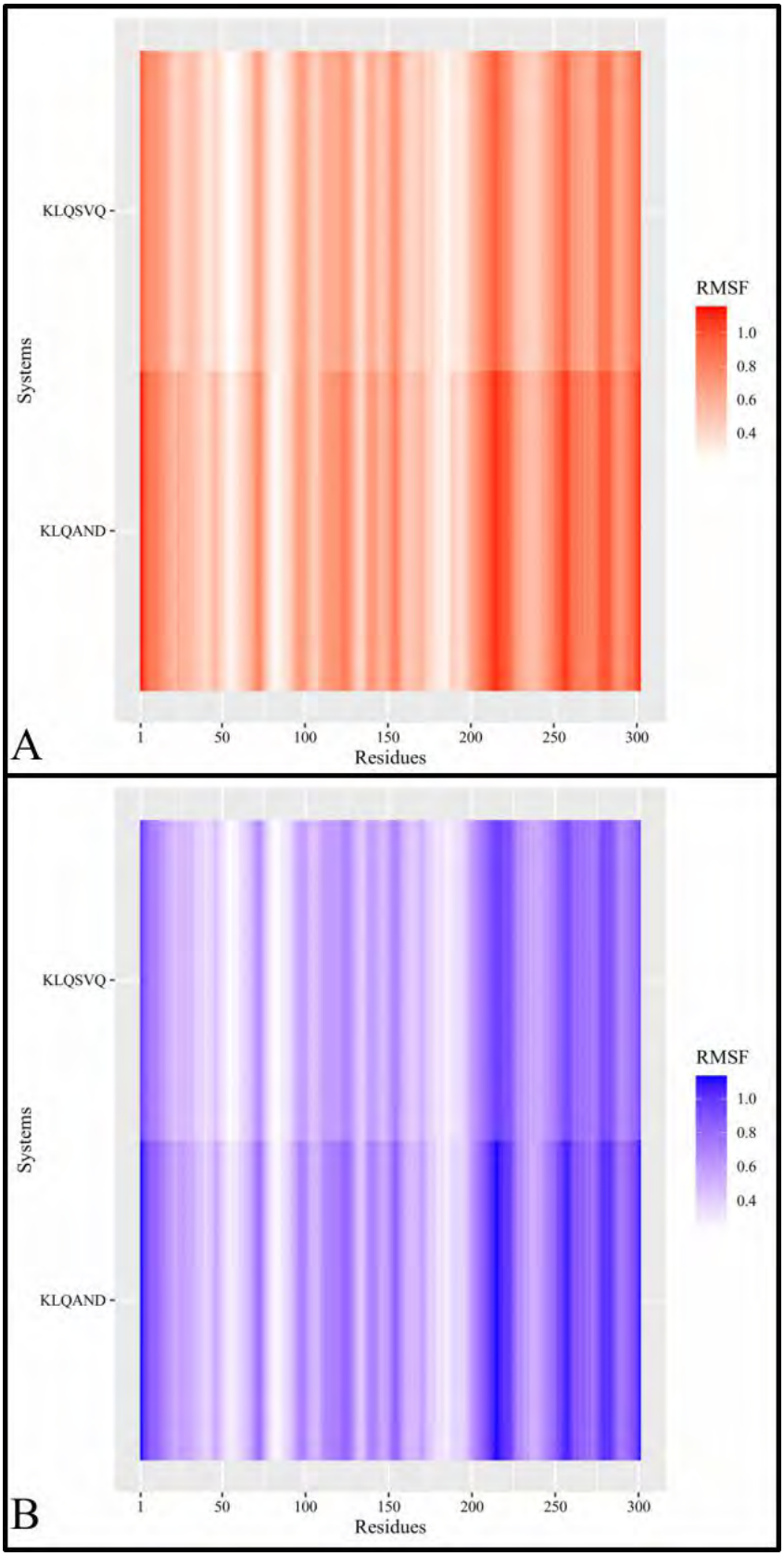


Figure 4.9. The local stability of the M^{pro} and M^{pro}-KLQ* Hexapeptide complexes**. RMSF of the backbone α-carbon atoms for the *apo*-protein, KLQAND and KLQSVQ hexapeptide bound M^{pro} systems during the 20 ns MD simulation. **A** and **B** refer to the chains of the M^{pro} homodimer. Images were created using RStudio.

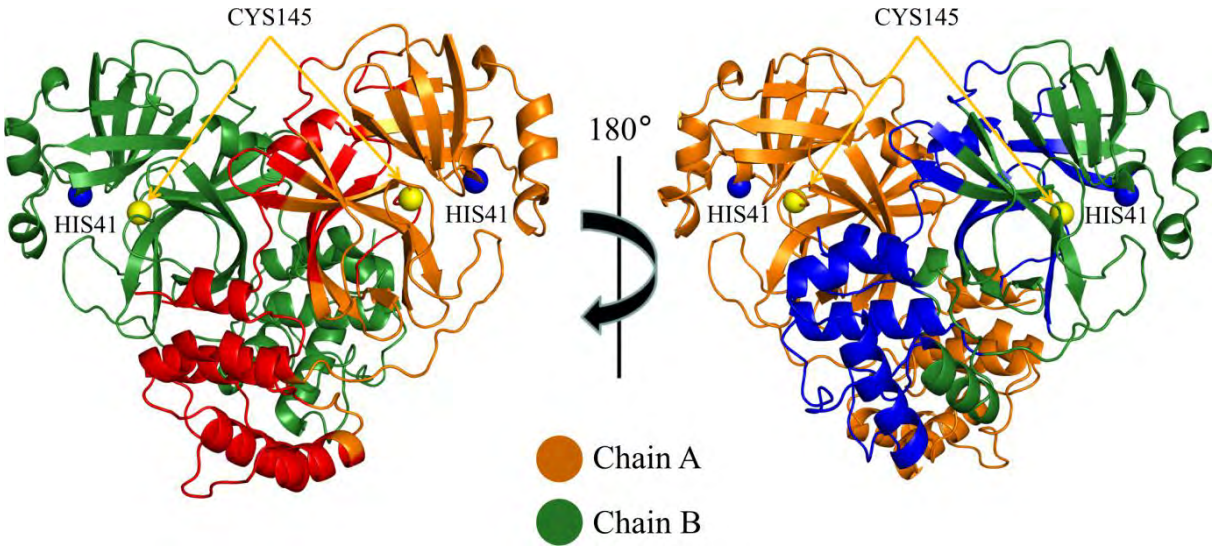


Figure 4.10. The localisation of high-fluctuation residues of the M^{pro} in *apo*-, KLQAND and KLQSVQ Hexapeptide bound systems. The 3D structure of the M^{pro} show chain A and chain B in orange and green, respectively. Residues displaying high fluctuations are shown in red in chain A, and in blue in chain B. Catalytic His41 is shown as blue spheres and catalytic Cys145 is shown as yellow spheres on each chain. Images were generated using PyMOL.

In conclusion, the RMSF results show that the M^{pro} residues display typical fluctuation patterns in the presence of the KLQ*** substrates, except for KLQAND and KLQSVQ, of course. The RMSF in both monomers approximated one another in magnitude and localisation, in terms of flexible regions of the M^{pro}. Curiously, the KLQAEQ and KLQSGA systems did not demonstrate atypical, or similar behaviours to KLQAND and KLQSVQ as previously shown in RMSD and Rg. The visualisation of the systems revealed the key events that account for these nuances/variations in the KLQAEQ, KLQSGA, KLQAND and KLQSVQ systems. The details are further discussed below.

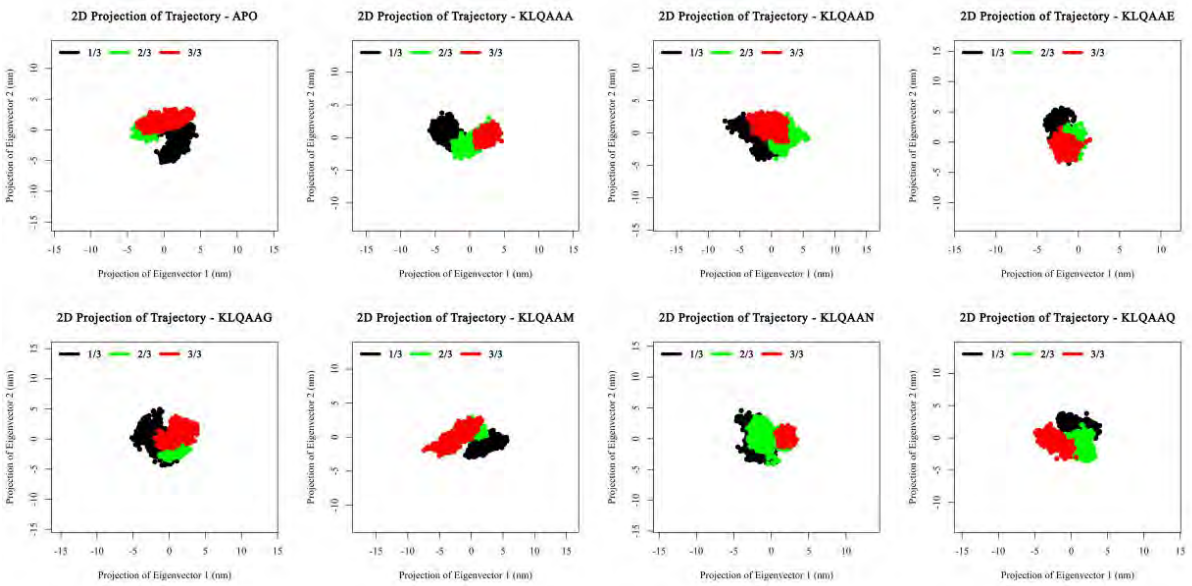
4.4.3 ANALYSIS OF THE PROMINENT MOTIONS OF THE M^{PRO} SYSTEMS

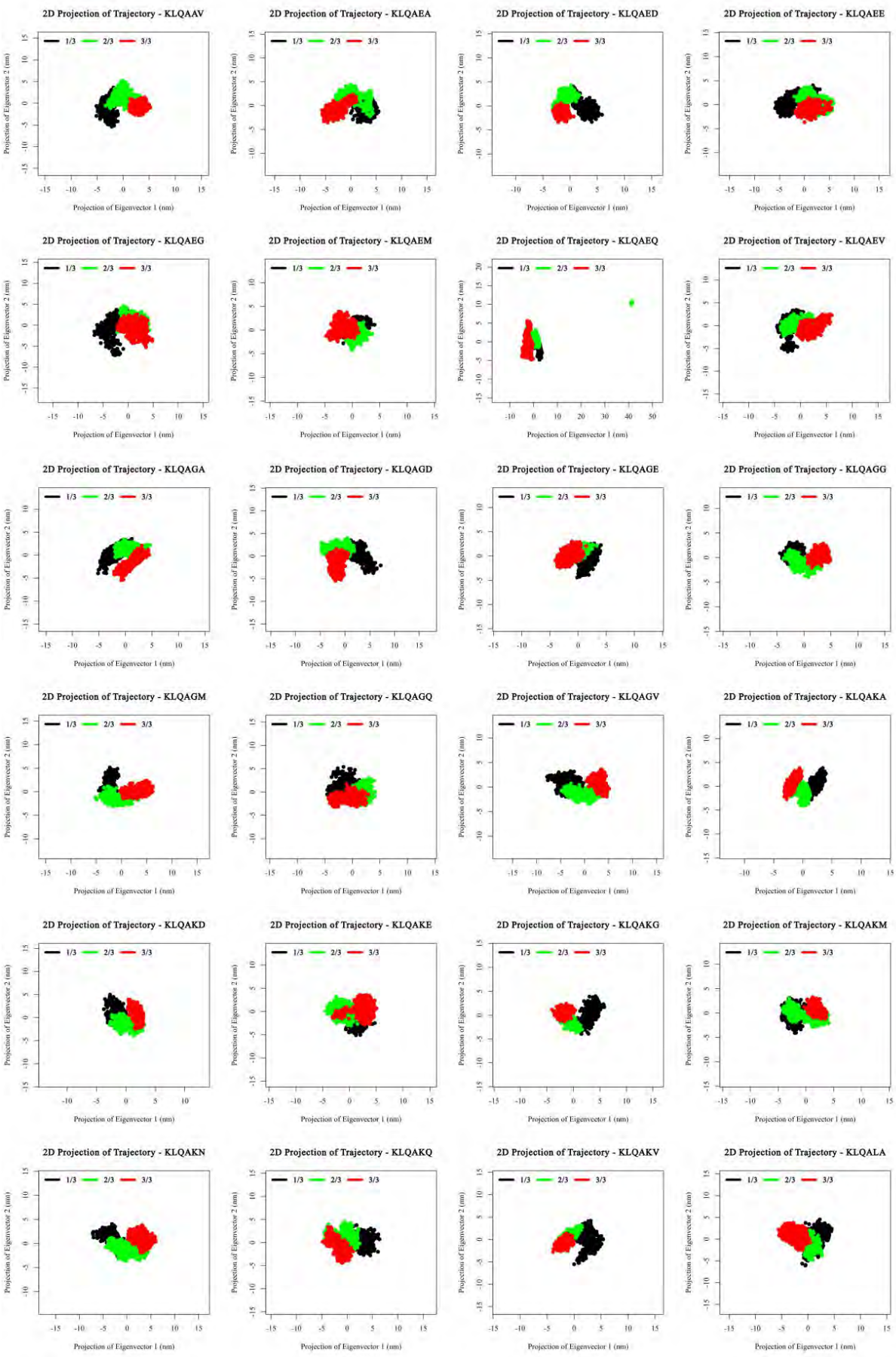
4.4.3.1 PCA

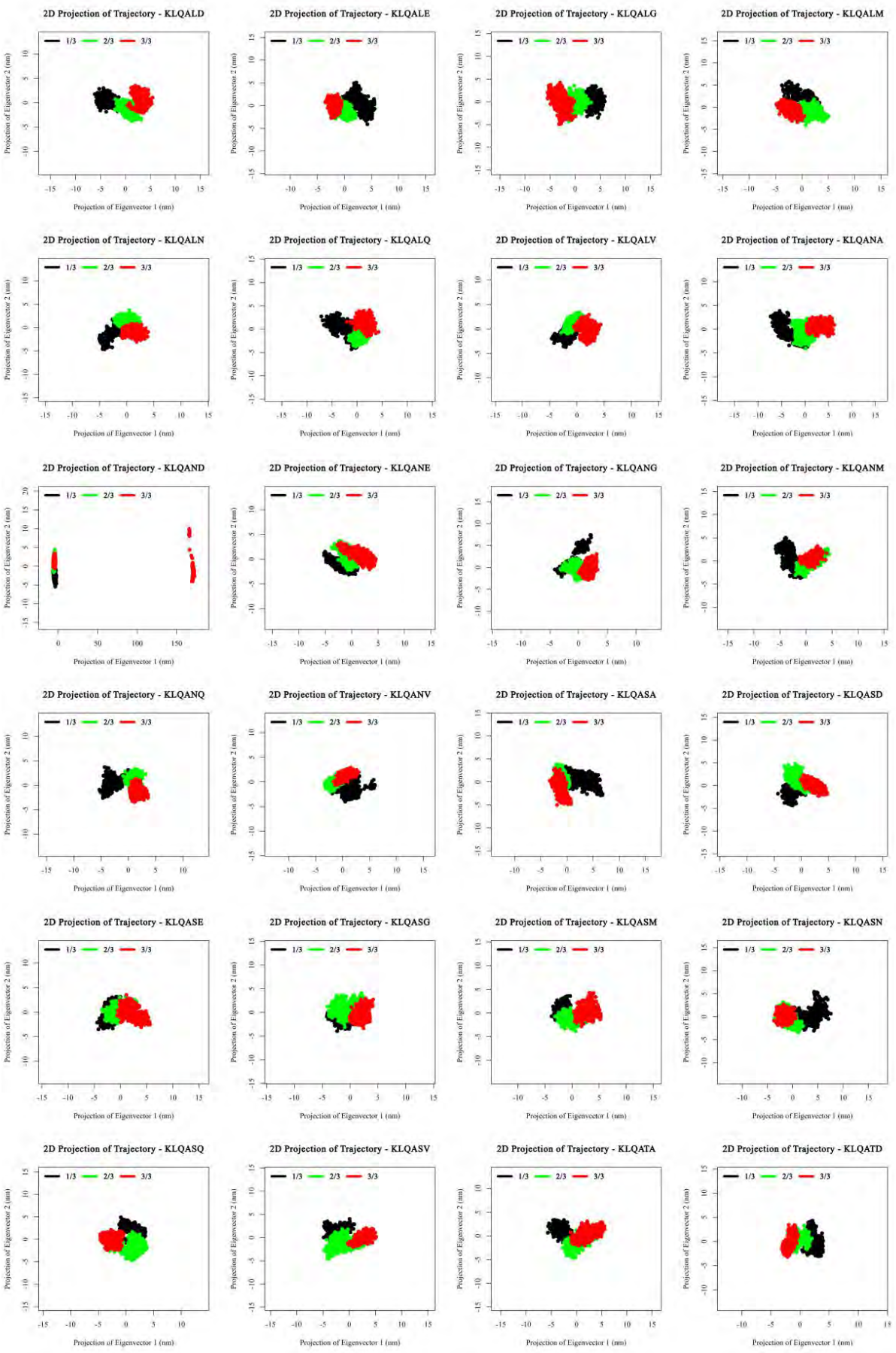
The prominent structural motions and conformational changes of the M^{pro} backbone during the 20 ns MD simulations were assessed using PCA calculations. PCA divided the overall motion of the trajectories into principal components that describe the essential functional protein motions during the simulation. Since the first two principal components, PC1 and PC2, retain the majority of the variance of the original data, they can be used to provide a meaningful description of the protein motions throughout the course of the simulations. Thus, 2D projections of these principal components were plotted using the Cartesian coordinates of all backbone atoms and used to visualise and examine these conformational changes (figure

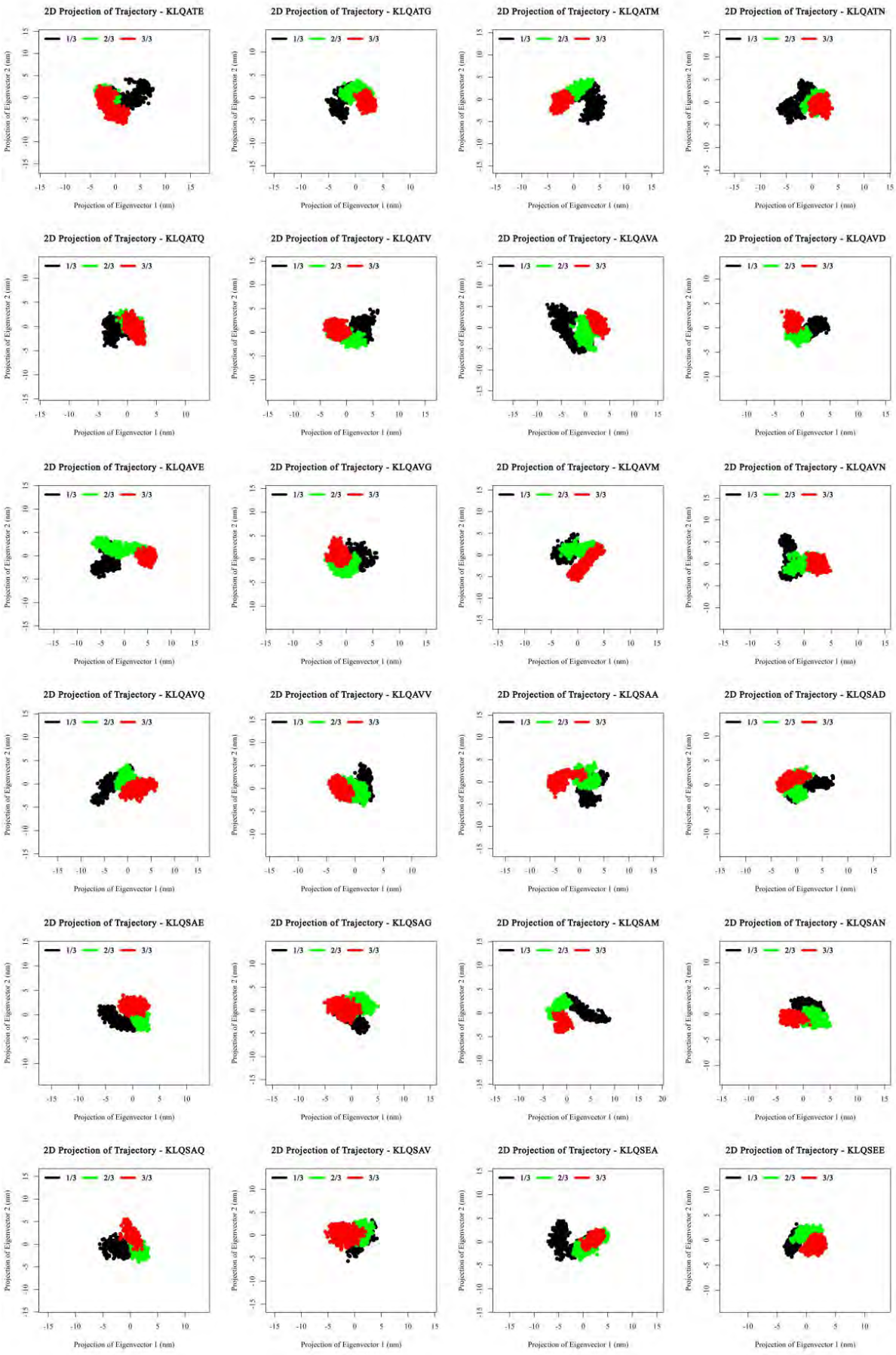
4.11).

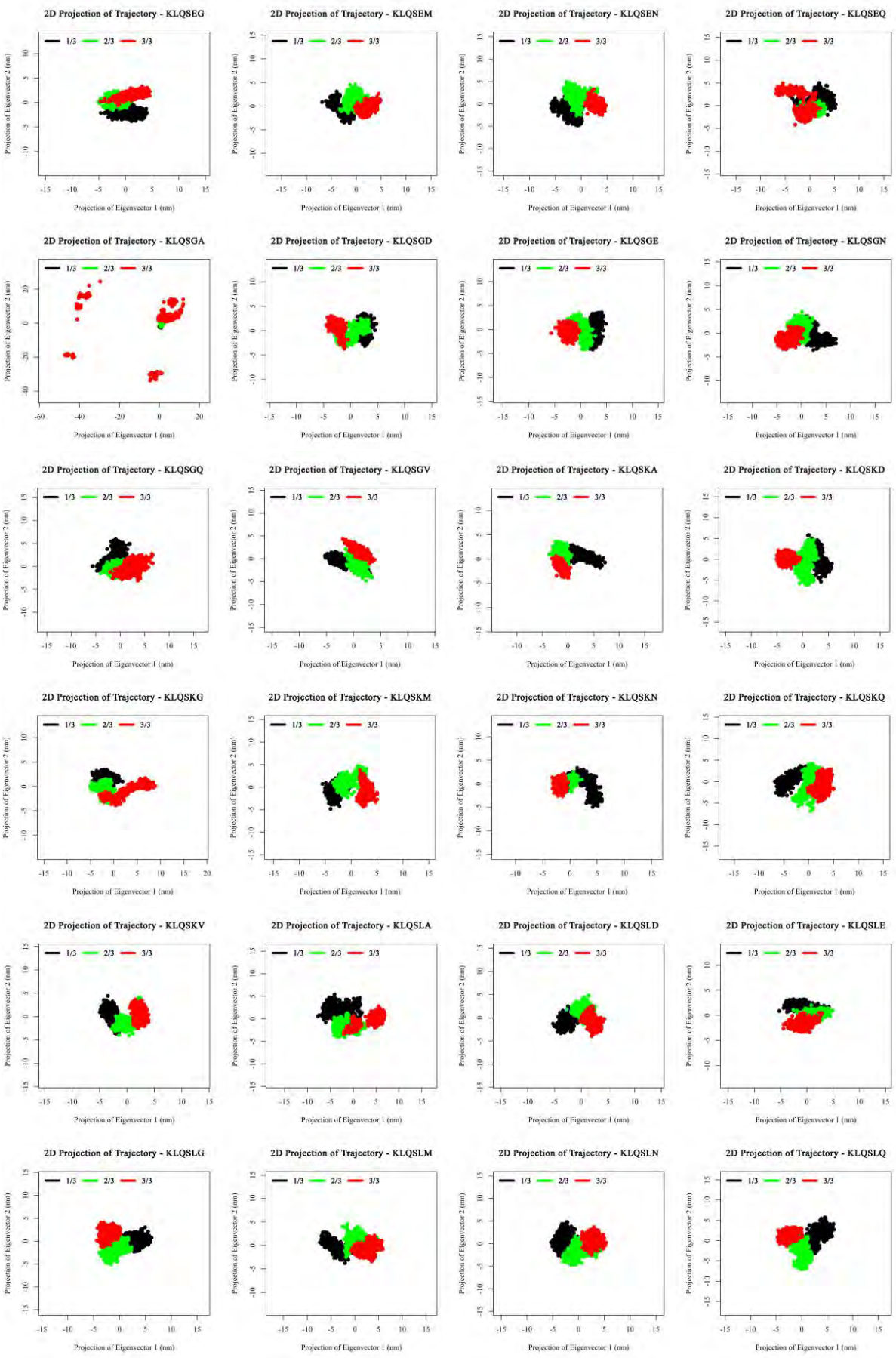
The coordinates were divided into three to observe the motions of the systems at various time periods in the duration of the simulation. The black points depict the backbone motions in the first third (frames 0-666), the green points indicate the motions in the second third (frames 667-1333), and the red points show the motions of the final third (frames 1334-2001) of the simulation, as shown in figure 4.11. The different systems demonstrated different patterns in protein motions and backbone conformational changes. The direction of change of the PCA plot according to timeframe was unique to each system, with some demonstrating a clockwise change in PCA (like KLQAAD and KLQAAE), whilst others displayed an anti-clockwise change in PCA trajectory (like *APO* and KLQAAA). Seemingly, the majority of systems retained steady conformational changes throughout the simulation as the distribution of the coordinates of the PCA were generally compact. Regardless of the patterns shown by the PC1 and PC2 coordinates, the typical range for the coordinates was between -5 and 5 for both PC1 and PC2. The KLQAEQ, KLQAND, KLQSGA and KLQSVQ systems displayed the most drastic conformational changes, indicating structural instabilities during the simulation. The time periods in which these rapid conformational changes occur corresponded with the timestamps of the backbone instability illustrated by their RMSD and Rg results.

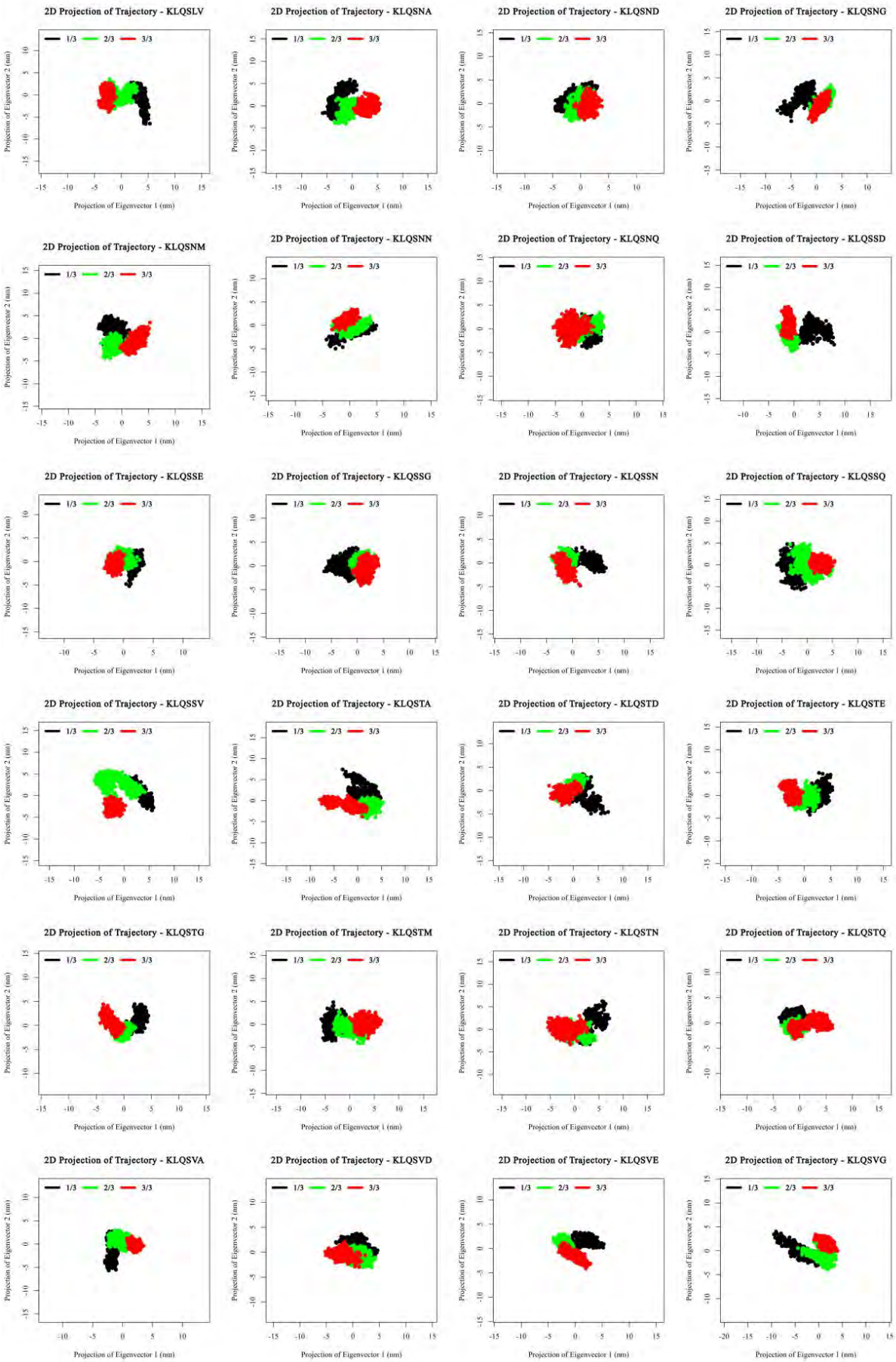



2D Projection of Trajectory - KLQAAV
2D Projection of Trajectory - KLQAEA
2D Projection of Trajectory - KLQAED
2D Projection of Trajectory - KLQAEE
2D Projection of Trajectory - KLQAEG
2D Projection of Trajectory - KLQAEM
2D Projection of Trajectory - KLQAEQ
2D Projection of Trajectory - KLQAEV
2D Projection of Trajectory - KLQAGA
2D Projection of Trajectory - KLQAGD
2D Projection of Trajectory - KLQAGE
2D Projection of Trajectory - KLQAGG
2D Projection of Trajectory - KLQAGM
2D Projection of Trajectory - KLQAGQ
2D Projection of Trajectory - KLQAGV
2D Projection of Trajectory - KLQAKA
2D Projection of Trajectory - KLQAKD
2D Projection of Trajectory - KLQAKE
2D Projection of Trajectory - KLQAKG
2D Projection of Trajectory - KLQAKM
2D Projection of Trajectory - KLQAKN
2D Projection of Trajectory - KLQAKQ
2D Projection of Trajectory - KLQAKV
2D Projection of Trajectory - KLQALA
1/3
2/3
3/3
Projection of Eigenvector 1 (nm)
Projection of Eigenvector 2 (nm)


2D Projection of Trajectory - KLQALD
2D Projection of Trajectory - KLQALE
2D Projection of Trajectory - KLQALG
2D Projection of Trajectory - KLQALM
2D Projection of Trajectory - KLQALN
2D Projection of Trajectory - KLQALQ
2D Projection of Trajectory - KLQALV
2D Projection of Trajectory - KLQANA
2D Projection of Trajectory - KLQAND
2D Projection of Trajectory - KLQANE
2D Projection of Trajectory - KLQANG
2D Projection of Trajectory - KLQANM
2D Projection of Trajectory - KLQANQ
2D Projection of Trajectory - KLQANV
2D Projection of Trajectory - KLQASA
2D Projection of Trajectory - KLQASD
2D Projection of Trajectory - KLQASE
2D Projection of Trajectory - KLQASG
2D Projection of Trajectory - KLQASM
2D Projection of Trajectory - KLQASN
2D Projection of Trajectory - KLQASQ
2D Projection of Trajectory - KLQASV
2D Projection of Trajectory - KLQATA
2D Projection of Trajectory - KLQATD
1/3
2/3
3/3
Projection of Eigenvector 1 (nm)
Projection of Eigenvector 2 (nm)


2D Projection of Trajectory - KLQATE
2D Projection of Trajectory - KLQATG
2D Projection of Trajectory - KLQATM
2D Projection of Trajectory - KLQATN
2D Projection of Trajectory - KLQATQ
2D Projection of Trajectory - KLQATV
2D Projection of Trajectory - KLQAVA
2D Projection of Trajectory - KLQAVD
2D Projection of Trajectory - KLQAVE
2D Projection of Trajectory - KLQAVG
2D Projection of Trajectory - KLQAVM
2D Projection of Trajectory - KLQAVN
2D Projection of Trajectory - KLQAVQ
2D Projection of Trajectory - KLQAVV
2D Projection of Trajectory - KLQSAA
2D Projection of Trajectory - KLQSAD
2D Projection of Trajectory - KLQSAE
2D Projection of Trajectory - KLQSAG
2D Projection of Trajectory - KLQSAM
2D Projection of Trajectory - KLQSAN
2D Projection of Trajectory - KLQSAQ
2D Projection of Trajectory - KLQSAV
2D Projection of Trajectory - KLQSEA
2D Projection of Trajectory - KLQSEE
1/3
2/3
3/3
Projection of Eigenvector 1 (nm)
Projection of Eigenvector 2 (nm)


2D Projection of Trajectory - KLQSEG
2D Projection of Trajectory - KLQSEM
2D Projection of Trajectory - KLQSEN
2D Projection of Trajectory - KLQSEQ
2D Projection of Trajectory - KLQSGA
2D Projection of Trajectory - KLQSGD
2D Projection of Trajectory - KLQSGE
2D Projection of Trajectory - KLQSGN
2D Projection of Trajectory - KLQSGQ
2D Projection of Trajectory - KLQSGV
2D Projection of Trajectory - KLQSKA
2D Projection of Trajectory - KLQSKD
2D Projection of Trajectory - KLQSKG
2D Projection of Trajectory - KLQSKM
2D Projection of Trajectory - KLQSKN
2D Projection of Trajectory - KLQSKQ
2D Projection of Trajectory - KLQSKV
2D Projection of Trajectory - KLQSLA
2D Projection of Trajectory - KLQSLD
2D Projection of Trajectory - KLQSLE
2D Projection of Trajectory - KLQSLG
2D Projection of Trajectory - KLQSLM
2D Projection of Trajectory - KLQSLN
2D Projection of Trajectory - KLQSLQ
1/3
2/3
3/3
Projection of Eigenvector 1 (nm)
Projection of Eigenvector 2 (nm)


2D Projection of Trajectory - KLQSLV
2D Projection of Trajectory - KLQSNA
2D Projection of Trajectory - KLQSND
2D Projection of Trajectory - KLQSNG
2D Projection of Trajectory - KLQSNM
2D Projection of Trajectory - KLQSNN
2D Projection of Trajectory - KLQSNQ
2D Projection of Trajectory - KLQSSD
2D Projection of Trajectory - KLQSSE
2D Projection of Trajectory - KLQSSG
2D Projection of Trajectory - KLQSSN
2D Projection of Trajectory - KLQSSQ
2D Projection of Trajectory - KLQSSV
2D Projection of Trajectory - KLQSTA
2D Projection of Trajectory - KLQSTD
2D Projection of Trajectory - KLQSTE
2D Projection of Trajectory - KLQSTG
2D Projection of Trajectory - KLQSTM
2D Projection of Trajectory - KLQSTN
2D Projection of Trajectory - KLQSTQ
2D Projection of Trajectory - KLQSVA
2D Projection of Trajectory - KLQSVD
2D Projection of Trajectory - KLQSVE
2D Projection of Trajectory - KLQSVG
1/3
2/3
3/3
Projection of Eigenvector 1 (nm)
Projection of Eigenvector 2 (nm)

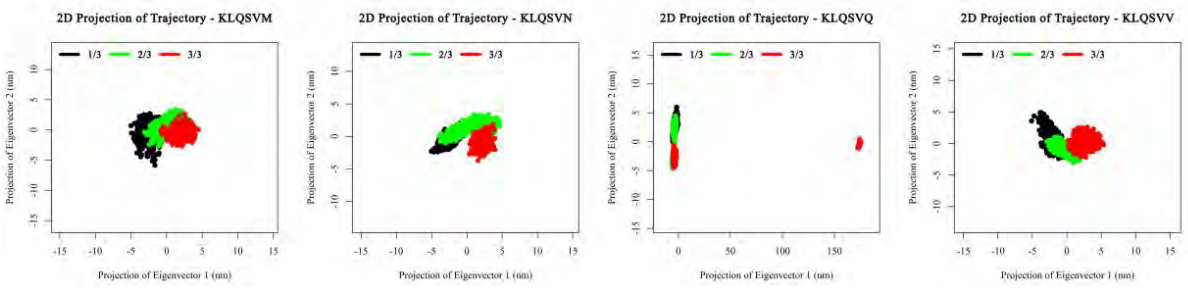


Figure 4.11. The 2D projections of the principal components for M^{pro} *apo* and KLQ*-substrate-bound systems over the duration of the 20 ns MD simulations**. The projection of the motion along phase space for PC1 and PC2 of M^{pro} *apo* and KLQ***-substrate-bound systems, showing the first third (black), second third (green) and final third (red) of the 20 ns simulation. Images were generated using Xmgrace (of Grace 5) and RStudio.

4.4.3.2 CLASSIFICATION OF THE PCA

In an attempt to classify the PCA data, a custom pairwise comparison of the M^{pro} systems was performed. The Cartesian coordinates of PC1 and PC2 for each system were fitted in a 5×5 grid to demarcate the protein motions as defined by PCA over the course of the simulation. In intervals of 2 ns (200 frames), the PC1 and PC2 coordinates of the protein were averaged and subsequently used to position the PCA within a grid. This was assigned a letter code to uniquely identify the position of the PCA in the particular time interval (Appendix P). The string of letters therefore uniquely identified the progression in PCA for a particular simulation (Appendix Q). The differences in these codes between simulations were used to create a pairwise comparison that included all the systems (Appendices R & S). This assumes that the motion described by PC1 and PC2 is the same in compared systems. However, this was in an attempt to identify systems with similar motion across 132 simulations; once identified, the similarity of motion could be validated through the superimposition of the structures.

The differences between these PCA codes were clustered using correlation as a measure of distance (figure 4.12). Figure 4.12 shows the cluster map, and this illustrates the correlation of the protein motions for all the systems, with respect to one another (bearing in mind the underlying assumption). The accompanying hierarchy also indicates four main clades of PCA plot, where progression of PCA within the clade is similar. This hierarchical clustering was performed to indicate hierarchical relationships between the systems, with regard to their dynamic motion (figure 4.13). Hierarchical clustering allowed for the arrangement of the systems based on similarity. The resultant dendrogram of the hierarchical clustering also clearly shows these four main groups of systems sharing similarities. The main clades are shown in table 4.1. While three groups of PCA plots share varying similarities within and between themselves, group 1 PCA (comprising of the KLQSVQ, KLQAEQ and KLQAND

systems) demonstrated the highest dissimilarity to the other groups. Considering the results of RMSD, Rg and even RMSF (for KLQAND and KLQSVQ systems), this is highly expected and relates to the dynamic events that indicate destabilisation of these systems during the simulations. This illustrated the ability of this clustering of PCA codes to identify highly dissimilar motion, this dissimilar motion is already confirmed from RMSD, Rg and RMSF. Interestingly, the KLQSGA system which also showed backbone instability in RMSD and Rg was arranged in group 3 and not alongside the KLQSVQ, KLQAEQ and KLQAND systems. However, the dendrogram in figure 4.13 indicates that KLQSGA displayed the highest dissimilarity with the rest of the systems within the group, delineating a weak correlation in terms of protein motions and conformational changes with "similar" systems.

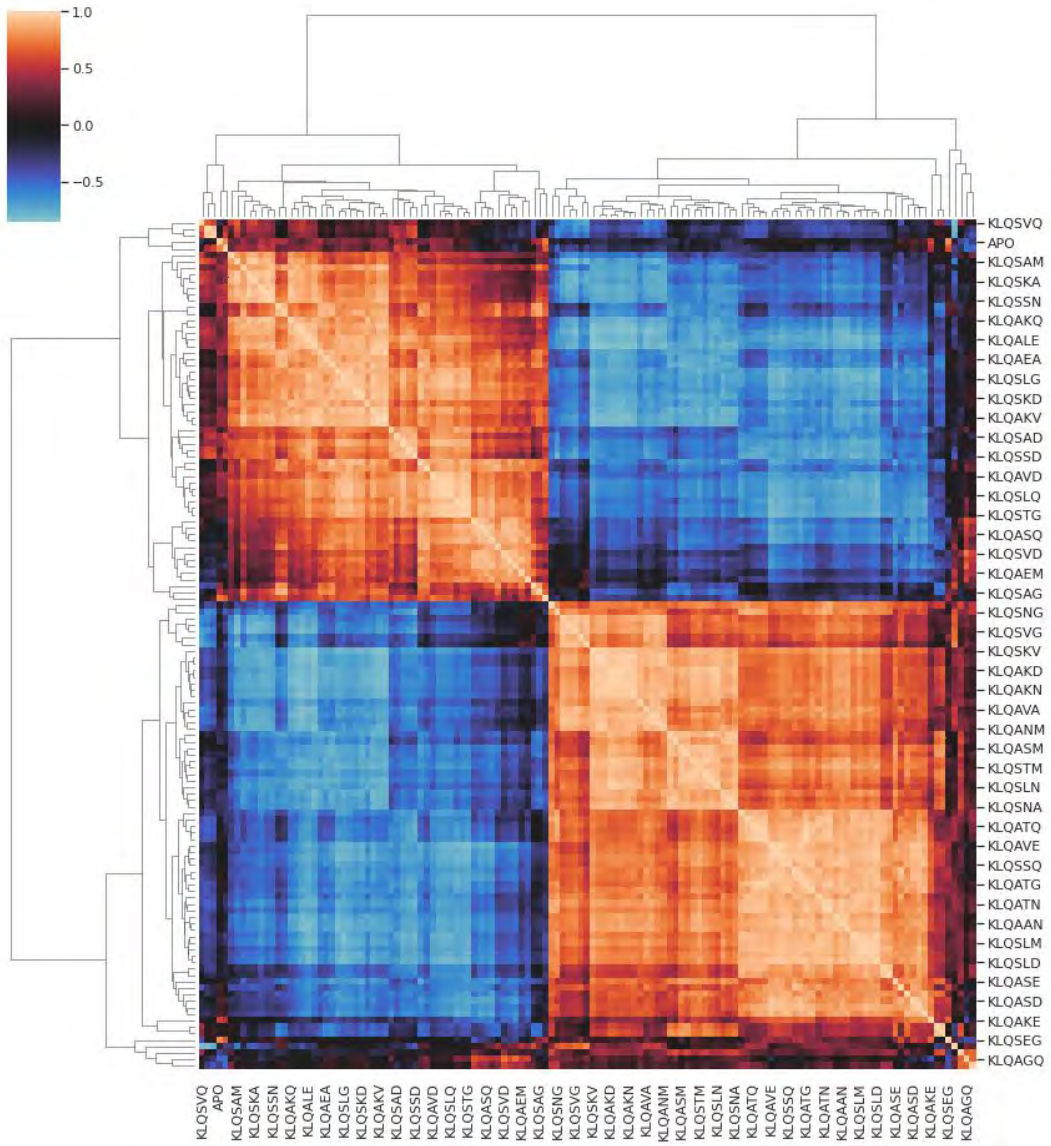


Figure 4.12. Determination of correlation in the protein motions of M^{pro} dynamic systems. The clustering of the differences in protein motions of M^{pro} systems using correlation as a measure of distance. The image was generated using Seaborn in Python.

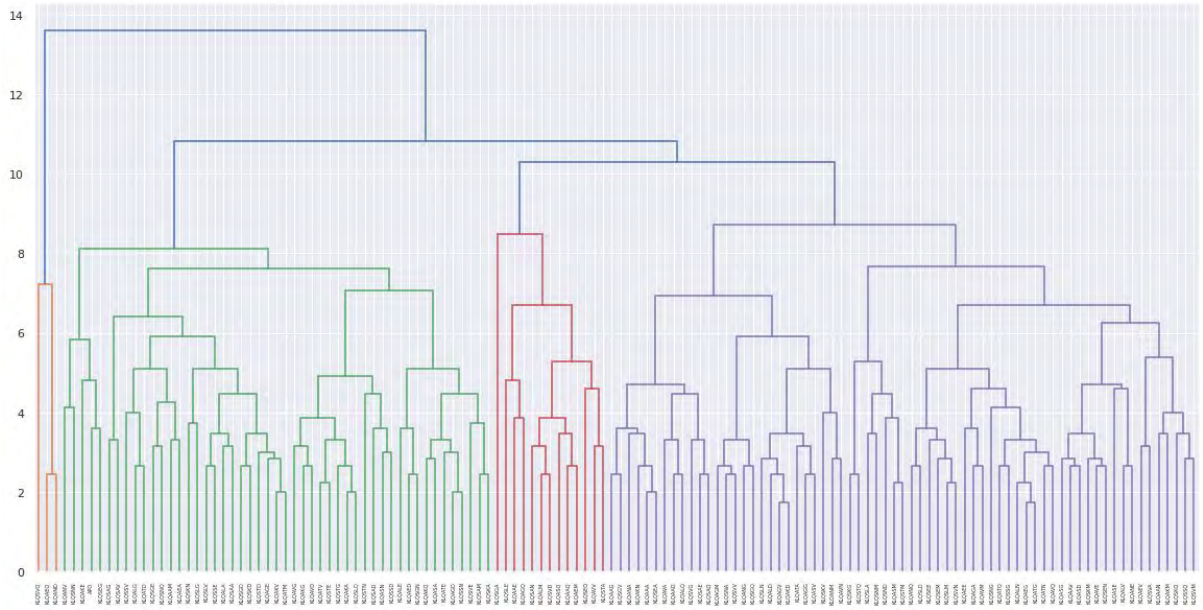


Figure 4.13. The similarity of the protein dynamic motion in the *apo*-M^{pro} and Hexapeptide-M^{pro} systems. The dendrogram shows the arrangement of the M^{pro} systems according to similarity. The image was generated using Seaborn in Python.

Table 4.1: The related M^{pro} systems based on conformational changes in the duration of the 20 ns simulation.

	Groups			
	1	2	3	4
Substrates	KLQSVQ; KLQAEQ; KLQAND	KLQANV; KLQSNN; KLQAKE; *APO*; KLQSEG; KLQSAG; KLQSAV; KLQSSV; KLQALG; KLQATD; KLQSGE; KLQSNQ; KLQAAM; KLQAEA; KLQSKN; KLQSLG; KLQSLV; KLQSSE; KLQALA; KLQSAA; KLQSGD; KLQSKD; KLQSTD; KLQAGE; KLQAKV; KLQATM; KLQAVG; KLQAKG; KLQAVD; KLQATV; KLQSTE; KLQSTG; KLQAKA; KLQSLQ; KLQSTN; KLQSAD; KLQASN; KLQSSD; KLQALE; KLQAED; KLQSGN; KLQAKQ; KLQASA; KLQATE; KLQAGD; KLQSSN; KLQSVE; KLQSAM; KLQSKA	KLQSGA; KLQSLE; KLQAAE; KLQAGQ; KLQSAN; KLQALM; KLQSVD; KLQASQ; KLQAAQ; KLQAEM; KLQSEQ; KLQAVV; KLQSTA	KLQAAG; KLQSGV; KLQANA; KLQAKN; KLQAAA; KLQSEA; KLQAVA; KLQAAD; KLQALQ; KLQSVG; KLQSAE; KLQSAQ; KLQAGM; KLQSNA; KLQSVV; KLQANG; KLQSGQ; KLQSLN; KLQALD; KLQAGV; KLQAKD; KLQATA; KLQAGG; KLQASV; KLQSKV; KLQANM; KLQAVN; KLQSKG; KLQSTQ; KLQSLA; KLQSNM; KLQSND; KLQASM; KLQSTM; KLQANQ; KLQSLD; KLQSEE; KLQSEM; KLQSLM; KLQSVN; KLQAEE; KLQAGA; KLQAVM; KLQSNG; KLQATQ; KLQSSG; KLQALN; KLQAEG; KLQATG; KLQATN; KLQAVQ; KLQASG; KLQAAV; KLQASD; KLQSKM; KLQAVE; KLQSEN; KLQASE; KLQALV; KLQANE; KLQAEV; KLQSVA; KLQAAN; KLQAKM; KLQSKQ; KLQSSQ; KLQSVM

The PCA results are in agreement with RMSD, Rg and RMSF observations. All the systems displayed similar behaviour that correlates with the trends of the RMSD and Rg, shown by the equilibration around 0.2 nm and 2.60 nm, respectively. Despite binding substrates with minor chemical differences, the proteins achieved and retained similar backbone flexibility and degree of compaction, and overall displayed similar stability in their trajectories. The SARS-CoV-2 M^{pro} has an intrinsic mechanism that enables the protein to bind different peptide substrates without conferring instability to the entire structure. In PCA, the majority of the systems seemingly occupied the same spaces throughout the course of the simulation. The exceptions to these trends have consistently included KLQAEQ, KLQAND, KLQSGA

and KLQSVQ systems. The hierarchical clustering even showed the high dissimilarity the protein motions in these systems share with the rest. However, the explanation for the partitioning of these select systems owes to the events that occur during the MD simulation. Thus, two systems from each hierarchical clade (or group) were selected for trajectory visualization using VMD.

4.4.3.3 VISUALISATION OF THE TRAJECTORIES

The visualisation of the trajectories was carried out to assess the similarities of protein motions implied in the hierarchical clustering and to monitor the events that caused the backbone destabilisation in the KLQAEQ, KLQAND, KLQSGA and KLQSVQ systems.

The visualisation of group 1 systems (KLQAEQ and KLQSVQ) provided insight into the instability of the backbone of the protein systems. Figure 4.14 shows that the systems appear to not overlap at any point of the simulation, indicating a weak relationship (in terms of movements) between the two protein systems. In KLQAEQ, the timestamps of steep increases in RMSD and Rg correspond to unbinding events of the substrate. The ejection of non-covalently bound ligands from the active site normally does not produce such extreme changes to the RMSD and Rg of the protein backbone. In this case, the peaks are shown as a result of including the substrates as a third chain of the M^{pro} before topology generation, as previously mentioned. Thus, the motions of the substrates contributed to and affected the values for RMSD and Rg. In terms of RMSF, the unbinding of the substrates is not visible in the RMSF plots, since these RMSF plots were limited to chains A and B. This explains why the KLQAEQ system was not included in RMSF with the extremely high-fluctuation systems. The RMSF only measured the deviations of the residues of each of the monomers. Hence, no high RMSF values were registered for the system. The spikes in RMSD and Rg for KLQSVQ corresponded with the dissociation of the M^{pro} dimer (figure 4.14). The movement apart of the monomers caused a great displacement of the subunit backbones and consequently, induced the high values for RMSD, Rg and RMSF.

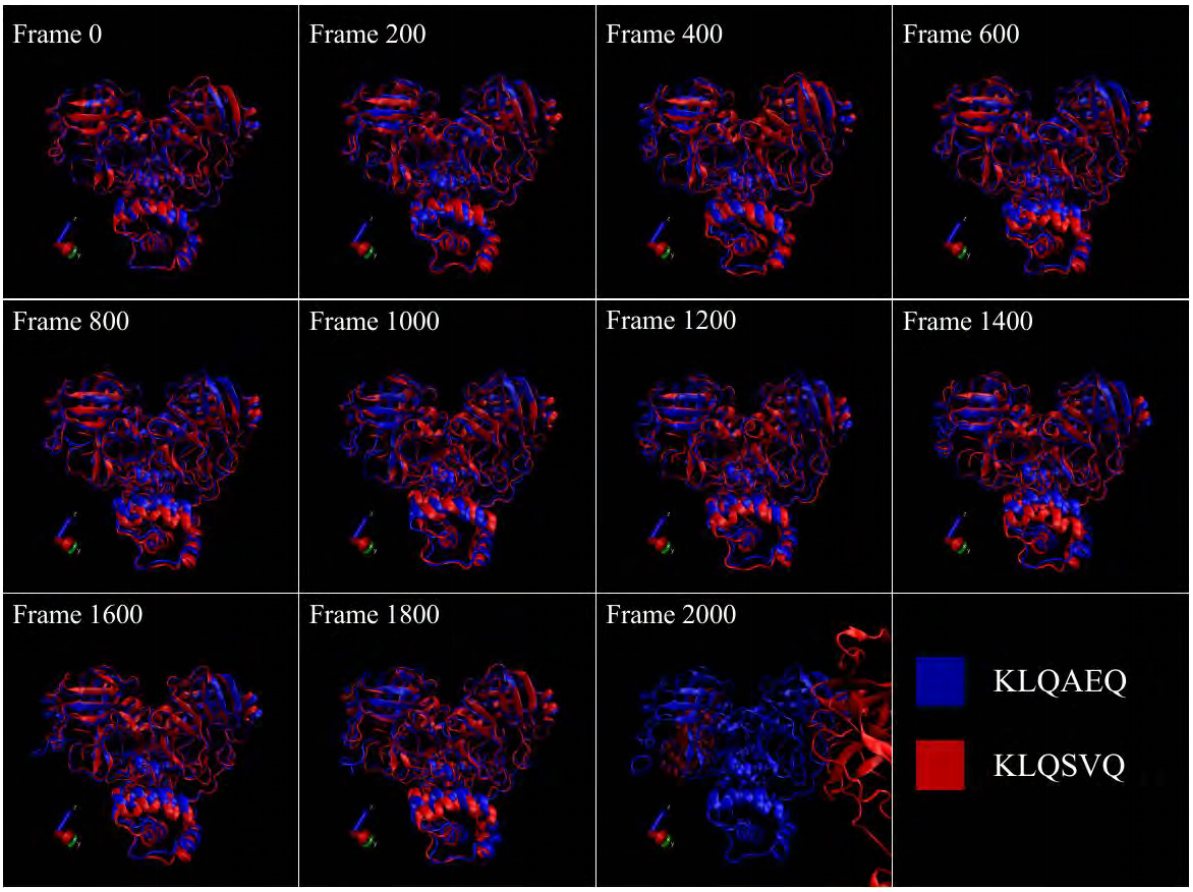


Figure 4.14. Visualisation of the MD trajectories for Group 1 systems. Protein systems are shown cartoon representation, showing M^{pro}-KLQAEQ in blue and M^{pro}-KLQSVQ in red. Images were generated using VMD.

The visualisation of the KLQAND system also revealed dimer dissociation at the timestamps corresponding to the steep spikes in RMSD and Rg (figure 4.15). The same explanation for the high values in RMSD, Rg and RMSF also applies in this system. The dimer dissociation resulted in great displacement of the monomers, thus increasing deviation from reference points and inducing high RMSD, Rg and RMSF values. Curiously, in both systems, the M^{pro} dimer was restored after initial dissociation and later further dissociated. The dissociation of the dimer in these systems proves to be an area of interest for future study. The identification of the underlying molecular interactions that governed this dissociation could prove beneficial in the investigation of antiviral agents.

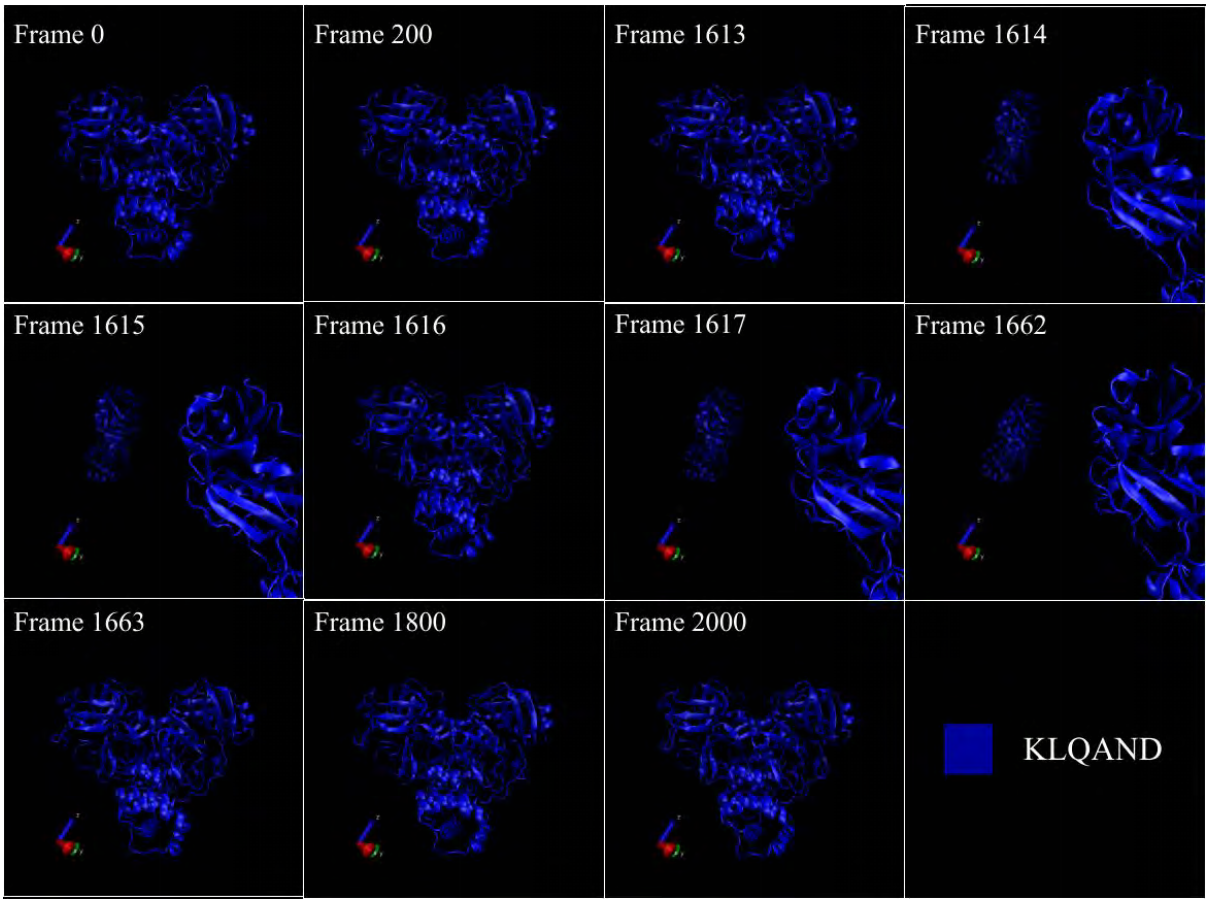


Figure 4.15. Visualisation of the MD trajectories for the KLQAND system. Protein systems are shown cartoon representation, shown in blue. The images were generated using VMD.

The visualisation of group 2 systems (the *apo*-M^{pro} and M^{pro}-KLQSEG) showed a strong relationship in protein motions as the structures overlapped throughout all frames of dynamics (figure 4.16). Particularly, Domains I and II (the chymotrypsin-like structure) consistently overlapped throughout the simulation, while the helices of Domain III showed the most difference. The mapping of high-flexibility residues indicated this domain to be highly flexible.

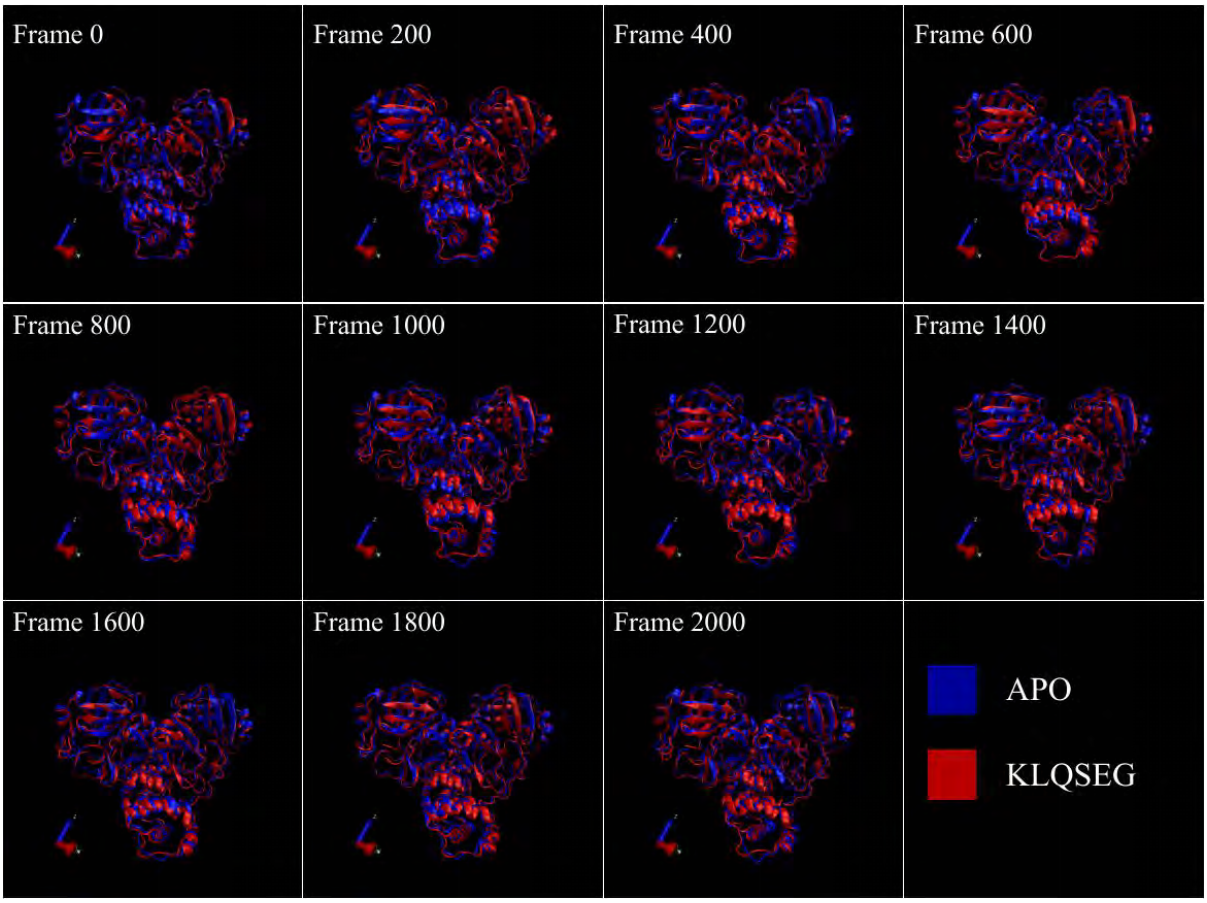


Figure 4.16. Visualisation of the MD trajectories for Group 2 systems. Protein systems are shown cartoon representation, showing *apo*-M^{pro} in blue and M^{pro}-KLQSVQ in red. Images were generated using VMD.

.

Group 3 systems (M^{pro}-KLQAVV and M^{pro}-KLQSGA) displayed little similarity in protein motions at any point of the simulation (figure 4.17). This is supported by the placement of these systems in dendrogram, where KLQSGA was shown to have high dissimilarity with the rest of the system in group 3. Nonetheless, the KLQAVV system demonstrated structural stability, as well as ligand stability throughout the simulation. The KLQSGA systems showed the unbinding of the substrate at corresponding time stamps with steep spikes in RMSD and Rg. This substrate ejection is seemingly the cause of the dissimilarity of KLQSGA with other systems in groups 3. However, the relation with other group 3 systems could possibly be as a result of overlapping M^{pro} motions. Much like KLQAEQ, the ejection of the substrate induces the high values in RMSD and Rg, but did not affect the RMSF values because only M^{pro} residues were considered in the calculations. This also explains as to why the spikes in KLQAEQ and KLQSGA were not as drastic like KLQAND and KLQSVQ. The spikes were a result of substrate displacement (or deviation), which was very small with respect to the M^{pro}.

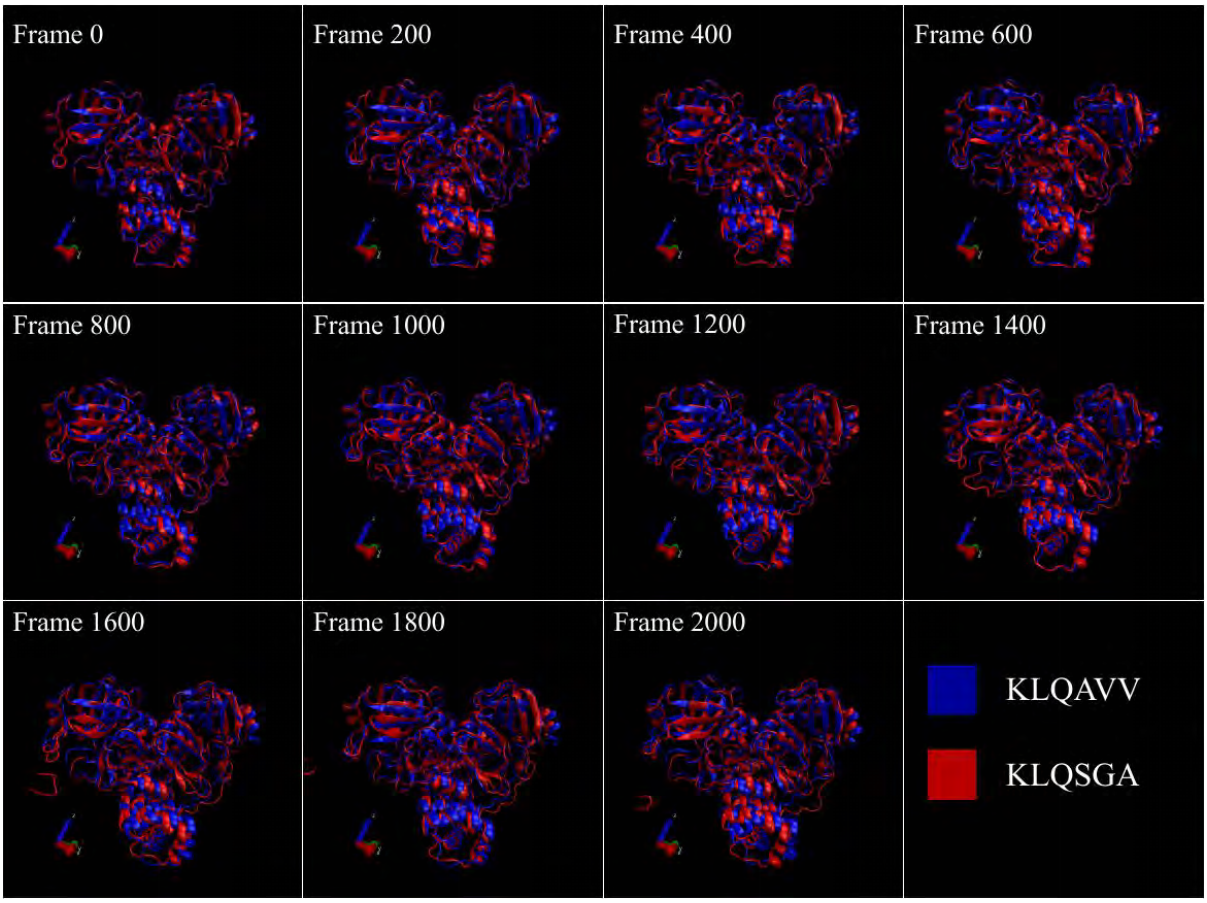


Figure 4.17. Visualisation of the MD trajectories for Group 3 systems. Protein systems are shown cartoon representation, showing M^{pro}-KLQAVV in blue and M^{pro}-KLQSGA in red. Images were generated using VMD.

Group 4 systems (M^{pro}-KLQAAA and M^{pro}-KLQSKG) also showed little similarity in motion (figure 4.18). Overlapping of structures was visible in the β-sheets of Domain I. Considering the placement of the KLQAAA and KLQSKG systems in the dendrogram (figure 4.13; table 4.1), a strong relation in protein motion was less likely. It could be that generally the motions are within a broad range, but visually it is not possible to see this detail, or it could be that simply the PC1 and PC2 motions are not the same in these systems.

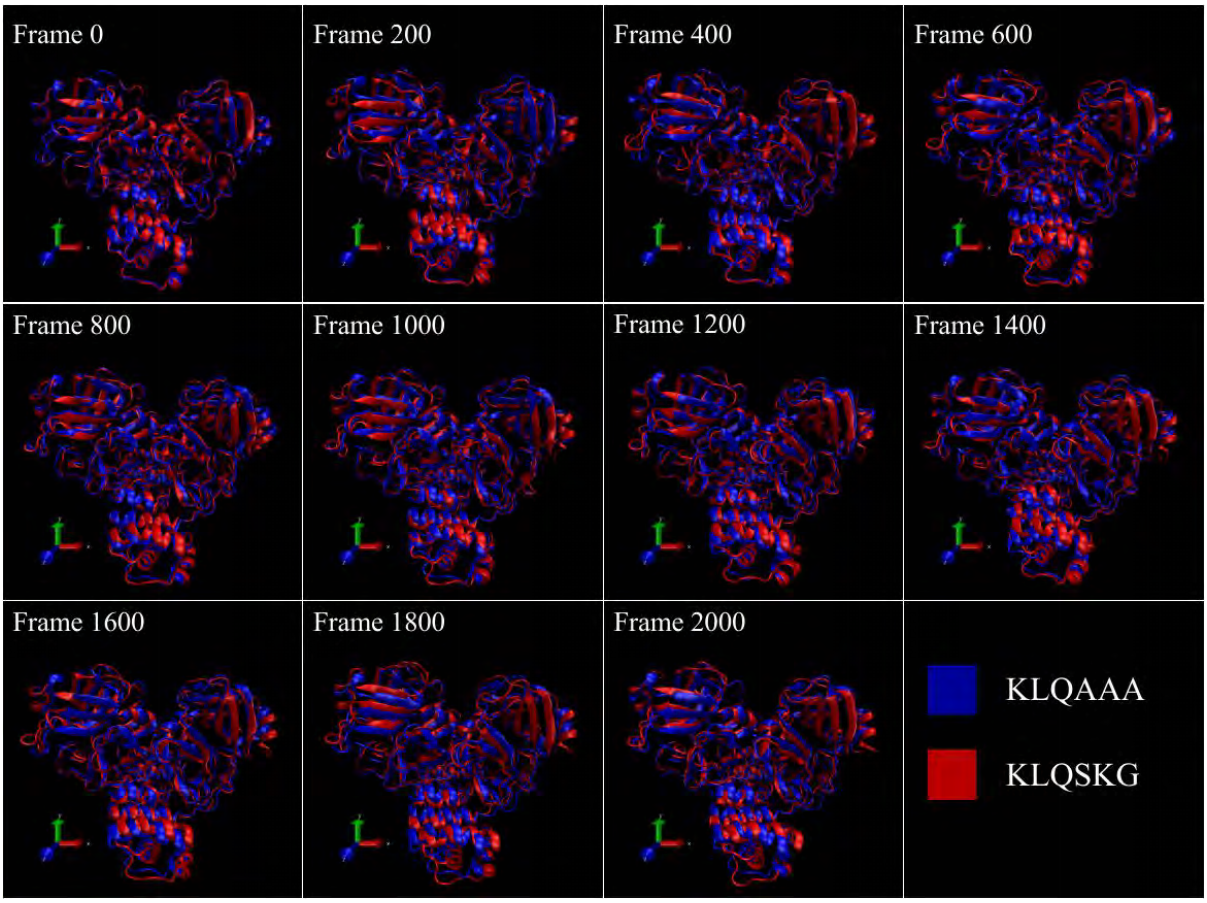


Figure 4.18. Visualisation of the MD trajectories for Group 3 systems. Protein systems are shown cartoon representation, showing M^{pro}-KLQAAA in blue and M^{pro}-KLQSKG in red. Images were generated using VMD.

Lastly, the visualisation of the systems from the different groups revealed little to no similarity in motion between the groups (figure 4.19). Structural overlaps rarely occurred during the simulation. These systems were representatives for each clade and each clade diverges from the highest point of dissimilarity as shown in figure 4.13. This observation further validates the outcomes of the hierarchical clustering since these systems exhibited no similar motions throughout the 20 ns simulation.

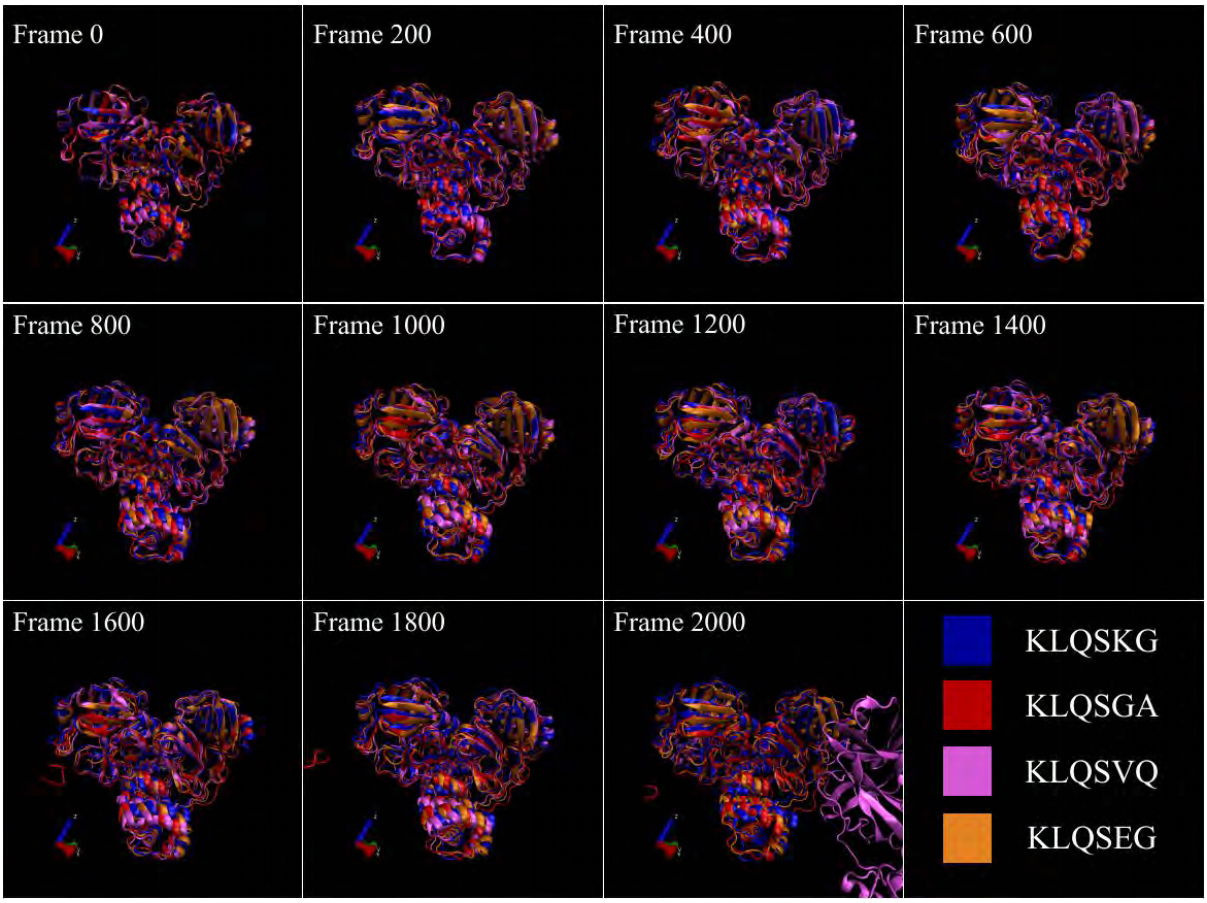


Figure 4.19. Visualisation of the MD trajectories for systems in all hierarchical groups. Protein systems are shown cartoon representation, showing M^{pro}-KLQSKQ in blue, M^{pro}-KLQSGA in red, M^{pro}-KLQSVQ in purple and M^{pro}-KLQSEG in orange. Images were generated using VMD.

This visualisation, of the four clades of PCA progression, allowed the inspection and confirmation of the similarities in protein motion of the systems. This shows that this type of PCA progression analysis is able, certainly in some cases to identify both similar and dissimilar motions of the protein. With refinement, it could provide for general use. In this case, it certainly provided focus on particular systems given the number of simulations to assess. For the second group of systems we have a set where the substrates affect the motion during dynamics in a very similar manner. These are listed in table 4.1 of this chapter. Furthermore, the dynamic events that indicated instability in the KLQAEQ, KLQAND, KLQSGA and KLQSVQ systems were assessed and this provided insight into the behaviour of the protein and the substrate in the respective systems.

4.5 CHAPTER SUMMARY

In this chapter, the objective to assess the behaviour and conformational changes of the M^{pro} in dynamic environments was addressed. MD simulations were performed on 132 M^{pro} systems (*apo*- and 131 KLQ*** hexapeptide bound), for a simulation period of 20 ns. The

trajectory files from these simulations were analysed in terms of RMSD, Rg, RMSF and PCA to monitor backbone fluctuation, protein folding stability, residue-level fluctuations and protein motion and conformational changes, respectively.

The majority of M^{pro} systems demonstrated relatively similar behaviour in terms of global stability of the backbone. Typical RMSD and Rg were equilibrated around 0.2 nm and 2.6 nm, respectively; indicating stable backbone flexibility and protein folding over the course of the 20 ns simulations. The exceptions to these trends were consistently the KLQAEQ, KLQAND, KLQSGA and KLQSVQ systems which displayed extreme increases in RMSD and Rg at various time frames when comparing the systems. These increases corresponded in timestamps in both RMSD and Rg within the same system. Despite having substrate docking performed exclusively on chain B of the M^{pro}, local fluctuation analysis saw the two M^{pro} chains approximately matching one another in magnitude and localisation of high-flexibility residues and regions of the protein. Localisation of high-flexibility residues showed high flexibility in loop regions joining β-sheets and the domains of each monomer, as well as the helices of Domain III. Semi-flexibility was observed in the catalytic domain which appeared to be functional fluctuation that accommodates the bound hexapeptide. Only the KLQAND and KLQSVQ systems registered disproportionately high RMSF values when compared to the rest of the systems in data.

In terms of protein motion and conformational changes, the different systems displayed unique PCA progression. Generally, the distribution of the coordinates of PC1 and PC2 were compact and ranged between -5 and 5. As with RMSD, Rg and RMSF, the KLQAEQ, KLQAND, KLQSGA and KLQSVQ systems registered extreme values for the coordinates of PC1 and PC2; in time periods corresponding to time frames of high RMSD, Rg and RMSF. In order to accommodate the analysis of 132 PCA plots, a strategy in terms of encoding PCA progression during the dynamic simulations was introduced, that could identify similar and dissimilar systems on the basis of protein motion and conformational changes. The strategy was able to identify that group 2 motions, according to sequences in Table 4.1, of the protein during dynamics is similar, and therefore the interaction between M^{pro} of these systems is likely similar. Taking results from RMSD and Rg, together with visualisation of the trajectories enabled the identification of the unstable systems where dimer dissociation was observed, in particular for systems KLQAND and KLQSVQ; and the unbinding of the substrates in KLQAEQ and KLQSGA systems.

CONCLUDING REMARKS AND FUTURE PROSPECTS

COVID-19, as a disease and pandemic, continues to cause devastation around the world. The strain on public health services is of extraordinary and unprecedented proportions. The disease is caused by SARS-CoV-2 infection. Existing measures placed to control the spread of the virus are losing efficacy due to the emergence of novel variants with increased virulence and immunological evasive mechanisms. The viral life cycle of SARS-CoV-2 relies greatly on the cleavage of polyproteins 1a and 1ab into mature non-structural proteins (nsps), facilitated by the main protease (M^{pro}) and papain-like proteases activity. After its autocleavage, the M^{pro} further cleaves downstream nsps at eleven sites, recognising the sequence Leu-Gln↓(Ser/Ala/Gly) (↓ shows the cleavage site). M^{pro} proves to be a promising drug target as it exhibits high degrees of conservation in sequence, structure and specificity. Therefore, this study sought to profile the binding of substrates in the context of hexapeptide substrates, onto SARS-CoV-2 M^{pro}.

In this study, a virtual multi-conformer substrate library was generated comprising 100 conformers of 810 unique hexapeptide sequences. Each hexapeptide was constructed to contain the recognition sequence and cleavage points and equally divided between the C- (P3-P1) and N-terminal (P1′-P3′) products. Terminal capping was successfully effected to safeguard the structural stability of each conformer.

The conformers were screened against chain B of the crystal structure of SARS-CoV-2 M^{pro} (PDB ID: 6XHM) using AutoDock Vina at high levels of exhaustiveness. After docking, the reproducibility of docking results was validated using the high-affinity poses. Calculation of ligand efficiency indices consistently showed residues Val, Ala, and Gly and Ala, to be efficient binders at P3, P1′, and P2′ and P3′, respectively. RLQ*** substrates exhibited the poorest binding efficiencies despite attaining the highest mean binding energy, and the best balance between BEI and SEI. Subsite mapping was performed to assess substrate recognition at the active site and the majority of hexapeptides showed appropriate binding modes. Resolution

of active site intermolecular interactions, as means to assess specificity, revealed a high prevalence of stabilising interactions, like hydrogen bonding, proving favourable binding and thus confirmed specificity. This specificity was also supported by the high-affinity binding of hexapeptides, as Vina scores ranged between -8.7 and -7.0 kcal.mol^{-1}. Hexapeptide binding modes and interactions showed optimal positioning of the substrates at the active site for proteolytic cleavage.

Complexed M^{pro} systems with 131 KLQ*** hexapeptides and an *apo*-M^{pro} were subjected to 20 ns MD runs to assess the strength of the interactions and the binding effect of the hexapeptides. System stability was assessed using RMSD, Rg and RMSF and revealed persistent stability in all but four systems. PCA was performed to assess the protein motion and conformational changes in the M^{pro} systems and showed a compact distribution of PC1 and PC2 in all but the same four systems. Custom pairwise comparison was conducted to quantify the PCA progression of each system and to subsequently determine similarities in PCA motion among the 132 systems through hierarchical clustering. Hierarchical clustering revealed four main clades (or groups) of similarity in the PCA progression. Trajectory visualization confirmed the calculated similarity within one group and verified dissimilarity across the groups. Visualization was also used to assess the dynamics of the four unstable systems and revealed substrate unbinding in KLQAEQ and KLQSVQ systems, and dimer dissociation in KLQAND and KLQSVQ systems.

This present study is a prelude for intended future studies which will seek to characterise the M^{pro} proteolytic mechanism using combined Quantum Mechanics/Molecular Mechanics techniques, as well as to explore and profile the conformational diversity of the hexapeptides, since they are fundamentally protein chains, using Replica Exchange Molecular Dynamics. Furthermore, future inhibition studies have many bases for rational drug design, such as position-specific efficient (binding) residues, binding modes (appropriate and inappropriate) of the hexapeptides, hexapeptide unbinding, dimer dissociating hexapeptides and so forth. Recommendations for future studies include permitting flexibility in M^{pro} active site residues during docking and longer MD simulations, allowing for more accurate

profiling of substrate binding and longer periods to assess the effect of the hexapeptides on the behaviour of the SARS-CoV-2 M^{pro}.

REFERENCES

Abad-Zapatero, C. and Metz, J.T., 2005. Ligand efficiency indices as guideposts for drug discovery. *Drug discovery today*, *10*(7), pp.464-469.

Abouzzohour, Y., 2020. COVID in the Maghreb: Responses and Impacts. *The COVID-19 Pandemic in the Middle East and North Africa*, pp.51-54.

Abraham, M.J., Murtola, T., Schulz, R., Páll, S., Smith, J.C., Hess, B. and Lindahl, E., 2015. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, *1*, pp.19-25.

Abramowitz, N., Schechter, I. and Berger, A., 1967. On the size of the active site in proteases II. Carboxypeptidase-A. *Biochemical and biophysical research communications*, *29*(6), pp.862-867.

Adcock, S.A. and McCammon, J.A., 2006. Molecular dynamics: survey of methods for simulating the activity of proteins. *Chemical reviews*, *106*(5), pp.1589-1615.

Adrian, T. and Natalucci, F., 2020a. COVID-19 Crisis Poses Threat to Financial Stability. [online] Available at: <https://blogs.imf.org/2020/04/14/covid-19-crisis-poses-threat-to-financial-stability/> [Accessed 14 July 2020].

Adrian, T. and Natalucci, F., 2020b. COVID-19 Worsens Pre-existing Financial Vulnerabilities. [online] Available at: <https://blogs.imf.org/2020/05/22/covid-19-worsens-pre-existing-financial-vulnerabilities/>[Accessed 14 July 2020].

Aleem, A., Samad, A.B.A. and Slenker, A.K., 2021. Emerging variants of SARS-CoV-2 and novel therapeutics against coronavirus (COVID-19). [online] Available at: <https://www.ncbi.nlm.nih.gov/books/NBK570580/> [Accessed 10 February 2022].

Alsaadi, E.A.J. and Jones, I.M., 2019. Membrane binding proteins of coronaviruses. *Future Virology*, *14*(4), pp.275-286.

Andreasen, M., Skeby, K.K., Zhang, S., Nielsen, E.H., Klausen, L.H., Frahm, H., Christiansen, G., Skrydstrup, T., Dong, M., Schiøtt, B. and Otzen, D., 2014. The importance of being capped: terminal capping of an amyloidogenic peptide affects fibrillation propensity and fibril morphology. *Biochemistry*, *53*(44), pp.6968-6980.

Ansari, M.A., Jamal, Q.M.S., Rehman, S., Almatroudi, A., Alzohairy, M.A., Alomary, M.N., Tripathi, T., Alharbi, A.H., Adil, S.F., Khan, M. and Malik, M.S., 2020. TAT-peptide conjugated repurposing drug against SARS-CoV-2 main protease (3CLpro): Potential therapeutic intervention to combat COVID-19. *Arabian journal of chemistry*, *13*(11), pp.8069-8079.

Awadasseid, A., Wu, Y., Tanaka, Y. and Zhang, W., 2021. Effective drugs used to combat SARS-CoV-2 infection and the current status of vaccines. *Biomedicine & Pharmacotherapy*, *137*, p.111330.

Azgari, C., Kilinc, Z., Turhan, B., Circi, D. and Adebali, O., 2021. The mutation profile of SARS-CoV-2 is primarily shaped by the host antiviral defense. *Viruses*, *13*(3), p.394.

Baden, L.R., El Sahly, H.M., Essink, B., Kotloff, K., Frey, S., Novak, R., Diemert, D., Spector, S.A., Rouphael, N., Creech, C.B. and McGettigan, J., 2021. Efficacy and safety of the mRNA-1273 SARS-CoV-2 vaccine. *New England Journal of Medicine*, *384*(5), pp.403-416.

Beigel, J.H., Tomashek, K.M., Dodd, L.E., Mehta, A.K., Zingman, B.S., Kalil, A.C., Hohmann, E., Chu, H.Y., Luetkemeyer, A., Kline, S. and Lopez de Castilla, D., 2020. Remdesivir for the treatment of Covid-19. *New England Journal of Medicine*, *383*(19), pp.1813-1826.

Berendsen, H.J., van der Spoel, D. and van Drunen, R., 1995. GROMACS: A message-passing parallel molecular dynamics implementation. *Computer physics communications*, *91*(1-3), pp.43-56.

Bhuiyan, A.I., Sakib, N., Pakpour, A.H., Griffiths, M.D. and Mamun, M.A., 2020. COVID-19-related suicides in Bangladesh due to lockdown and economic factors: case study evidence from media reports. *International Journal of Mental Health and Addiction*, pp.1-6.

Bianchi, M., Benvenuto, D., Giovanetti, M., Angeletti, S., Ciccozzi, M. and Pascarella, S., 2020. Sars-CoV-2 Envelope and Membrane Proteins: Structural Differences Linked to Virus Characteristics?. *BioMed Research International*, *2020*.

Binder, K., Horbach, J., Kob, W., Paul, W. and Varnik, F., 2004. Molecular dynamics simulations. *Journal of Physics: Condensed Matter*, *16*(5), p.S429.

Biniossek, M.L., Niemer, M., Maksimchuk, K., Mayer, B., Fuchs, J., Huesgen, P.F., McCafferty, D.G., Turk, B., Fritz, G., Mayer, J. and Haecker, G., 2016. Identification of protease specificity by combining proteome-derived peptide libraries and quantitative proteomics. *Molecular & cellular proteomics*, *15*(7), pp.2515-2524.

BioChemCoRe 2018, 2021. RMSD/RMSF Analysis. [online] Available at: <https://ctlee.github.io/BioChemCoRe-2018/rmsd-rmsf/> [Accessed 03 February 2022].

Boulware, K.T. and Daugherty, P.S., 2006. Protease specificity determination by using cellular libraries of peptide substrates (CLiPS). *Proceedings of the National Academy of Sciences*, *103*(20), pp.7583-7588.

Brooks, B.R., Brooks III, C.L., Mackerell Jr, A.D., Nilsson, L., Petrella, R.J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S. and Caflisch, A., 2009. CHARMM: the biomolecular simulation program. *Journal of computational chemistry*, *30*(10), pp.1545-1614.

Buheji, M., da Costa Cunha, K., Beka, G., Mavric, B., de Souza, Y.L., da Costa Silva, S.S., Hanafi, M. and Yein, T.C., 2020. The extent of covid-19 pandemic socio-economic impact on global poverty. a global integrative multidisciplinary review. *American Journal of Economics*, *10*(4), pp.213-224.

Cascella, M., Rajnik, M., Cuomo, A., Dulebohn, S.C. and Di Napoli, R., 2020. Features, evaluation and treatment coronavirus (COVID-19). *Statpearls*. StatPearls Publishing.

Case, D.A., Cheatham III, T.E., Darden, T., Gohlke, H., Luo, R., Merz Jr, K.M., Onufriev, A., Simmerling, C., Wang, B. and Woods, R.J., 2005. The Amber biomolecular simulation programs. *Journal of computational chemistry*, *26*(16), pp.1668-1688.

Chang, G.G., 2010. Quaternary structure of the SARS coronavirus main protease. In *Molecular Biology of the SARS-Coronavirus* (pp. 115-128). Springer, Berlin, Heidelberg.

Chen, H., Zhou, X., Gao, Y. and Zhou, J., 2017. Fragment-based drug design: Strategic advances and lessons learned. In *Drug Discovery Technologies* (pp. 212-232). Elsevier Inc..

Chen, J., Wang, R., Gilby, N.B. and Wei, G.W., 2021. Omicron (B. 1.1. 529): Infectivity, vaccine breakthrough, and antibody resistance. [online] Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8647651/> [Accessed 10 February 2022].

Chen, Y., Liu, Q. and Guo, D., 2020. Emerging coronaviruses: genome structure, replication, and pathogenesis. *Journal of medical virology*, *92*(4), pp.418-423.

Christen, M., Hünenberger, P.H., Bakowies, D., Baron, R., Bürgi, R., Geerke, D.P., Heinz, T.N., Kastenholz, M.A., Kräutler, V., Oostenbrink, C. and Peter, C., 2005. The GROMOS software for biomolecular simulation: GROMOS05. *Journal of computational chemistry*, *26*(16), pp.1719-1751.

Crooks, G.E., Hon, G., Chandonia, J.M. and Brenner, S.E., 2004. WebLogo: a sequence logo generator. Genome research, 14(6), pp.1188-1190.

David, C.C. and Jacobs, D.J., 2014. Principal component analysis: a method for determining the essential dynamics of proteins. In *Protein dynamics* (pp. 193-226). Humana Press, Totowa, NJ.

Daylight, 2019a. 3. SMILES - A Simplified Chemical Language. [online] Available at: <https://www.daylight.com/dayhtml/doc/theory/theory.smiles.html> [Accessed 15 March 2021].

Daylight, 2019b. 4. SMARTS - A Language for Describing Molecular Patterns. [online] Available at: <https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html> [Accessed 15 March 2021].

DeLano, W.L., 2002. PyMOL. [online] Available at: <https://www.ccp4.ac.uk/newsletters/newsletter40/11_pymol.pdf> [Accessed 25 July 2020].

DeLano, W.L., 2013. PyMOL. [online] Available at: <http://virology.wisc.edu/acp/Classes/DropFolders/Drop660_lectures/2013_660/L01_PyMOL_2013r.pdf> [Accessed 12 March 2021].

Di Micco, S., Musella, S., Scala, M.C., Sala, M., Campiglia, P., Bifulco, G. and Fasano, A., 2021. In silico Analysis Revealed Potential Anti-SARS-CoV-2 Main Protease Activity by the Zonulin Inhibitor Larazotide Acetate. *Frontiers in chemistry*, *8*, p.1271.

Do, P.C., Lee, E.H. and Le, L., 2018. Steered molecular dynamics simulation in rational drug design. *Journal of Chemical Information and Modeling*, *58*(8), pp.1473-1482.

Dong, Y.W., Liao, M.L., Meng, X.L. and Somero, G.N., 2018. Structural flexibility and protein adaptation to temperature: Molecular dynamics analysis of malate dehydrogenases of marine molluscs. *Proceedings of the National Academy of Sciences*, *115*(6), pp.1274-1279.

Du, X., Li, Y., Xia, Y.L., Ai, S.M., Liang, J., Sang, P., Ji, X.L. and Liu, S.Q., 2016. Insights into protein–ligand interactions: mechanisms, models, and methods. *International journal of molecular sciences*, *17*(2), p.144.

Duan, Y., Wu, C., Chowdhury, S., Lee, M.C., Xiong, G., Zhang, W., Yang, R., Cieplak, P., Luo, R., Lee, T. and Caldwell, J., 2003. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *Journal of computational chemistry*, *24*(16), pp.1999-2012.

Duan, Y., Wu, C., Chowdhury, S., Lee, M.C., Xiong, G., Zhang, W., Yang, R., Cieplak, P., Luo, R., Lee, T. and Caldwell, J., 2003. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *Journal of computational chemistry*, *24*(16), pp.1999-2012.

Eberhardt, J., Santos-Martins, D., Tillack, A. and Forli, S., 2021. AutoDock Vina 1.2. 0: new docking methods, expanded force field, and Python bindings. *Journal of Chemical Information and Modeling*.

Falsey, A.R., Frenck Jr, R.W., Walsh, E.E., Kitchin, N., Absalon, J., Gurtman, A., Lockhart, S., Bailey, R., Swanson, K.A., Xu, X. and Koury, K., 2021. SARS-CoV-2 neutralization with BNT162b2 vaccine dose 3. *New England Journal of Medicine*, *385*(17), pp.1627-1629.

Fang, W.J., Yakovleva, T. and Aldrich, J.V., 2011. A convenient approach to synthesizing peptide C-terminal N-alkyl amides. *Peptide Science*, *96*(6), pp.715-722.

Gahlawat, A., Kumar, N., Kumar, R., Sandhu, H., Singh, I.P., Singh, S., Sjöstedt, A. and Garg, P., 2020. Structure-based virtual screening to discover potential lead molecules for the SARS-CoV-2 main protease. *Journal of chemical information and modeling*, *60(12)*, pp.5781-5793.

Galeazzi, R., 2009. Molecular dynamics as a tool in rational drug design: current status and some major applications. *Current Computer-Aided Drug Design*, *5*(4), pp.225-240.

Gasteiger, J., Martin, Y., Nicholls, A., Oprea, T.I. and Stouch, T., 2018. Leaving us with fond memories, smiles, SMILES and, alas, tears: a tribute to David Weininger, 1952–2016. *Journal of Computer-Aided Molecular Design*, *32*, pp.313–319.

González, M.A., 2011. Force fields and molecular dynamics simulations. *École thématique de la Société Française de la Neutronique*, *12*, pp.169-200.

Gowers, R.J., Linke, M., Barnoud, J., Reddy, T.J.E., Melo, M.N., Seyler, S.L., Domanski, J., Dotson, D.L., Buchoux, S., Kenney, I.M. and Beckstein, O., 2019. *MDAnalysis: a Python package for the rapid analysis of molecular dynamics simulations* (No. LA-UR-19-29136). Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

Goyal, B. and Goyal, D., 2020. Targeting the dimerization of main protease of coronaviruses: A potential broad-spectrum therapeutic strategy. *ACS Combinatorial Science*, *22*(6), pp.297–305.

Grigalunas, M., Burhop, A., Christoforow, A. and Waldmann, H., 2020. Pseudo-natural products and natural product-inspired methods in chemical biology and drug discovery. *Current opinion in chemical biology*.

GROMACS, 2015. Acceleration and parallelization. [online] Available at: <https://www.gromacs.org/Documentation/Acceleration_and_parallelization> [Accessed 09 February 2022].

Guedes, I.A., de Magalhães, C.S. and Dardenne, L.E., 2014. Receptor–ligand molecular docking. *Biophysical reviews*, *6*(1), pp.75-87.

Guo, G., Ye, L., Pan, K., Chen, Y., Xing, D., Yan, K., Chen, Z., Ding, N., Li, W., Huang, H. and Zhang, L., 2020. New insights of emerging SARS-CoV-2: epidemiology, etiology, clinical features, clinical treatment, and prevention. *Frontiers in Cell and Developmental Biology*, *8*, p.410.

Gupta, S., Singh, A.K., Kushwaha, P.P., Prajapati, K.S., Shuaib, M., Senapati, S. and Kumar, S., 2021. Identification of potential natural inhibitors of SARS-CoV2 main protease by molecular docking and simulation studies. *Journal of Biomolecular Structure and Dynamics*, *39*(12), pp.4334-4345.

Ha, J.H. and Loh, S.N., 2012. Protein conformational switches: from nature to design. *Chemistry–A European Journal*, *18*(26), pp.7984-7999.

Handoko, S.D., Ouyang, X., Su, C.T.T., Kwoh, C.K. and Ong, Y.S., 2012. QuickVina: accelerating AutoDock Vina using gradient-based heuristics for global optimization. *IEEE/ACM transactions on computational biology and bioinformatics*, *9*(5), pp.1266-1272.

Hara, T., Takeda, T.A., Takagishi, T., Fukue, K., Kambe, T. and Fukada, T., 2017. Physiological roles of zinc transporters: molecular and genetic importance in zinc homeostasis. *The Journal of Physiological Sciences*, *67*(2), pp.283-301.

Harvey, W.T., Carabelli, A.M., Jackson, B., Gupta, R.K., Thomson, E.C., Harrison, E.M., Ludden, C., Reeve, R., Rambaut, A., Peacock, S.J. and Robertson, D.L., 2021. SARS-CoV-2 variants, spike mutations and immune escape. *Nature Reviews Microbiology*, *19*(7), pp.409-424.

Heinz, H., Lin, T.J., Kishore Mishra, R. and Emami, F.S., 2013. Thermodynamically consistent force fields for the assembly of inorganic, organic, and biological nanostructures: the INTERFACE force field. *Langmuir*, *29*(6), pp.1754-1765.

Hernik-Magoń, A., Fedorczyk, B., Dec, R., Puławski, W., Misicka, A. and Dzwolak, W., 2017. Effects of terminal capping on the fibrillation of short (L-Glu) n peptides. *Colloids and Surfaces B: Biointerfaces*, *159*, pp.861-868.

Herschlag, D. and Pinney, M.M., 2018. Hydrogen bonds: Simple after all?. *Biochemistry*, *57*(24), pp.3338-3352.

Hoffmann, M., Kleine-Weber, H., Schroeder, S., Krüger, N., Herrler, T., Erichsen, S., Schiergens, T.S., Herrler, G., Wu, N.H., Nitsche, A. and Müller, M.A., 2020. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *Cell, 181,* pp.271–280.

Hopkins, A.L., Groom, C.R. and Alex, A., 2004. Ligand efficiency: a useful metric for lead selection. *Drug discovery today*, *9*(10), pp.430-431.

Horowitz, S. and Trievel, R.C., 2012. Carbon-oxygen hydrogen bonding in biological structure and function. *Journal of Biological Chemistry*, *287*(50), pp.41576-41582.

Hsu, M.F., Kuo, C.J., Chang, K.T., Chang, H.C., Chou, C.C., Ko, T.P., Shr, H.L., Chang, G.G., Wang, A.H.J. and Liang, P.H., 2005. Mechanism of the Maturation Process of SARS-CoV 3CL Protease. *Journal of Biological Chemistry*, *280*(35), pp.31257-31266.

Humphrey, W., Dalke, A. and Schulten, K., 1996. VMD: visual molecular dynamics. *Journal of molecular graphics*, *14*(1), pp.33-38.

IMGT, 2021. Amino acids. [online] Available at: <http://www.imgt.org/IMGTeducation/Aide-memoire/_UK/aminoacids/IMGTclasses.html> [Accessed 03 January 2022].

International Monetary Fund and Capital Markets Department, 2020. *Global Financial Stability Report: Markets in the Time of COVID-19*. International Monetary Fund.

Ivry, S.L., Meyer, N.O., Winter, M.B., Bohn, M.F., Knudsen, G.M., O'Donoghue, A.J. and Craik, C.S., 2018. Global substrate specificity profiling of post-translational modifying enzymes. *Protein Science*, *27*(3), pp.584-594.

Jaghoori, M.M., Bleijlevens, B. and Olabarriaga, S.D., 2016. 1001 Ways to run AutoDock Vina for virtual screening. *Journal of computer-aided molecular design*, *30*(3), pp.237-249.

Johnson, Z.L. and Chen, J., 2017. Structural basis of substrate recognition by the multidrug resistance protein MRP1. *Cell*, *168*(6), pp.1075-1085.

Khan, S.A., Zia, K., Ashraf, S., Uddin, R. and Ul-Haq, Z., 2020. Identification of chymotrypsin-like protease inhibitors of SARS-CoV-2 via integrated computational approach. *Journal of Biomolecular Structure and Dynamics*, pp.1-10.

Kim, D., Lee, J.Y., Yang, J.S., Kim, J.W., Kim, V.N. and Chang, H., 2020. The architecture of SARS-CoV-2 transcriptome. *Cell.*

Klebe, G., 2006. Virtual ligand screening: strategies, perspectives and limitations. *Drug discovery today*, *11*(13-14), pp.580-594.

Kumar, S., Nyodu, R., Maurya, V.K. and Saxena, S.K., 2020b. Morphology, Genome Organization, Replication, and Pathogenesis of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2). In *Coronavirus Disease 2019 (COVID-19)* (pp. 23-31). Springer, Singapore.

Kumar, V., Dhanjal, J.K., Kaul, S.C., Wadhwa, R. and Sundar, D., 2020a. Withanone and caffeic acid phenethyl ester are predicted to interact with main protease (M^{pro}) of SARS-CoV-2 and inhibit its activity. *Journal of Biomolecular Structure and Dynamics*, pp.1-17.

Lamiable, A., Thévenet, P., Rey, J., Vavrusa, M., Derreumaux, P. and Tufféry, P., 2016. PEP-FOLD3: faster de novo structure prediction for linear peptides in solution and in complex. *Nucleic acids research*, *44*(W1), pp.W449-W454.

Landrum, G., 2013a. RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling. [online] Available at: <http://www.rdkit.org/RDKit_Overview.pdf> [Accessed 15 February 2021].

Landrum, G., 2013b. RDKit Documentation *Release 2019.09.1*. [online] Available at: <https://buildmedia.readthedocs.org/media/pdf/rdkit/latest/rdkit.pdf> [Accessed 15 February 2021].

Leach, A.R., Shoichet, B.K. and Peishoff, C.E., 2006. Prediction of protein−ligand interactions. Docking and scoring: successes and gaps. *Journal of medicinal chemistry*, *49*(20), pp.5851-5855.

Lee, C.C., Nayak, A., Sethuraman, A., Belfort, G. and McRae, G.J., 2007. A three-stage kinetic model of amyloid fibrillation. *Biophysical journal*, *92*(10), pp.3448-3458.

Lee, J., Worrall, L.J., Vuckovic, M., Rosell, F.I., Gentile, F., Ton, A.T., Caveney, N.A., Ban, F., Cherkasov, A., Paetzel, M. and Strynadka, N.C., 2020. Crystallographic structure of wild-type SARS-CoV-2 main protease acyl-enzyme intermediate with physiological C-terminal autoprocessing site. *Nature communications*, *11*(1), pp.1-9.

Lei, J., Kusov, Y. and Hilgenfeld, R., 2018. Nsp3 of coronaviruses: Structures and functions of a large multi-domain protein. *Antiviral research*, *149*, pp.58-74.

Lin, X., Li, X. and Lin, X., 2020. A review on applications of computational methods in drug screening and design. *Molecules*, *25*(6), p.1375.

Lindahl, E.R., 2008. Molecular dynamics simulations. In *Molecular modeling of proteins* (pp. 3-23). Humana Press.

Lippert, T. and Rarey, M., 2009. Fast automated placement of polar hydrogen atoms in protein-ligand complexes. *Journal of cheminformatics*, *1*(1), pp.1-12.

Liu, C., Zhou, Q., Li, Y., Garner, L.V., Watkins, S.P., Carter, L.J., Smoot, J., Gregg, A.C., Daniels, A.D., Jervey, S. and Albaiu, D., 2020. Research and development on therapeutic agents and vaccines for COVID-19 and related human coronavirus diseases. *ACS Central Science*, *6*, pp.315−331.

Liu, Y., Liu, J., Xia, H., Zhang, X., Fontes-Garfias, C.R., Swanson, K.A., Cai, H., Sarkar, R., Chen, W., Cutler, M. and Cooper, D., 2021. Neutralizing activity of BNT162b2-elicited serum. *New England Journal of Medicine*, *384*(15), pp.1466-1468.

Lobanov, M.Y., Bogatyreva, N.S. and Galzitskaya, O.V., 2008. Radius of gyration as an indicator of protein structure compactness. *Molecular Biology*, *42*(4), pp.623-628.

Lu, I.L., Mahindroo, N., Liang, P.H., Peng, Y.H., Kuo, C.J., Tsai, K.C., Hsieh, H.P., Chao, Y.S. and Wu, S.Y., 2006. Structure-based drug design and structural biology study of novel nonpeptide inhibitors of severe acute respiratory syndrome coronavirus main protease. *Journal of medicinal chemistry*, *49*(17), pp.5154-5161.

Margreitter, C., Petrov, D. and Zagrovic, B., 2013. Vienna-PTM web server: a toolkit for MD simulations of protein post-translational modifications. *Nucleic acids research*, *41*(W1), pp.W422-W426.

Maupetit, J., Derreumaux, P. and Tuffery, P., 2009. PEP-FOLD: an online resource for de novo peptide structure prediction. *Nucleic acids research*, *37*(suppl_2), pp.W498-W503.

McIntosh, K. and Perlman, S., 2015. Coronaviruses, including severe acute respiratory syndrome (SARS) and Middle East respiratory syndrome (MERS). *Mandell, Douglas, and Bennett's Principles and Practice of Infectious Diseases*, p.1928.

Meng, X.Y., Mezei, M. and Cui, M., 2014. Computational approaches for modeling GPCR dimerization. *Current pharmaceutical biotechnology*, *15*(10), pp.996-1006.

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.

Mengist, H.M., Dilnessa, T. and Jin, T., 2021. Structural basis of potential inhibitors targeting SARS-CoV-2 main protease. *Frontiers in Chemistry*, *9*.

Michaud-Agrawal, N., Denning, E.J., Woolf, T.B. and Beckstein, O., 2011. MDAnalysis: a toolkit for the analysis of molecular dynamics simulations. *Journal of computational chemistry*, *32*(10), pp.2319-2327.

Milewska, A., Kula-Pacurar, A., Wadas, J., Suder, A., Szczepanski, A., Dabrowska, A., Owczarek, K., Marcello, A., Ochman, M., Stacel, T. and Rajfur, Z., 2020. Replication of SARS-CoV-2 in human respiratory epithelium. *Journal of Virology, 94*(15), pp.1-7.

Mishra, A. and Dey, S., 2019. Molecular Docking Studies of a Cyclic Octapeptide-Cyclosaplin from Sandalwood. *Biomolecules*, *9*(11), p.740 (pp.1-18).

Morris, G.M., Huey, R., Lindstrom, W., Sanner, M.F., Belew, R.K., Goodsell, D.S. and Olson, A.J., 2009. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of computational chemistry*, *30*(16), pp.2785-2791.

Mousavizadeh, L. and Ghasemi, S., 2020. Genotype and phenotype of COVID-19: Their roles in pathogenesis. *Journal of Microbiology, Immunology and Infection*.

Muramatsu, T., Takemoto, C., Kim, Y.T., Wang, H., Nishii, W., Terada, T., Shirouzu, M. and Yokoyama, S., 2016. SARS-CoV 3CL protease cleaves its C-terminal autoprocessing site by novel subsite cooperativity. *Proceedings of the National Academy of Sciences*, *113*(46), pp.12997-13002.

Nhean, S., Varela, M.E., Nguyen, Y.N., Juarez, A., Huynh, T., Udeh, D. and Tseng, A.L., 2021. COVID-19: a review of potential treatments (corticosteroids, Remdesivir, tocilizumab, bamlanivimab/etesevimab, and casirivimab/imdevimab) and pharmacological considerations. *Journal of pharmacy practice*, p.08971900211048139.

Nicola, M., Alsafi, Z., Sohrabi, C., Kerwan, A., Al-Jabir, A., Iosifidis, C., Agha, M. and Agha, R., 2020. The socio-economic implications of the coronavirus and COVID-19 pandemic: a review. *International Journal of Surgery, 78,* pp.185-193.

Norel, R., Petrey, D., Wolfson, H.J. and Nussinov, R., 1999. Examination of shape complementarity in docking of unbound proteins. *Proteins: Structure, Function, and Bioinformatics*, *36*(3), pp.307-317.

O'Boyle, N.M., Banck, M., James, C.A., Morley, C., Vandermeersch, T. and Hutchison, G.R., 2011. Open Babel: An open chemical toolbox. *Journal of cheminformatics*, *3*(1), pp.1-14.

Orita, M., Ohno, K., Warizaya, M., Amano, Y. and Niimi, T., 2011. Lead generation and examples: opinion regarding how to follow up hits. *Methods in enzymology*, *493*, pp.383-419.

Orleans, L.A., is Vice, H. and Manchikanti, L., 2020. Expanded umbilical cord mesenchymal stem cells (UC-MSCs) as a therapeutic strategy in managing critically ill COVID-19 patients: the case for compassionate use. *Pain physician*, *23*, pp.71-83.

Pantsar, T. and Poso, A., 2018. Binding affinity via docking: fact and fiction. *Molecules*, *23*(8), p.1899.

Phillips, J.C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R.D., Kale, L. and Schulten, K., 2005. Scalable molecular dynamics with NAMD. *Journal of computational chemistry*, *26*(16), pp.1781-1802.

Porto, W.F., 2021. Virtual screening of peptides with high affinity for SARS-CoV-2 main protease. *Computers in biology and medicine*, *133*, p.104363.

Prieto-Martínez, F.D., Arciniega, M. and Medina-Franco, J.L., 2018. Molecular docking: current advances and challenges. *TIP. Revista especializada en ciencias químico-biológicas*, *21*.

Qi, E., Wang, D., Gao, B., Li, Y. and Li, G., 2017. Block-based characterization of protease specificity from substrate sequence profile. *BMC bioinformatics*, *18*(1), pp.1-9.

Quiroga, R. and Villarreal, M.A., 2016. Vinardo: A scoring function based on autodock vina improves scoring, docking, and virtual screening. *PloS one*, *11*(5), p.e0155183.

Sadoff, J., Le Gars, M., Shukarev, G., Heerwegh, D., Truyers, C., de Groot, A.M., Stoop, J., Tete, S., Van Damme, W., Leroux-Roels, I. and Berghmans, P.J., 2021. Interim results of a phase 1–2a trial of Ad26. COV2. S Covid-19 vaccine. *New England Journal of Medicine*, *384*(19), pp.1824-1835.

Salmaso, V. and Moro, S., 2018. Bridging molecular docking to molecular dynamics in exploring ligand-protein recognition process: an overview. *Frontiers in pharmacology*, *9*, p.923.

Salsbury Jr, F.R., 2010. Molecular dynamics simulations of protein dynamics and their relevance to drug discovery. *Current opinion in pharmacology*, *10*(6), pp.738-744.

Schauperl, M., Fuchs, J.E., Waldner, B.J., Huber, R.G., Kramer, C. and Liedl, K.R., 2015. Characterizing protease specificity: how many substrates do we need?. *PLOS one*, *10*(11), p.e0142658.

Schuler, L.D., Daura, X. and Van Gunsteren, W.F., 2001. An improved GROMOS96 force field for aliphatic hydrocarbons in the condensed phase. *Journal of computational chemistry*, *22*(11), pp.1205-1218.

Shang, J., Wan, Y., Luo, C., Ye, G., Geng, Q., Auerbach, A. and Li, F., 2020. Cell entry mechanisms of SARS-CoV-2. *Proceedings of the National Academy of Sciences*, *117*(21), pp.11727-11734.

Sheik Amamuddy, O., Verkhivker, G.M. and Tastan Bishop, O., 2020. Impact of early pandemic stage mutations on molecular dynamics of SARS-CoV-2 M^{pro}. *Journal of chemical information and modeling*, *60*(10), pp.5080-5102.

Shen, Y., Maupetit, J., Derreumaux, P. and Tufféry, P., 2014. Improved PEP-FOLD approach for peptide and miniprotein structure prediction. *Journal of chemical theory and computation*, *10*(10), pp.4745-4758.

Shi, D., An, X., Bai, Q., Bing, Z., Zhou, S., Liu, H. and Yao, X., 2019. Computational insight into the small molecule intervening PD-L1 dimerization and the potential structure-activity relationship. *Frontiers in Chemistry*, *7*, p.764 (pp.1-15).

Sironi, M., Hasnain, S.E., Phan, T., Luciani, F., Shaw, M.A., Sallum, M.A., Mirhashemi, M.E., Morand, S. and González-Candelas, F., 2020. SARS-CoV-2 and COVID-19: A genetic, epidemiological, and evolutionary perspective. *Infection, Genetics and Evolution, 84*, pp.1-15).

Skipper, L., 2005. Encyclopedia of Analytical Science (Second Edition, pp. 344-352). Elsevier.

Sliwoski, G., Kothiwale, S., Meiler, J. and Lowe, E.W., 2014. Computational methods in drug discovery. *Pharmacological reviews*, *66*(1), pp.334-395.

Sneha, P. and Doss, C.G.P., 2016. Molecular dynamics: new frontier in personalized medicine. *Advances in protein chemistry and structural biology*, *102*, pp.181-224.

Srikumar, P.S., Rohini, K. and Rajesh, P.K., 2014. Molecular dynamics simulations and principal component analysis on human laforin mutation W32G and W32G/K87A. *The protein journal*, *33*(3), pp.289-295.

Statista, 2022. Number of coronavirus (COVID-19) cases, recoveries, and deaths worldwide as of February 7, 2022. [online] Available at: <https://www.statista.com/statistics/1087466/covid19-cases-recoveries-deaths-world wide/> [Accessed 09 February 2022].

Suárez, D. and Díaz, N., 2020. SARS-CoV-2 main protease: A molecular dynamics study. *Journal of chemical information and modeling*, *60*(12), pp.5815-5831.

Swiderek, K. and Moliner, V., 2020. Revealing the Molecular Mechanisms of Proteolysis of SARS-CoV-2 M^{pro} from QM/MM Computational Methods. *Chemical Science*, pp.1-5.

Tanchuk, V.Y., Tanin, V.O., Vovk, A.I. and Poda, G., 2016. A new, improved hybrid scoring function for molecular docking and scoring based on AutoDock and AutoDock Vina. *Chemical biology & drug design*, *87*(4), pp.618-625.

Tobi, D. and Bahar, I., 2005. Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state. *Proceedings of the National Academy of Sciences*, *102*(52), pp.18908-18913.

Trott, O. and Olson, A.J., 2010. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, *31*(2), pp.455-461.

Tu, Y.F., Chien, C.S., Yarmishyn, A.A., Lin, Y.Y., Luo, Y.H., Lin, Y.T., Lai, W.Y., Yang, D.M., Chou, S.J., Yang, Y.P. and Wang, M.L., 2020. A review of SARS-CoV-2 and the ongoing clinical trials. *International journal of molecular sciences*, *21*(7), 2657, pp.1-19.

Uliana, F., Vizovišek, M., Acquasaliente, L., Ciuffa, R., Fossati, A., Frommelt, F., Goetze, S., Wollscheid, B., Gstaiger, M., De Filippis, V. and auf dem Keller, U., 2021. Mapping specificity, cleavage entropy, allosteric changes and substrates of blood proteases in a high-throughput screen. *Nature communications*, *12*(1), pp.1-18.

Ullrich, S. and Nitsche, C., 2020. The SARS-CoV-2 Main Protease as Drug Target. *Bioorganic & Medicinal Chemistry Letters, 30*(*19*), p.127377(pp.1-8).

van der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A.E. and Berendsen, H.J., 2005. GROMACS: fast, flexible, and free. *Journal of computational chemistry*, *26*(16), pp.1701-1718.

Vanommeslaeghe, K. and Guvench, O., 2014. Molecular mechanics. *Current pharmaceutical design*, *20*(20), pp.3281-3292.

Vasilevskaya, T., Khrenova, M.G., Nemukhin, A.V. and Thiel, W., 2016. Methodological aspects of QM/MM calculations: A case study on matrix metalloproteinase-2. *Journal of computational chemistry*, *37*(19), pp.1801-1809.

Venkatraman, P., Balakrishnan, S., Rao, S., Hooda, Y. and Pol, S., 2009. A sequence and structure based method to predict putative substrates, functions and regulatory networks of endo proteases. *PloS one*, *4*(5), p.e5700.

Vieira, T.F. and Sousa, S.F., 2019. Comparing AutoDock and Vina in ligand/decoy discrimination for virtual screening. *Applied Sciences*, *9*(21), p.4538.

Vizovišek, M., Vidmar, R., Drag, M., Fonović, M., Salvesen, G.S. and Turk, B., 2018. Protease specificity: towards in vivo imaging applications and biomarker discovery. *Trends in biochemical sciences*, *43*(10), pp.829-844.

Wang, Y., Wang, Y., Chen, Y. and Qin, Q., 2020. Unique epidemiological and clinical features of the emerging 2019 novel coronavirus pneumonia (COVID-19) implicate special control measures. *Journal of medical virology*, *92*(6), pp.568-576.

Weininger, D., 1988. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, *28*(1), pp.31-36.

Werner Reckien, E., 2017. xtb - An extended tight-binding semi-empirical program package. [online] Available at: <https://www.chemie.uni-bonn.de/pctc/mulliken-center/software/xtb/xtb> [Accessed 10 February 2022].

WHO, 2020. Coronavirus disease (COVID-19) pandemic. [online] Available at:<https://www.who.int/emergencies/diseases/novel-coronavirus-2019>[Accessed 14 July 2020].

WHO, 2022. Coronavirus disease (COVID-19) pandemic. [online] Available at: <https://www.who.int/emergencies/diseases/novel-coronavirus-2019> [Accessed 09 February 2022].

Woo, P.C., Huang, Y., Lau, S.K. and Yuen, K.Y., 2010. Coronavirus genomics and bioinformatics analysis. *viruses*, *2(8)*, pp.1804-1820.

Wu, C., Liu, Y., Yang, Y., Zhang, P., Zhong, W., Wang, Y., Wang, Q., Xu, Y., Li, M., Li, X. and Zheng, M., 2020a. Analysis of therapeutic targets for SARS-CoV-2 and discovery of potential drugs by computational methods. *Acta Pharmaceutica Sinica B*.

Wu, F., Zhao, S., Yu, B., Chen, Y.M., Wang, W., Song, Z.G., Hu, Y., Tao, Z.W., Tian, J.H., Pei, Y.Y. and Yuan, M.L., 2020b. A new coronavirus associated with human respiratory disease in China. *Nature*, *579*(7798), pp.265-269.

Yan, C.H., Faraji, F., Prajapati, D.P., Boone, C.E. and DeConde, A.S., 2020, April. Association of chemosensory dysfunction and Covid‐19 in patients presenting with influenza‐like symptoms. *International Forum of Allergy & Rhinology*, *10(7)*, pp.803-813.

Yang, L.Q., Sang, P., Tao, Y., Fu, Y.X., Zhang, K.Q., Xie, Y.H. and Liu, S.Q., 2014. Protein dynamics and motions in relation to their functions: several case studies and the underlying mechanisms. *Journal of Biomolecular Structure and Dynamics*, *32*(3), pp.372-393.

Yongye, A.B., Bender, A. and Martínez-Mayorga, K., 2010. Dynamic clustering threshold reduces conformer ensemble size while maintaining a biologically relevant ensemble. *Journal of computer-aided molecular design*, *24*(8), pp.675-686.

Zhang, L., Lin, D., Sun, X., Curth, U., Drosten, C., Sauerhering, L., Becker, S., Rox, K. and Hilgenfeld, R., 2020. Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors. *Science*, *368*(6489), pp.409-412.

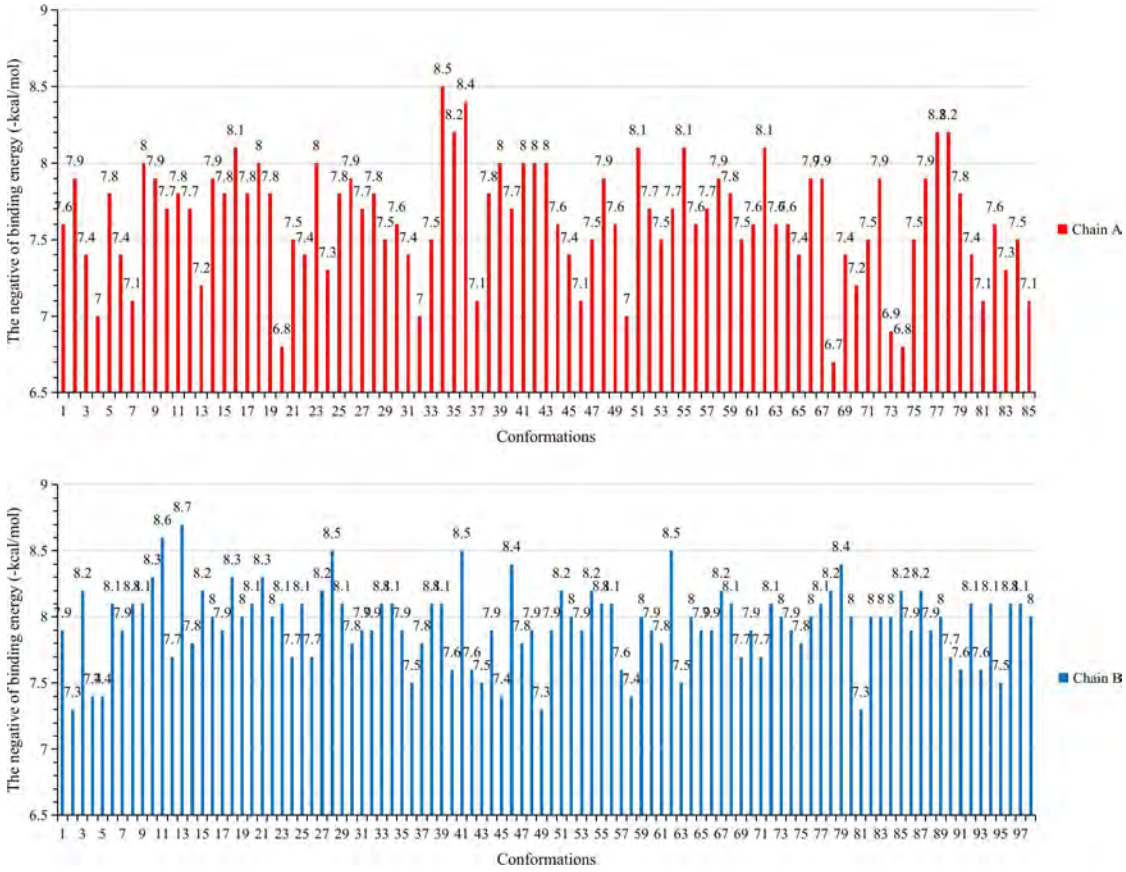
Zhou, J., Li, S., Leung, K.K., O'Donovan, B., Zou, J.Y., DeRisi, J.L. and Wells, J.A., 2020. Deep profiling of protease substrate specificity enabled by dual random and scanned human proteome substrate phage libraries. *Proceedings of the National Academy of Sciences*, *117*(41), pp.25464-25475.

Zuo, Z., Gandhi, N.S., Arndt, K.M. and Mancera, R.L., 2012. Free energy calculations of the interactions of c-Jun-based synthetic peptides with the c-Fos protein. *Biopolymers*, *97*(11), pp.899-909.

Zuo, Z.L., Guo, L. and Mancera, R.L., 2014. Free energy of binding of coiled-coil complexes with different electrostatic environments: the influence of force field polarisation and capping. *Natural products and bioprospecting*, *4*(5), pp.285-295.

Çakır, B., Okuyan, B., Şener, G. and Tunali-Akbay, T., 2021. Investigation of beta-lactoglobulin derived bioactive peptides against SARS-CoV-2 (COVID-19): in silico analysis. *European Journal of Pharmacology*, *891*, p.17378.

SUPPLEMENTARY MATERIAL



Supplementary figure 3.1. Preliminary docking studies to determine the protein chain to prioritise for docking studies. The negative of the docking scores of the RLQAAN conformers were plotted on bar graphs, showing the docking scores of chain A (red) and chain B (blue). Image was generated using WPS Spreadsheet 2019.

Supplementary table 3.1: Summary of docking results.

Substrate	Conformers Docked	Best Binding Conformer	ΔG (kcal/mol)
KLQAAA	100	17	-8.3
KLQAAD	100	100	-8.2
KLQAAE	100	3	-8.1
KLQAAF	100	27	-8.2
KLQAAG	57	27	-8.1
KLQAAM	19	32	-7.6
KLQAAN	10	32	-7.9
KLQAAQ	10	33	-7.9

KLQAAV	100	16	-8.1
KLQAEA	40	14	-8.3
KLQAED	40	10	-8.1
KLQAEE	24	16	-7.8
KLQAEF	24	18	-8.3
KLQAEG	40	12	-7.9
KLQAEM	10	31	-7.1
KLQAEN	40	24	-8.1
KLQAEQ	14	38	-7.7
KLQAEV	24	17	-8.0
KLQAGA	24	1	-8.4
KLQAGD	14	32	-7.7
KLQAGE	24	1	-7.8
KLQAGF	14	31	-8.0
KLQAGG	24	15	-8.2
KLQAGM	40	29	-8.0
KLQAGN	24	32	-8.2
KLQAGQ	24	15	-7.9
KLQAGV	14	33	-8.0
KLQAKA	14	35	-7.7
KLQAKD	14	1	-7.8
KLQAKE	10	33	-7.4
KLQAKF	40	22	-8.0
KLQAKG	10	31	-7.6
KLQAKM	24	15	-7.3
KLQAKN	40	19	-7.8
KLQAKQ	10	32	-7.4
KLQAKV	40	11	-7.9
KLQALA	10	32	-8.0
KLQALD	40	12	-8.1
KLQALE	24	1	-7.9
KLQALF	10	32	-7.9
KLQALG	24	1	-7.9
KLQALM	10	32	-7.4
KLQALN	35	16	-8.2
KLQALQ	14	34	-7.9

KLQALV	10	32	-7.9
KLQANA	24	19	-8.1
KLQAND	10	31	-8.1
KLQANE	40	15	-8.0
KLQANF	24	15	-8.4
KLQANG	19	1	-8.0
KLQANM	10	31	-7.4
KLQANN	24	16	-8.2
KLQANQ	10	31	-7.9
KLQANV	10	36	-8.0
KLQASA	24	16	-8.1
KLQASD	16	31	-8.2
KLQASE	10	35	-7.9
KLQASF	40	24	-8.1
KLQASG	24	18	-7.9
KLQASM	14	1	-7.5
KLQASN	35	1	-8.2
KLQASQ	10	34	-7.9
KLQASV	10	31	-8.0
KLQATA	10	32	-8.2
KLQATD	10	31	-8.1
KLQATE	10	33	-7.8
KLQATF	40	1	-8.2
KLQATG	10	34	-8.1
KLQATM	14	36	-7.7
KLQATN	10	31	-8.2
KLQATQ	10	31	-7.8
KLQATV	35	19	-8.0
KLQAVA	14	2	-8.1
KLQAVD	24	1	-8.2
KLQAVE	24	15	-7.9
KLQAVF	10	31	-8.0
KLQAVG	10	33	-8.1
KLQAVM	40	11	-7.7
KLQAVN	14	2	-8.1
KLQAVQ	24	15	-8.1

KLQAVV	24	15	-8.0
KLQSAA	40	27	-8.1
KLQSAD	14	33	-7.6
KLQSAE	14	1	-8.0
KLQSAF	24	15	-8.0
KLQSAG	35	2	-8.1
KLQSAM	40	14	-7.6
KLQSAN	24	17	-7.9
KLQSAQ	10	31	-7.5
KLQSAV	100	28	-8.0
KLQSEA	40	1	-7.9
KLQSED	10	32	-7.9
KLQSEE	40	21	-7.8
KLQSEF	24	2	-7.6
KLQSEG	35	21	-7.9
KLQSEM	24	17	-7.2
KLQSEN	24	33	-8.0
KLQSEQ	14	32	-7.3
KLQSEV	24	1	-7.8
KLQSGA	24	37	-8.0
KLQSGD	40	15	-8.0
KLQSGE	10	31	-7.4
KLQSGF	24	33	-7.7
KLQSGG	10	31	-8.0
KLQSGM	24	2	-7.5
KLQSGN	10	32	-7.9
KLQSGQ	10	34	-7.7
KLQSGV	10	31	-7.7
KLQSKA	24	18	-7.8
KLQSKD	14	2	-7.2
KLQSKE	10	33	-7.3
KLQSKF	40	26	-7.7
KLQSKG	14	33	-7.8
KLQSKM	10	31	-7.0
KLQSKN	24	15	-7.6
KLQSKQ	10	32	-7.3

KLQSKV	40	19	-7.6
KLQSLA	14	2	-7.8
KLQSLD	10	31	-7.5
KLQSLE	10	32	-7.4
KLQSLF	24	2	-7.7
KLQSLG	10	32	-8.1
KLQSLM	19	32	-7.4
KLQSLN	10	33	-8.1
KLQSLQ	14	32	-7.6
KLQSLV	10	32	-7.6
KLQSNA	14	1	-8.1
KLQSND	40	13	-8.2
KLQSNE	24	2	-7.5
KLQSNF	10	32	-7.6
KLQSNG	40	14	-8.0
KLQSNM	10	31	-7.3
KLQSNN	24	18	-8.0
KLQSNQ	24	17	-7.9
KLQSNV	24	16	-8.0
KLQSSA	10	32	-8.1
KLQSSD	24	15	-7.9
KLQSSE	40	1	-7.7
KLQSSF	40	15	-8.0
KLQSSG	40	22	-8.0
KLQSSM	21	16	-7.5
KLQSSN	40	19	-8.0
KLQSSQ	24	16	-7.7
KLQSSV	24	33	-7.9
KLQSTA	14	34	-8.0
KLQSTD	100	33	-8.0
KLQSTE	10	31	-7.6
KLQSTF	40	10	-8.0
KLQSTG	24	31	-7.9
KLQSTM	14	1	-7.6
KLQSTN	24	2	-7.9
KLQSTQ	10	35	-7.8

KLQSTV	10	31	-7.8
KLQSVA	10	32	-8.0
KLQSVD	10	35	-7.8
KLQSVE	10	32	-7.6
KLQSVF	24	16	-7.7
KLQSVG	10	32	-7.9
KLQSVM	40	18	-7.6
KLQSVN	40	14	-7.9
KLQSVQ	10	31	-7.7
KLQSVV	14	34	-7.8
MLQAAA	24	17	-7.9
MLQAAD	40	27	-7.8
MLQAAE	10	35	-7.5
MLQAAF	14	31	-8.0
MLQAAG	10	32	-7.8
MLQAAM	40	17	-7.7
MLQAAN	10	32	-7.8
MLQAAQ	19	2	-7.8
MLQAAV	40	1	-7.9
MLQAEA	10	32	-7.9
MLQAED	10	31	-7.6
MLQAEE	10	32	-7.6
MLQAEF	10	32	-7.9
MLQAEG	24	31	-7.6
MLQAEM	24	37	-7.7
MLQAEN	10	34	-7.9
MLQAEQ	14	1	-7.7
MLQAEV	10	31	-8.0
MLQAGA	24	15	-8.0
MLQAGD	10	32	-7.8
MLQAGE	10	33	-7.7
MLQAGF	10	31	-7.6
MLQAGG	14	2	-7.9
MLQAGM	40	30	-7.8
MLQAGN	14	1	-7.9
MLQAGQ	14	32	-7.8

MLQAGV	24	18	-7.9
MLQAKA	24	18	-7.8
MLQAKD	10	32	-7.5
MLQAKE	40	16	-7.4
MLQAKF	10	35	-7.5
MLQAKG	10	32	-7.4
MLQAKM	24	18	-7.2
MLQAKN	24	18	-7.5
MLQAKQ	40	11	-7.4
MLQAKV	10	31	-7.9
MLQALA	40	23	-8.5
MLQALD	10	34	-8.0
MLQALE	10	32	-7.6
MLQALF	40	12	-8.0
MLQALG	10	31	-7.7
MLQALM	10	35	-7.5
MLQALN	40	14	-7.9
MLQALQ	24	15	-8.0
MLQALV	10	31	-7.7
MLQANA	40	29	-8.2
MLQAND	24	36	-8.0
MLQANE	10	35	-7.9
MLQANF	10	31	-8.0
MLQANG	24	34	-7.8
MLQANM	14	1	-7.4
MLQANN	14	32	-7.9
MLQANQ	24	1	-8.1
MLQANV	40	19	-7.9
MLQASA	40	18	-7.8
MLQASD	24	18	-7.9
MLQASE	10	31	-7.5
MLQASF	10	32	-8.0
MLQASG	24	19	-7.7
MLQASM	24	39	-7.6
MLQASN	10	31	-7.8
MLQASQ	14	31	-7.6

MLQASV	24	18	-7.8
MLQATA	10	31	-7.9
MLQATD	35	23	-8.0
MLQATE	24	15	-7.7
MLQATF	10	33	-7.9
MLQATG	24	16	-7.8
MLQATM	10	31	-7.6
MLQATN	35	23	-8.1
MLQATQ	14	1	-7.7
MLQATV	10	33	-7.9
MLQAVA	14	1	-8.1
MLQAVD	24	16	-7.8
MLQAVE	40	14	-7.7
MLQAVF	40	11	-8.0
MLQAVG	24	15	-7.8
MLQAVM	19	1	-7.6
MLQAVN	14	1	-8.0
MLQAVQ	40	11	-7.7
MLQAVV	40	16	-7.8
MLQSAA	10	31	-7.8
MLQSAD	35	17	-7.6
MLQSAE	10	34	-7.5
MLQSAF	35	16	-7.8
MLQSAG	40	12	-7.7
MLQSAM	40	22	-7.5
MLQSAN	10	33	-7.8
MLQSAQ	24	16	-7.7
MLQSAV	14	31	-7.6
MLQSEA	14	37	-7.7
MLQSED	14	36	-7.9
MLQSEE	35	30	-7.6
MLQSEF	35	26	-7.9
MLQSEG	19	33	-7.7
MLQSEM	10	33	-7.6
MLQSEN	14	31	-7.8
MLQSEQ	14	36	-7.7

MLQSEV	40	12	-7.8
MLQSGA	14	1	-7.9
MLQSGD	24	1	-7.5
MLQSGE	24	19	-7.7
MLQSGF	24	33	-7.8
MLQSGG	40	10	-8.0
MLQSGM	14	2	-7.4
MLQSGN	24	17	-7.8
MLQSGQ	10	31	-7.6
MLQSGV	24	35	-7.8
MLQSKA	24	18	-7.5
MLQSKD	14	1	-7.4
MLQSKE	24	18	-7.4
MLQSKF	14	1	-7.4
MLQSKG	10	31	-7.4
MLQSKM	14	2	-7.2
MLQSKN	24	15	-7.6
MLQSKQ	24	15	-7.6
MLQSKV	14	1	-7.5
MLQSLA	35	30	-7.8
MLQSLD	10	32	-7.5
MLQSLE	10	32	-7.4
MLQSLF	10	31	-7.6
MLQSLG	40	15	-7.8
MLQSLM	35	22	-7.5
MLQSLN	100	30	-7.9
MLQSLQ	24	2	-7.6
MLQSLV	10	33	-7.8
MLQSNA	10	34	-7.6
MLQSND	24	19	-7.8
MLQSNE	14	1	-7.5
MLQSNF	40	17	-8.1
MLQSNG	10	32	-7.6
MLQSNM	40	17	-7.6
MLQSNN	10	36	-7.8
MLQSNQ	40	13	-7.8

MLQSNV	10	31	-7.6
MLQSSA	19	33	-7.9
MLQSSD	24	15	-7.6
MLQSSE	24	1	-7.6
MLQSSF	10	37	-7.7
MLQSSG	24	16	-7.8
MLQSSM	40	15	-7.5
MLQSSN	24	2	-7.8
MLQSSQ	40	10	-7.7
MLQSSV	10	31	-7.7
MLQSTA	14	1	-7.6
MLQSTD	35	22	-7.8
MLQSTE	10	31	-7.6
MLQSTF	10	32	-8.1
MLQSTG	10	31	-7.8
MLQSTM	24	32	-7.4
MLQSTN	24	36	-7.9
MLQSTQ	40	24	-7.7
MLQSTV	10	32	-7.7
MLQSVA	35	18	-8.0
MLQSVD	35	17	-7.9
MLQSVE	40	15	-7.6
MLQSVF	35	21	-8.3
MLQSVG	24	16	-7.7
MLQSVM	100	14	-7.6
MLQSVN	10	33	-7.7
MLQSVQ	10	32	-7.6
MLQSVV	10	33	-7.6
RLQAAA	10	31	-8.1
RLQAAD	19	1	-8.5
RLQAAE	10	35	-7.9
RLQAAF	35	20	-8.6
RLQAAG	10	33	-8.2
RLQAAM	24	19	-8.1
RLQAAN	100	11	-8.6
RLQAAQ	24	17	-8.2

RLQAAV	24	17	-8.4
RLQAEA	10	31	-7.9
RLQAED	10	33	-7.9
RLQAEE	40	15	-8.1
RLQAEF	10	32	-8.4
RLQAEG	14	1	-8.0
RLQAEM	35	16	-8.0
RLQAEN	10	33	-8.3
RLQAEQ	24	18	-8.2
RLQAEV	19	2	-8.3
RLQAGA	10	31	-8.6
RLQAGD	24	15	-8.4
RLQAGE	10	33	-7.6
RLQAGF	40	14	-8.5
RLQAGG	24	16	-8.4
RLQAGM	10	31	-7.6
RLQAGN	14	31	-7.9
RLQAGQ	40	13	-8.2
RLQAGV	10	31	-8.1
RLQAKA	40	23	-8.0
RLQAKD	10	31	-7.7
RLQAKE	24	1	-7.6
RLQAKF	14	1	-8.1
RLQAKG	14	1	-8.2
RLQAKM	14	2	-7.7
RLQAKN	10	33	-7.9
RLQAKQ	24	35	-7.8
RLQAKV	10	35	-8.0
RLQALA	10	32	-7.8
RLQALD	10	32	-8.1
RLQALE	35	21	-8.1
RLQALF	10	31	-8.1
RLQALG	40	14	-8.6
RLQALM	24	18	-7.8
RLQALN	14	33	-8.0
RLQALQ	40	11	-8.2

RLQALV	10	33	-8.3
RLQANA	10	32	-8.3
RLQAND	24	15	-8.1
RLQANE	10	31	-7.8
RLQANF	10	32	-8.5
RLQANG	14	1	-8.1
RLQANM	14	1	-8.0
RLQANN	14	1	-8.2
RLQANQ	10	31	-7.8
RLQANV	10	31	-8.2
RLQASA	10	36	-8.4
RLQASD	10	34	-7.9
RLQASE	40	14	-8.1
RLQASF	24	2	-8.4
RLQASG	24	15	-8.2
RLQASM	40	15	-8.0
RLQASN	14	31	-8.0
RLQASQ	40	2	-8.2
RLQASV	24	17	-8.1
RLQATA	35	1	-8.4
RLQATD	14	2	-8.1
RLQATE	100	54	-8.3
RLQATF	24	34	-8.7
RLQATG	24	15	-8.3
RLQATM	24	17	-8.0
RLQATN	10	35	-8.2
RLQATQ	14	1	-8.2
RLQATV	40	17	-8.4
RLQAVA	24	18	-8.2
RLQAVD	10	33	-8.2
RLQAVE	24	16	-8.2
RLQAVF	10	33	-8.1
RLQAVG	24	16	-8.4
RLQAVM	10	32	-7.8
RLQAVN	35	22	-8.6
RLQAVQ	10	33	-8.1

RLQAVV	40	20	-8.4
RLQSAA	10	34	-8.2
RLQSAD	35	22	-8.1
RLQSAE	40	17	-8.2
RLQSAF	10	35	-8.5
RLQSAG	10	31	-8.2
RLQSAM	19	35	-7.9
RLQSAN	19	32	-8.1
RLQSAQ	100	26	-8.2
RLQSAV	14	2	-8.0
RLQSEA	40	20	-8.1
RLQSED	10	31	-7.8
RLQSEE	24	15	-7.9
RLQSEF	10	31	-8.0
RLQSEG	24	17	-7.9
RLQSEM	14	33	-7.8
RLQSEN	10	32	-7.8
RLQSEQ	24	2	-8.2
RLQSEV	10	31	-7.9
RLQSGA	100	19	-8.7
RLQSGD	24	17	-7.8
RLQSGE	10	32	-7.8
RLQSGF	100	100	-8.4
RLQSGG	40	18	-8.5
RLQSGM	24	15	-7.9
RLQSGN	14	2	-8.0
RLQSGQ	24	17	-7.9
RLQSGV	40	15	-8.3
RLQSKA	24	18	-7.9
RLQSKD	24	18	-7.9
RLQSKE	10	35	-7.5
RLQSKF	10	33	-7.9
RLQSKG	14	1	-7.7
RLQSKM	14	31	-7.8
RLQSKN	10	33	-7.5
RLQSKQ	14	2	-7.6

RLQSKV	10	32	-8.0
RLQSLA	10	33	-7.8
RLQSLD	14	34	-8.0
RLQSLE	10	31	-7.6
RLQSLF	40	18	-8.1
RLQSLG	24	15	-8.0
RLQSLM	40	12	-7.9
RLQSLN	35	19	-8.1
RLQSLQ	40	20	-8.2
RLQSLV	40	17	-8.3
RLQSNA	40	1	-8.3
RLQSND	14	2	-7.9
RLQSNE	24	15	-8.0
RLQSNF	14	2	-8.2
RLQSNG	40	14	-8.3
RLQSNM	40	14	-7.9
RLQSNN	10	34	-8.0
RLQSNQ	24	16	-8.1
RLQSNV	10	32	-7.9
RLQSSA	100	29	-8.4
RLQSSD	14	1	-8.3
RLQSSE	40	18	-8.2
RLQSSF	10	31	-8.3
RLQSSG	24	16	-8.1
RLQSSM	10	31	-7.6
RLQSSN	24	1	-8.1
RLQSSQ	35	22	-8.4
RLQSSV	24	1	-8.1
RLQSTA	35	19	-8.4
RLQSTD	10	31	-7.8
RLQSTE	10	33	-7.9
RLQSTF	14	1	-8.7
RLQSTG	24	1	-8.3
RLQSTM	40	11	-7.8
RLQSTN	14	33	-8.2
RLQSTQ	24	16	-8.2

RLQSTV	24	31	-8.1
RLQSVA	14	1	-8.2
RLQSVD	10	31	-8.0
RLQSVE	40	27	-8.3
RLQSVF	24	1	-8.5
RLQSVG	24	15	-8.4
RLQSVM	24	16	-7.9
RLQSVN	40	15	-8.2
RLQSVQ	40	13	-8.3
RLQSVV	10	31	-8.2
TLQAAA	10	36	-8.2
TLQAAD	24	17	-8.2
TLQAAE	19	38	-7.9
TLQAAF	14	38	-8.1
TLQAAG	10	31	-8.3
TLQAAM	14	31	-7.7
TLQAAN	24	19	-8.4
TLQAAQ	35	23	-8.1
TLQAAV	10	31	-8.2
TLQAEA	24	17	-8.2
TLQAED	10	33	-7.7
TLQAEE	40	13	-8.0
TLQAEF	10	31	-7.8
TLQAEG	10	33	-8.2
TLQAEM	24	16	-7.7
TLQAEN	10	33	-8.4
TLQAEQ	24	19	-8.0
TLQAEV	40	1	-8.1
TLQAGA	14	32	-8.4
TLQAGD	10	33	-7.6
TLQAGE	24	19	-8.0
TLQAGF	35	32	-8.6
TLQAGG	24	32	-8.3
TLQAGM	10	34	-7.9
TLQAGN	40	19	-8.2
TLQAGQ	14	35	-8.2

TLQAGV	24	17	-8.4
TLQAKA	10	32	-7.6
TLQAKD	10	32	-7.6
TLQAKE	24	32	-7.6
TLQAKF	14	32	-7.8
TLQAKG	14	1	-8.0
TLQAKM	10	32	-7.5
TLQAKN	10	31	-7.5
TLQAKQ	35	22	-8.0
TLQAKV	10	35	-7.6
TLQALA	10	33	-7.9
TLQALD	24	1	-8.1
TLQALE	40	11	-7.9
TLQALF	10	33	-7.7
TLQALG	10	31	-7.9
TLQALM	10	31	-7.7
TLQALN	40	14	-8.2
TLQALQ	24	32	-7.9
TLQALV	35	16	-8.1
TLQANA	14	1	-7.9
TLQAND	24	39	-8.2
TLQANE	35	23	-7.9
TLQANF	14	1	-7.9
TLQANG	24	31	-8.2
TLQANM	40	18	-7.7
TLQANN	24	18	-8.5
TLQANQ	10	33	-8.2
TLQANV	24	19	-8.1
TLQASA	14	33	-8.1
TLQASD	10	33	-7.9
TLQASE	40	17	-8.2
TLQASF	10	31	-8.0
TLQASG	40	25	-8.2
TLQASM	24	39	-7.7
TLQASN	14	2	-8.3
TLQASQ	40	11	-8.0

TLQASV	24	17	-8.1
TLQATA	14	35	-8.4
TLQATD	10	33	-7.9
TLQATE	10	31	-8.3
TLQATF	24	32	-8.4
TLQATG	24	17	-8.2
TLQATM	10	32	-7.8
TLQATN	14	31	-8.2
TLQATQ	24	15	-8.2
TLQATV	10	32	-8.2
TLQAVA	10	31	-8.6
TLQAVD	40	13	-8.3
TLQAVE	24	15	-8.1
TLQAVF	10	34	-8.3
TLQAVG	40	24	-8.4
TLQAVM	10	31	-8.2
TLQAVN	10	34	-7.9
TLQAVQ	10	32	-8.0
TLQAVV	24	16	-8.4
TLQSAA	10	32	-8.2
TLQSAD	40	12	-8.1
TLQSAE	10	32	-8.2
TLQSAF	40	14	-8.4
TLQSAG	10	37	-8.0
TLQSAM	40	18	-7.7
TLQSAN	14	32	-8.1
TLQSAQ	24	16	-8.2
TLQSAV	40	22	-8.0
TLQSEA	40	24	-8.1
TLQSED	10	31	-7.9
TLQSEE	40	12	-8.2
TLQSEF	24	18	-7.9
TLQSEG	14	32	-8.0
TLQSEM	40	11	-7.9
TLQSEN	24	15	-8.0
TLQSEQ	24	16	-7.9

TLQSEV	10	32	-7.6
TLQSGA	40	1	-8.4
TLQSGD	24	16	-7.9
TLQSGE	24	17	-8.1
TLQSGF	10	32	-8.2
TLQSGG	10	33	-8.1
TLQSGM	14	32	-7.9
TLQSGN	24	38	-8.3
TLQSGQ	10	35	-7.9
TLQSGV	10	31	-8.3
TLQSKA	14	34	-8.0
TLQSKD	24	31	-7.9
TLQSKE	40	12	-8.1
TLQSKF	24	19	-7.4
TLQSKG	14	32	-8.0
TLQSKM	40	13	-7.7
TLQSKN	14	1	-7.4
TLQSKQ	35	1	-7.6
TLQSKV	24	16	-8.0
TLQSLA	14	33	-8.0
TLQSLD	40	10	-7.7
TLQSLE	24	17	-7.7
TLQSLF	40	14	-8.1
TLQSLG	14	1	-8.0
TLQSLM	10	32	-7.1
TLQSLN	40	14	-8.0
TLQSLQ	24	17	-8.0
TLQSLV	35	29	-8.0
TLQSNA	24	2	-7.9
TLQSND	10	33	-7.9
TLQSNE	24	31	-7.7
TLQSNF	24	15	-8.1
TLQSNG	10	31	-7.9
TLQSNM	40	12	-7.9
TLQSNN	14	2	-7.7
TLQSNQ	10	34	-7.7

TLQSNV	40	13	-7.9
TLQSSA	40	13	-7.9
TLQSSD	10	31	-7.7
TLQSSE	24	15	-7.9
TLQSSF	40	28	-7.9
TLQSSG	40	11	-8.1
TLQSSM	40	20	-7.9
TLQSSN	40	13	-8.1
TLQSSQ	24	16	-8.1
TLQSSV	40	14	-8.2
TLQSTA	40	12	-8.3
TLQSTD	40	11	-8.2
TLQSTE	24	18	-7.9
TLQSTF	10	31	-8.0
TLQSTG	40	14	-8.3
TLQSTM	10	31	-7.8
TLQSTN	35	20	-8.3
TLQSTQ	10	32	-7.8
TLQSTV	24	15	-8.2
TLQSVA	24	17	-8.1
TLQSVD	24	19	-8.0
TLQSVE	24	16	-7.7
TLQSVF	14	31	-7.9
TLQSVG	40	10	-8.1
TLQSVM	35	18	-7.8
TLQSVN	35	20	-8.4
TLQSVQ	10	32	-7.9
TLQSVV	24	15	-7.9
VLQAAA	10	31	-8.1
VLQAAD	10	32	-8.2
VLQAAE	35	15	-8.2
VLQAAF	35	25	-8.6
VLQAAG	24	34	-8.5
VLQAAM	24	15	-7.8
VLQAAN	10	34	-7.9
VLQAAQ	14	2	-8.1

VLQAAV	14	1	-8.2
VLQAEA	24	35	-8.2
VLQAED	35	15	-8.2
VLQAEE	40	10	-7.9
VLQAEF	40	13	-7.9
VLQAEG	10	32	-7.9
VLQAEM	35	22	-7.8
VLQAEN	40	14	-8.4
VLQAEQ	10	31	-7.8
VLQAEV	10	34	-8.0
VLQAGA	10	34	-8.5
VLQAGD	24	17	-8.1
VLQAGE	40	12	-8.0
VLQAGF	14	34	-8.2
VLQAGG	10	31	-8.1
VLQAGM	24	16	-7.9
VLQAGN	24	19	-8.3
VLQAGQ	10	34	-8.2
VLQAGV	10	32	-8.0
VLQAKA	24	32	-8.0
VLQAKD	10	31	-7.2
VLQAKE	10	31	-7.5
VLQAKF	24	17	-8.3
VLQAKG	10	34	-7.8
VLQAKM	24	38	-7.7
VLQAKN	40	23	-8.1
VLQAKQ	40	14	-7.8
VLQAKV	24	31	-7.8
VLQALA	10	32	-8.2
VLQALD	10	34	-8.2
VLQALE	24	37	-7.6
VLQALF	10	31	-8.0
VLQALG	10	31	-8.0
VLQALM	40	1	-7.6
VLQALN	10	33	-8.1
VLQALQ	35	17	-8.3

VLQALV	24	16	-8.1
VLQANA	24	1	-8.4
VLQAND	10	32	-8.4
VLQANE	10	34	-8.0
VLQANF	10	32	-8.1
VLQANG	24	32	-8.2
VLQANM	24	19	-7.8
VLQANN	10	33	-8.2
VLQANQ	14	1	-8.2
VLQANV	10	31	-8.1
VLQASA	10	36	-8.3
VLQASD	14	1	-7.9
VLQASE	10	33	-8.0
VLQASF	100	37	-8.2
VLQASG	24	19	-8.3
VLQASM	10	32	-7.5
VLQASN	24	31	-8.4
VLQASQ	40	28	-8.0
VLQASV	10	33	-8.1
VLQATA	40	11	-8.4
VLQATD	10	32	-8.2
VLQATE	24	18	-8.2
VLQATF	24	15	-8.1
VLQATG	14	2	-8.1
VLQATM	10	32	-7.7
VLQATN	10	31	-8.4
VLQATQ	10	31	-8.2
VLQATV	24	1	-8.2
VLQAVA	10	33	-8.2
VLQAVD	24	2	-8.3
VLQAVE	14	1	-8.0
VLQAVF	40	21	-8.6
VLQAVG	40	29	-8.4
VLQAVM	40	10	-7.9
VLQAVN	100	26	-8.4
VLQAVQ	24	15	-8.2

VLQAVV	40	22	-8.2
VLQSAA	24	16	-8.2
VLQSAD	24	18	-8.2
VLQSAE	10	31	-7.9
VLQSAF	10	34	-7.8
VLQSAG	24	2	-8.2
VLQSAM	24	16	-7.9
VLQSAN	24	17	-8.2
VLQSAQ	35	19	-8.1
VLQSAV	24	18	-8.3
VLQSEA	10	31	-7.9
VLQSED	24	31	-8.1
VLQSEE	24	15	-7.7
VLQSEF	10	31	-7.8
VLQSEG	10	33	-7.9
VLQSEM	24	31	-7.6
VLQSEN	14	2	-7.9
VLQSEQ	24	18	-7.9
VLQSEV	40	30	-8.0
VLQSGA	10	33	-8.2
VLQSGD	100	2	-8.1
VLQSGE	10	35	-7.8
VLQSGF	40	3	-8.4
VLQSGG	14	1	-8.3
VLQSGM	40	26	-8.0
VLQSGN	35	1	-8.2
VLQSGQ	10	32	-8.0
VLQSGV	24	15	-8.3
VLQSKA	24	17	-8.1
VLQSKD	35	16	-7.9
VLQSKE	10	33	-7.3
VLQSKF	35	19	-7.8
VLQSKG	10	31	-7.7
VLQSKM	24	39	-7.4
VLQSKN	40	15	-8.0
VLQSKQ	10	31	-7.5

VLQSKV	40	11	-7.9
VLQSLA	40	10	-8.3
VLQSLD	10	32	-8.1
VLQSLE	24	2	-7.6
VLQSLF	14	31	-7.9
VLQSLG	10	33	-8.5
VLQSLM	10	31	-7.8
VLQSLN	24	17	-8.0
VLQSLQ	10	34	-7.7
VLQSLV	35	31	-8.1
VLQSNA	14	36	-8.2
VLQSND	19	2	-7.8
VLQSNE	14	2	-7.7
VLQSNF	14	2	-8.0
VLQSNG	10	31	-8.2
VLQSNM	24	38	-7.7
VLQSNN	10	31	-8.1
VLQSNQ	14	31	-8.2
VLQSNV	24	33	-8.0
VLQSSA	10	38	-8.2
VLQSSD	10	31	-7.9
VLQSSE	10	31	-7.9
VLQSSF	10	33	-8.0
VLQSSG	24	16	-8.0
VLQSSM	10	32	-7.6
VLQSSN	10	31	-7.8
VLQSSQ	10	33	-7.8
VLQSSV	24	15	-8.0
VLQSTA	24	16	-8.4
VLQSTD	10	33	-8.1
VLQSTE	24	19	-8.2
VLQSTF	10	31	-7.8
VLQSTG	24	18	-8.0
VLQSTM	24	34	-7.7
VLQSTN	24	15	-8.2
VLQSTQ	40	17	-8.2

VLQSTV	24	33	-8.3
VLQSVA	14	2	-8.0
VLQSVD	40	15	-8.2
VLQSVE	10	31	-7.5
VLQSVF	24	35	-8.5
VLQSVG	40	11	-8.1
VLQSVM	10	31	-7.6
VLQSVN	24	37	-8.2
VLQSVQ	10	35	-8.1
VLQSVV	10	32	-7.9

Supplementary table 3.2: Intermolecular interactions of SARS-CoV-2 M^{pro} complexed with substrates RLQATF, RLQSGA and TLQSTF.

Substrate	Hydrogen Bonds*	van der Waals Interactions	Other Interactions*	Docking Score ($kcal.mol^{-1}$)
RLQATF	Thr24; Thr25; Thr26; Tyr54; Leu141; Asn142; Gly143; Ser144; Cys145; His164; Met165; Glu166; Asp187; Gln189;	Ser(A)1; His163; Phe140; Thr190; Asn119; Leu27; His41; Met49; Arg188; Pro52	Cys45; Gly143	
RLQSGA	Thr26; His41; Asn119; Phe140; Gly143; Ser144; Cys145; His164; Glu166; Gln189; Thr190	Ser(A)1; Thr24; Thr25; Leu27; Ser46; Leu141; Asn142; His163; Leu167; Pro168; His172; Asp187; Arg188; Ala191; Gln192	Met49; Met165	-8.7
RLQSTF	Ser(A)1; Thr25; Thr26; Ser46; Met49; Phe140; Leu141; Asn142; Gly143; Ser144; Cys145; His164; Glu166; Gln189	Thr24; Leu27; Tyr54; Tyr118; Asn119; Asp187; Arg188; Thr190; Ala191;	His41; Met165	

Blue: Catalytic His41; Red: Catalytic Cys145; Purple: Oxyanion Loop
*Hydrogen bonds include : Conventional, Carbon, Pi-Donor
* Other interactions includes: Alkyl, Pi-Alkyl, Pi-Sigma, Sulfur-X, Unfavourable Donor-Donor, Unfavourable Bump

Supplementary table 3.3: Intermolecular interactions of SARS-CoV-2 M^{pro} complexed with substrates RLQAAF, RLQAAN, RLQAGA, RLQALG, RLQAVN, TLQAGF, TLQAVA, VLQAAF and VLQAVF.

Substrate	Hydrogen Bonds*	van der Waals Interactions	Other Interactions*	Docking Score ($kcal.mol^{-1}$)
RLQAAF	Thr26; His41; Tyr54; Leu141; Gly143; Ser144; Cys145; His164; Met165;mGlu166; Gln189	Ser(A)1; Thr24; Thr25; Met49; Pro52; Asn119; Phe140; Asn142; His163; Arg188; Thr190	Leu27; His41; Cys145	
RLQAAN	Thr24; Thr25; Thr26; Ser46; Phe140; Asn142; Gly143; Ser144; Cys145; His163; Glu166; Gln189	Ser(A)1; Cys44; Thr45; Leu141; His164; Leu167; Pro168; His172; Asp187; Arg188; Thr190; Ala191; Gln192	Leu27; His41; Met49; Met165	
RLQAGA	Thr26; Asn119; Phe140; Gly143; Ser144; Cys145; His163; His164; Met165; Glu166; Gln189	Ser(A)1; Thr24; Thr25; Tyr54; Leu141; Asn142; His172; Phe181; Asp187; Arg188	Leu27; His41; Met49; Cys145; Met165; Gln189	
RLQALG	Thr26; Asn119; Phe140; Gly143; Ser144; Cys145; His164; Met165; Glu166; Gln189	Ser(A)1; Thr24; Thr25; Ser46; Leu141; Asn142; His163; His172; Phe181; Asp187; Arg188	Leu27; His41; Met29; Gln189	-8,6
RLQAVN	Thr24; Thr25; Thr26; Phe140; Gly143; Ser144; Cys145; His164; Glu166; Arg188; Gln189	Ser(A)1; Thr45; Ser46; Asn119; Leu141; Asn142; His163; Leu167; Pro168; His172; Phe181; aAsp187; Ala191; Gln192	Leu27; His41; Met49; Met165; Thr190	
TLQAGF	Thr26; Asn119; Phe140; Asn142; Gly143; Ser144; Glu166; Gln189	Ser(A)1; Thr24; Thr25; Tyr118; Leu141; His163; His164; Leu167; Pro168; His172; Asp187; Arg188; Thr190; Ala191; Gln192	Leu27; His41; Met49; Cys145; Met165	
TLQAVA	Thr24; Thr25; Thr26; His41; Phe140; Gly143; Ser144; Cys145; Glu166; Met165; Gln189	Ser(A)1; Leu27; Ser46; Leu141; Asn142; His163; His164; Phe181; Val186; Asp187; Arg188	His41; Met49; Cys145; Met165	

VLQAAF	Thr26; Phe140; Gly143; Ser144; Cys145; Glu166; Gln189	Ser(A)1; Thr24; Thr25; Ser46; Asn119; Leu141; Asn142; His163; His164; His172; Asp187; Arg188	Leu27; His41; Met49; Cys145; Met165
VLQAVF	Thr24; Thr26; Phe140; Gly143; Ser144; Cys145; His163; His164; Met165; Glu166; Gln189	Ser(A)1; Thr25; Ser46; Tyr118; Leu141; Asn142; His172; Asp187; Arg188	Leu27; His41; Met49; Met165

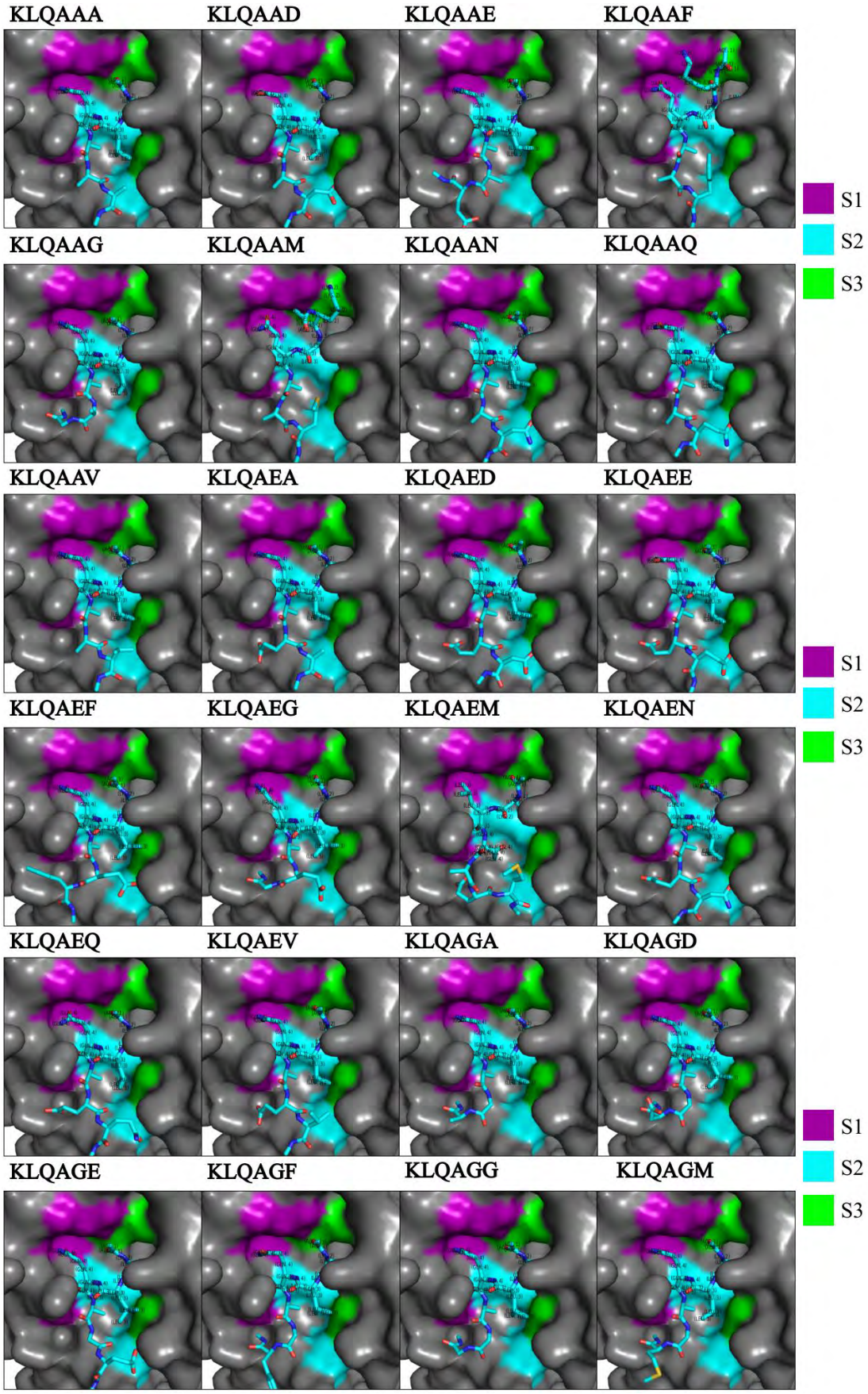
Blue: Catalytic His41; Red: Catalytic Cys145; Purple: Oxyanion Loop
*Hydrogen bonds include : Conventional, Carbon, Pi-Donor
* Other interactions includes: Alkyl, Pi-Alkyl, Pi-Sigma, Sulfur-X, Unfavourable Donor-Donor, Unfavourable Bump

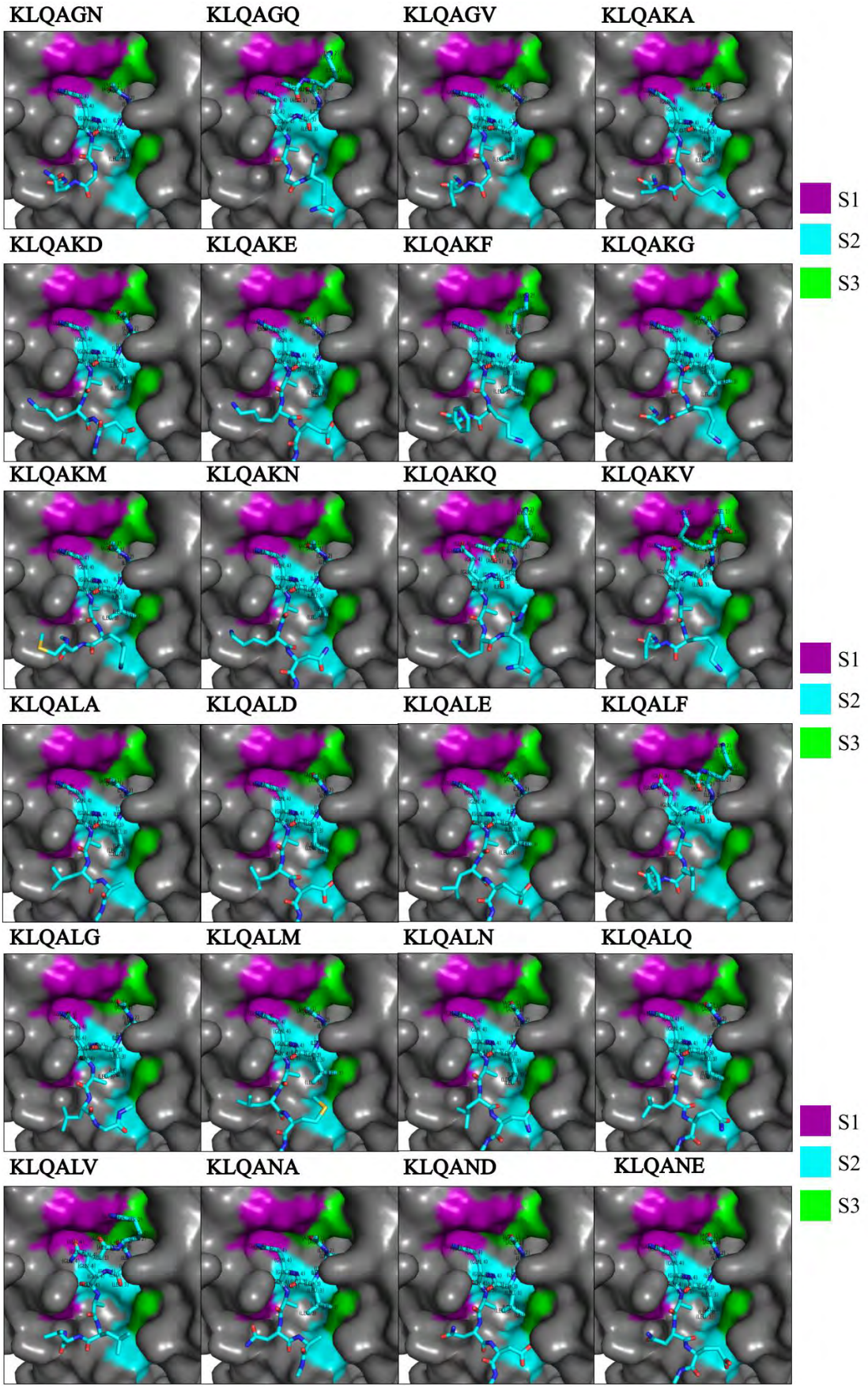
Supplementary table 3.4: Intermolecular interactions of SARS-CoV-2 M^{pro} in complexed with substrates KLQSKM, KLQAEM, TLQSLM, KLQSEM, KLQSKD, MLQAKM, MLQSKM and VLQAKD.

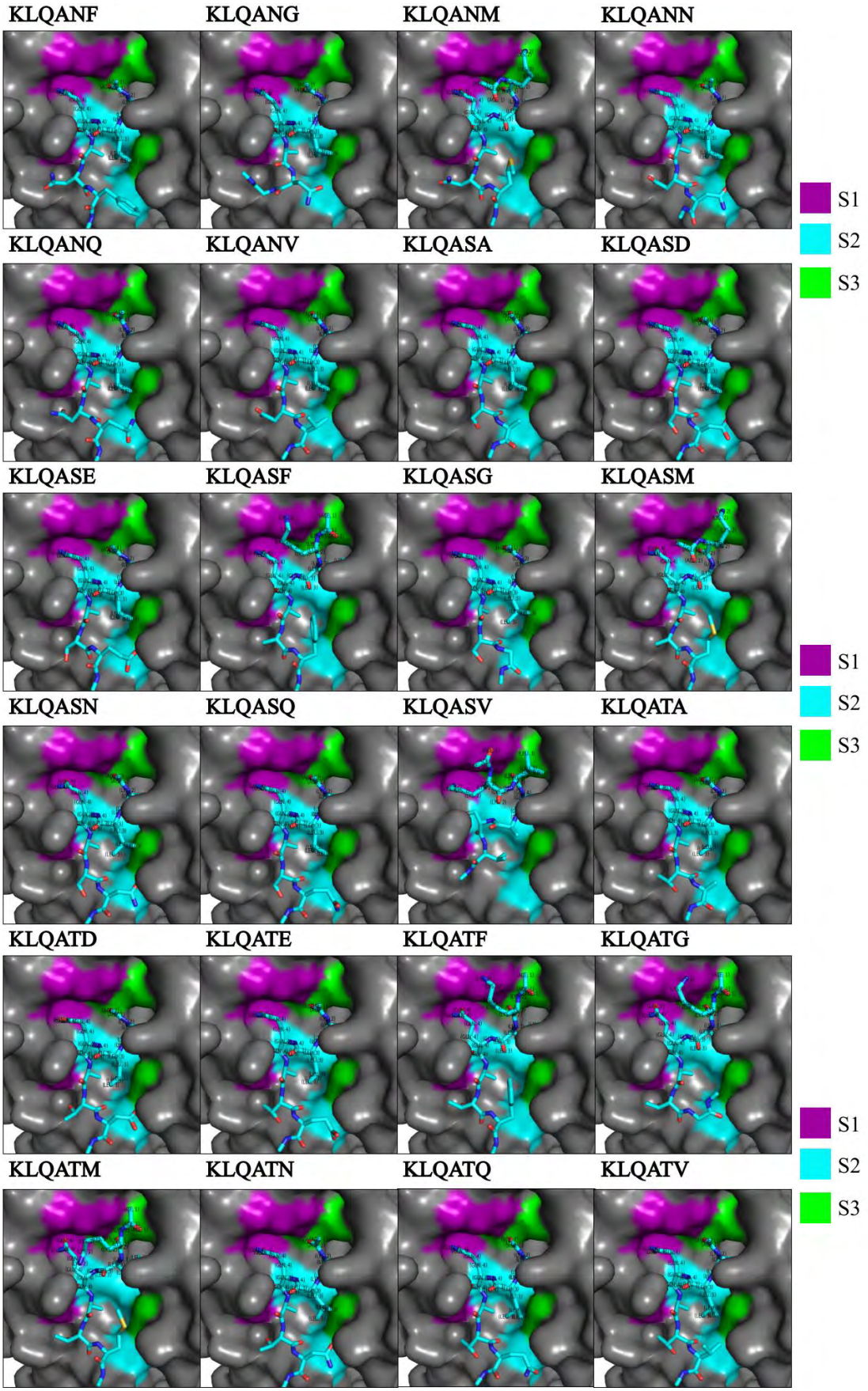
Substrate	Hydrogen Bonds*	van der Waals Interactions	Other Interactions*	Docking Score (kcal.mol^{-1})
KLMSKM	Thr24; Thr26; Phe140; Gly143; Ser144; Cys145; His163; Glu166; Gln189	Ser(A)1; Thr25; Leu27; Ser46; Asn119; Leu141; Asn142; His164; Leu167; Pro168; His172; Asp187; Arg188; Thr190; Ala191	His41; Met49; Met165	-7.0
KLQAEM	Thr24; Thr25; Thr26; His41; Tyr54; Asn119; Gly143; Cys145; Glu166; Asp187; Gln189	Leu27; Thr45; Ser46; Pro52; Phe140; Leu141; Asn142; Ser144; His163; His164; Met165; His172; Arg188	His41; Met49	-7.1
TLQSLM	His41; Leu141; Gly143; Ser144; Cys145; Glu166; Gln189	Thr25; Thr26; Ser46; Asn119; Phe140; Asn142; His163; His164; Met165; Leu167; Pro168; Arg188; Thr190	Leu27; His41; Met49; Cys145	
KLQSEM	Thr24; Thr26; Asn119; Leu141; Gly143; Ser144; Cys145; His163; Met165; Glu166; Gln189; Thr190	Thr25; Leu27; Thr45; Ser46; Phe140; Asn142; His164; Leu167; Pro168; His172; Asp187; Arg188; Ala191; Gln192	His41; Met49; Met165	-7.2
KLQSKD	Thr24; Thr26; Asn119; Phe140; Asn142; Gly143; Ser144; Cys145; His163; Glu166; Gln189	Thr25; Leu27; Tyr188; Asn119; Leu141; His164; Leu167; Pro168; His172; Asp187; Arg188; Thr190; Ala191	His41; Met49; Met165	
MLQAKM	Thr24; Thr26; Leu141; Gly143; Ser144; Cys145; His163; His164; Glu166; Gln189	Thr25; Ser46; Tyr118; Phe140; Asn142; Met165; Asp187; Arg188; Thr190; Ala191	Leu27; His41; Met49; Cys145	
MLQSKM	Thr24; Thr25; Thr26; His41; Phe140; Asn142; Gly143; Ser144; His163; Glu166; Gln189	Ser(A)1; Leu27; Ser46; Tyr118; Asn119; Leu141; Cys145; His164; Leu167; Pro168; His172; Asp187; Thr190; Ala191; Gln192	His41; Met49; Met165	

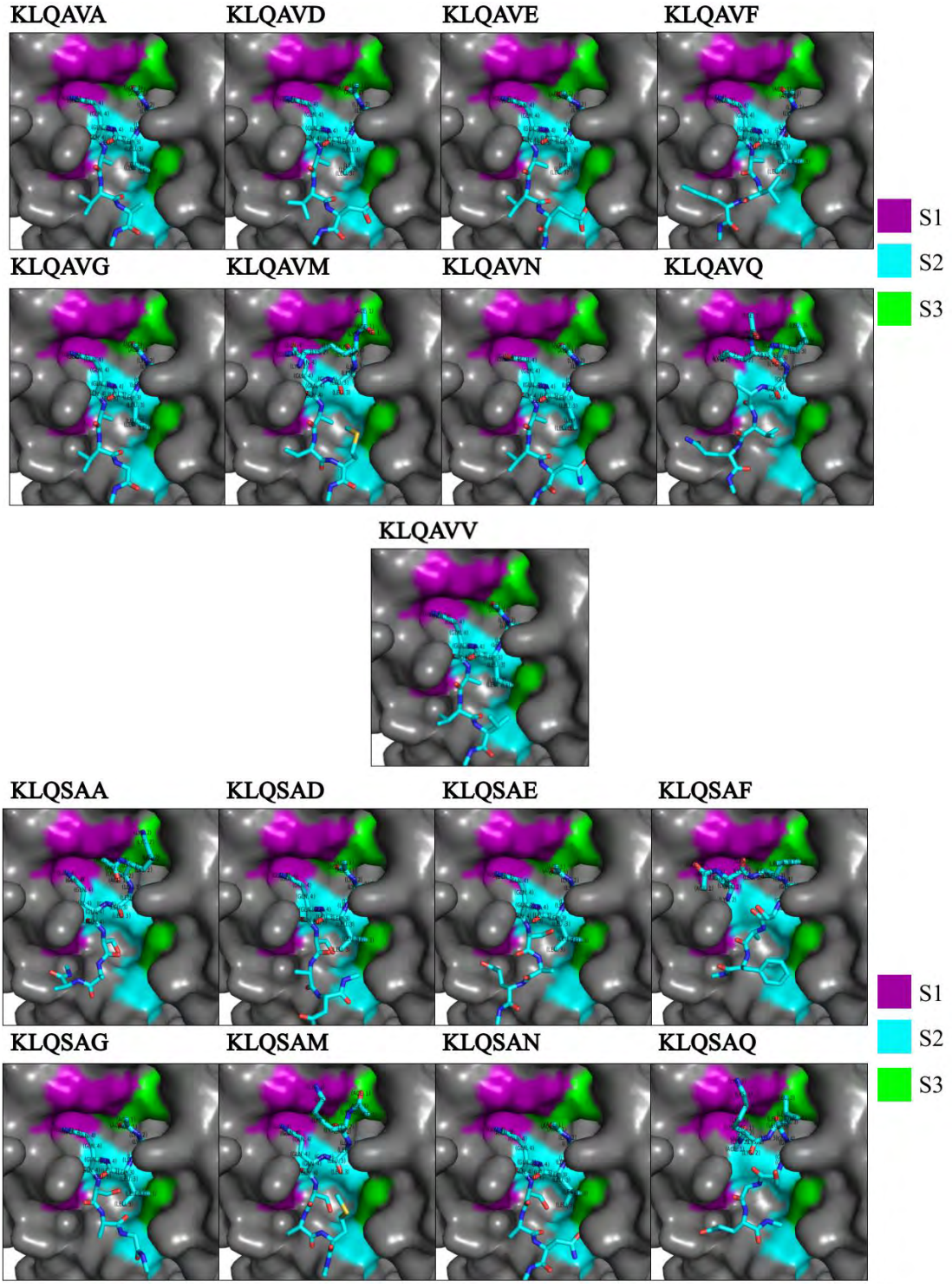
VLQAKD	His41; Phe140; Asn142; Gly143; His163; His164; Glu166; His172; Gln189	Ser(A)1; Thr25; Thr26; Ser46; Tyr54; Leu141; Ser144; Leu167; Pro168; Phe181; Asp187; Arg188; Thr190	Leu27; Met49; Cys145; Met165; Ala191

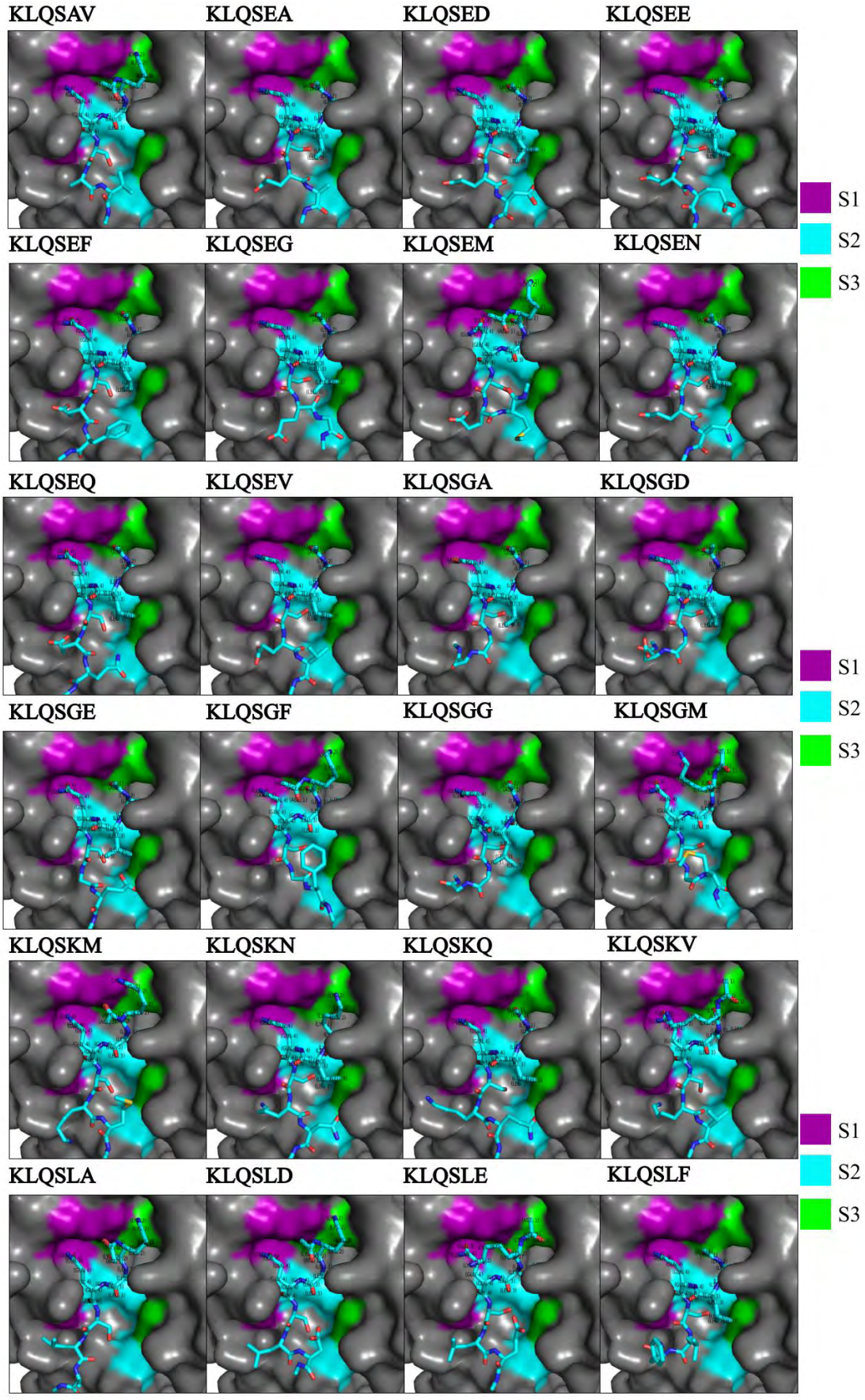
Blue: Catalytic His41; Red: Catalytic Cys145; Purple: Oxyanion Loop
*Hydrogen bonds include : Conventional, Carbon, Pi-Donor
* Other interactions includes: Alkyl, Pi-Alkyl, Pi-Sigma, Sulfur-X, Unfavourable Donor-Donor, Unfavourable Bump

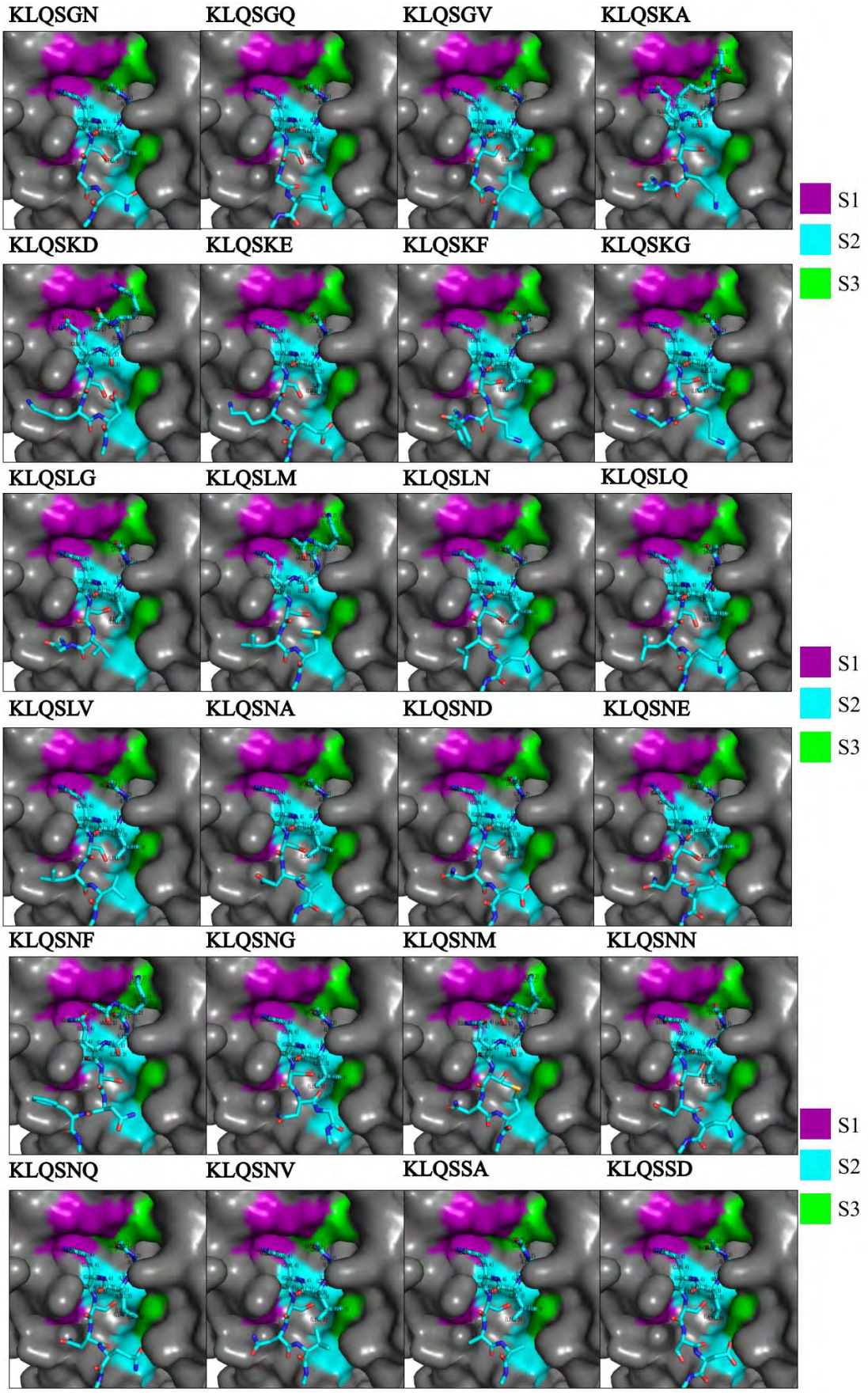

KLQAAA
KLQAAD
KLQAAE
KLQAAF
KLQAAG
KLQAAM
KLQAAN
KLQAAQ
KLQAAV
KLQAEA
KLQAED
KLQAEE
KLQAEF
KLQAEG
KLQAEM
KLQAEN
KLQAEQ
KLQAEV
KLQAGA
KLQAGD
KLQAGE
KLQAGF
KLQAGG
KLQAGM
S1
S2
S3
S1
S2
S3
S1
S2
S3

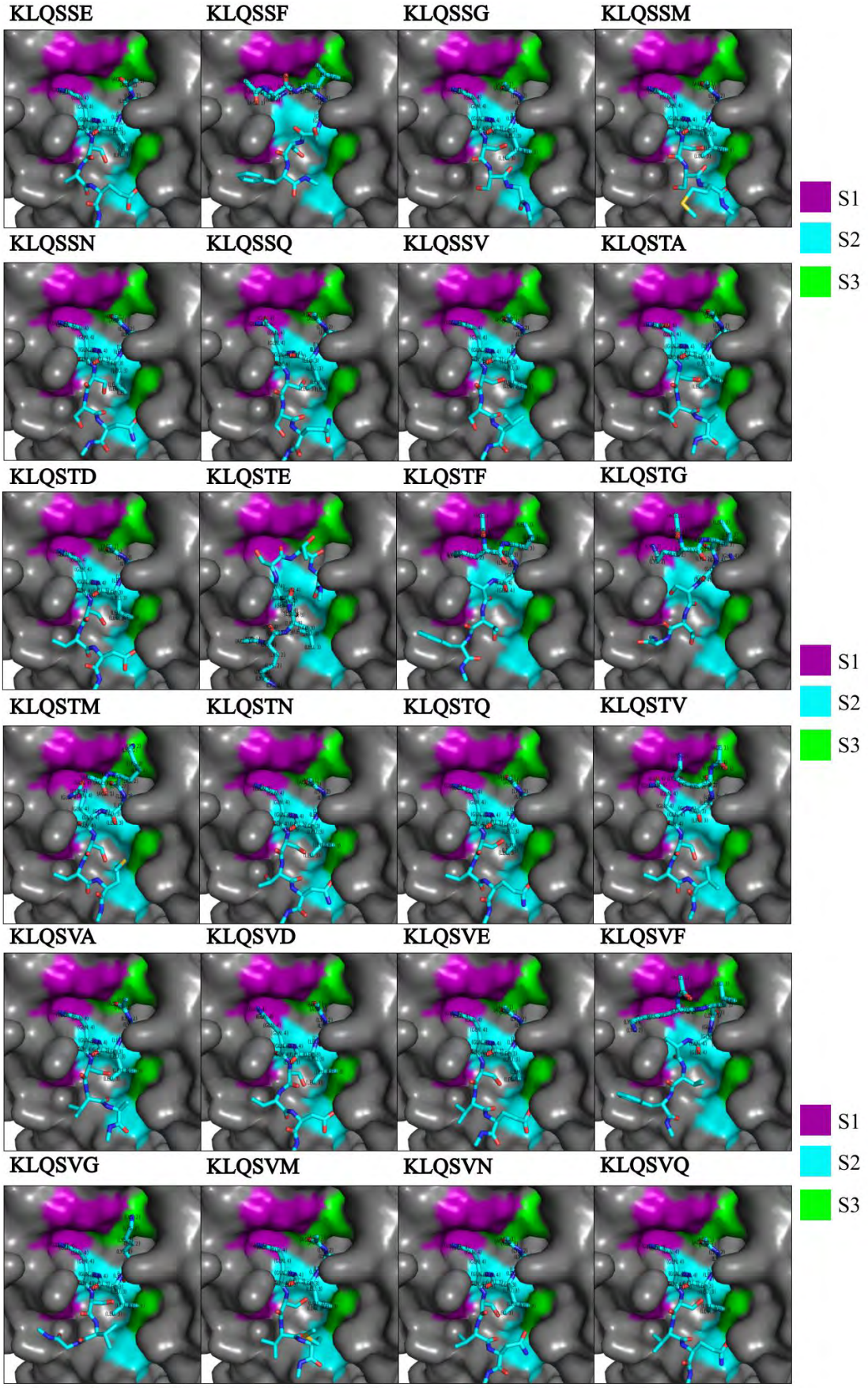

KLQAGN
KLQAGQ
KLQAGV
KLQAKA
KLQAKD
KLQAKE
KLQAKF
KLQAKG
KLQAKM
KLQAKN
KLQAKQ
KLQAKV
KLQALA
KLQALD
KLQALE
KLQALF
KLQALG
KLQALM
KLQALN
KLQALQ
KLQALV
KLQANA
KLQAND
KLQANE
S1
S2
S3


KLQANF
KLQANG
KLQANM
KLQANN
KLQANQ
KLQANV
KLQASA
KLQASD
KLQASE
KLQASF
KLQASG
KLQASM
KLQASN
KLQASQ
KLQASV
KLQATA
KLQATD
KLQATE
KLQATF
KLQATG
KLQATM
KLQATN
KLQATQ
KLQATV
S1
S2
S3
S1
S2
S3
S1
S2
S3

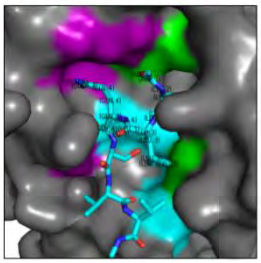

KLQAVA
KLQAVD
KLQAVE
KLQAVF
KLQAVG
KLQAVM
KLQAVN
KLQAVQ
KLQAVV
KLQSAA
KLQSAD
KLQSAE
KLQSAF
KLQSAG
KLQSAM
KLQSAN
KLQSAQ
S1
S2
S3
S1
S2
S3


KLQSAV
KLQSEA
KLQSED
KLQSEE
KLQSEF
KLQSEG
KLQSEM
KLQSEN
KLQSEQ
KLQSEV
KLQSGA
KLQSGD
KLQSGE
KLQSGF
KLQSGG
KLQSGM
KLQSKM
KLQSKN
KLQSKQ
KLQSKV
KLQSLA
KLQSLD
KLQSLE
KLQSLF
S1
S2
S3


KLQSGN
KLQSGQ
KLQSGV
KLQSKA
KLQSKD
KLQSKE
KLQSKF
KLQSKG
KLQSLG
KLQSLM
KLQSLN
KLQSLQ
KLQSLV
KLQSNA
KLQSND
KLQSNE
KLQSNF
KLQSNG
KLQSNM
KLQSNN
KLQSNQ
KLQSNV
KLQSSA
KLQSSD
S1
S2
S3

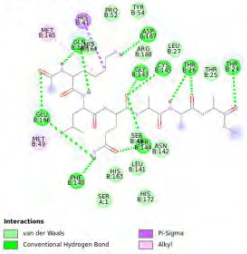

KLQSSE
KLQSSF
KLQSSG
KLQSSM
S1
S2
S3
KLQSSN
KLQSSQ
KLQSSV
KLQSTA
KLQSTD
KLQSTE
KLQSTF
KLQSTG
S1
S2
S3
KLQSTM
KLQSTN
KLQSTQ
KLQSTV
KLQSVA
KLQSVD
KLQSVE
KLQSVF
S1
S2
S3
KLQSVG
KLQSVM
KLQSVN
KLQSVQ

KLQSVV

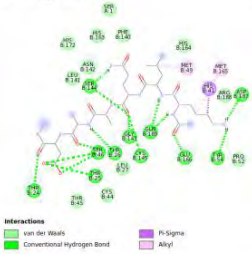


Supplementary figure 3.2. Confirmation of SARS-CoV-2 M^{pro} substrate recognition in binding poses for KLQ* substrates**. The surface of SARS-CoV-2 M^{pro} (PDB ID:6XHM) showing docked substrates and substrate binding subsites color-coded as follows: Purple: S1, Cyan: S2; Green: S3. The images were generated using PyMOL.

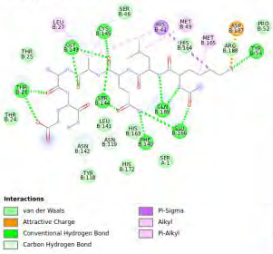
KLQAAA



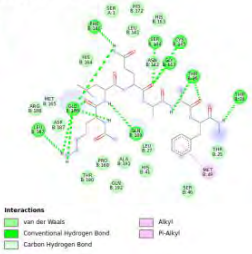
KLQAAD



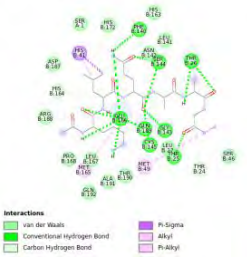
KLQAAE



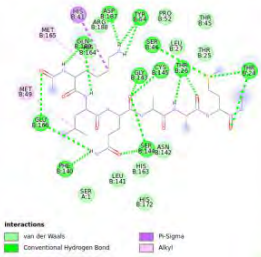
KLQAAF



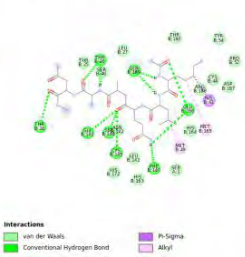
KLQAAG



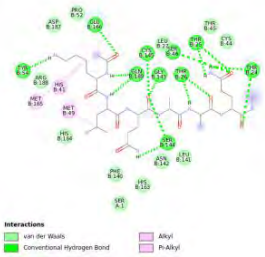
KLQAAM



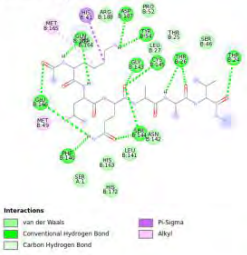
KLQAAN



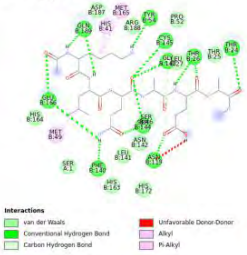
KLQAAQ



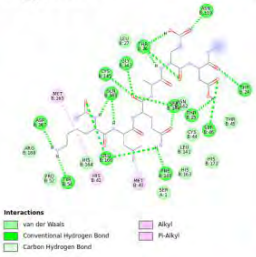
KLQAAV



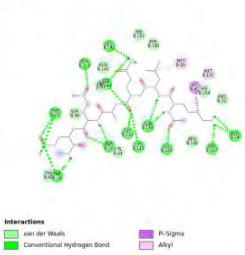
KLQAEA



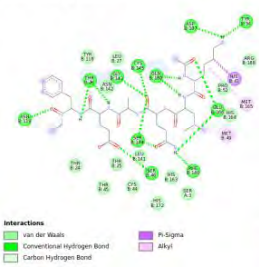
KLQAED



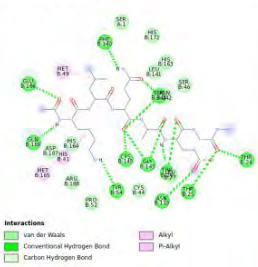
KLQAEE



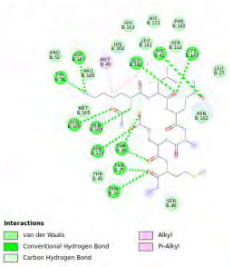
KLQAEF



KLQAEG

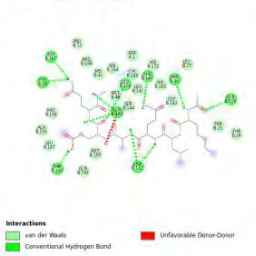


KLQAEM

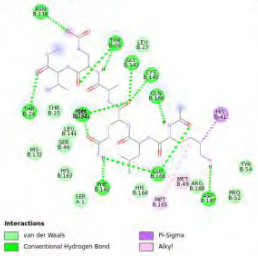


KLQAEN

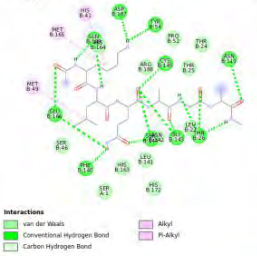
KLQAEQ



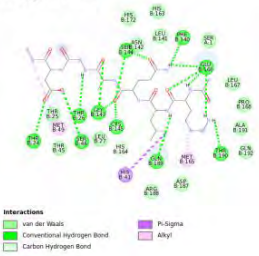
KLQAEV



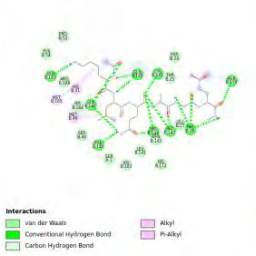
KLQAGA



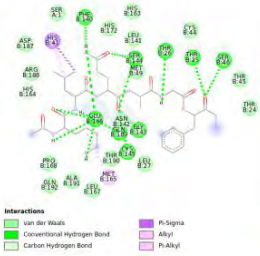
KLQAGD



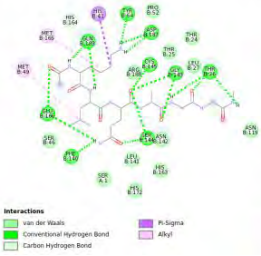
KLQAGE



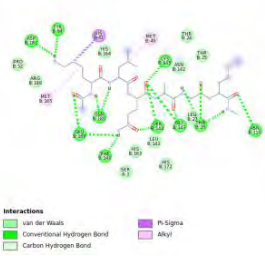
KLQAGF



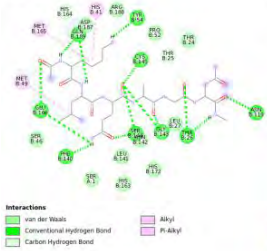
KLQAGG



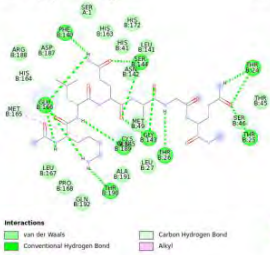
KLQAGM



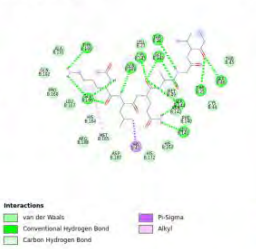
KLQAGN



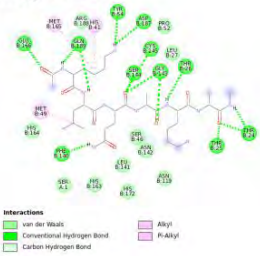
KLQAGQ



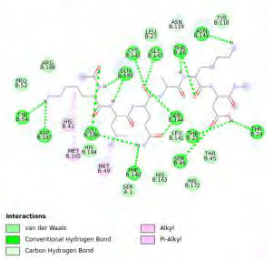
KLQAGV



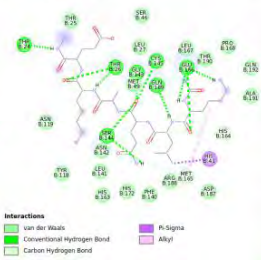
KLQAKA



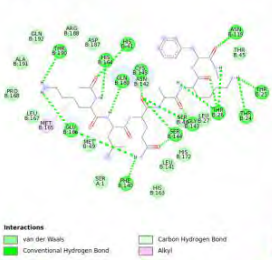
KLQAKD



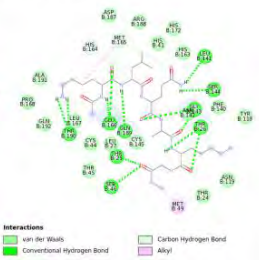
KLQAKE



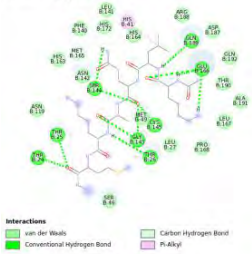
KLQAKF



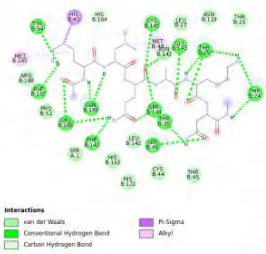
KLQAKG



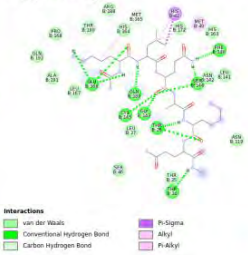
KLQAKM



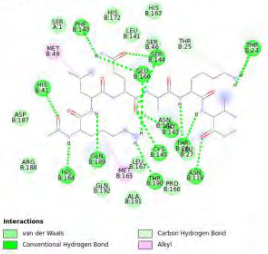
KLQAKN



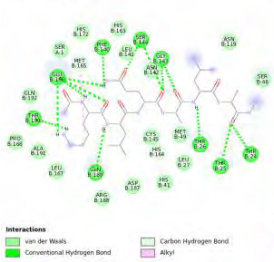
KLQAKQ



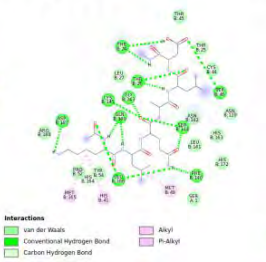
KLQAKV



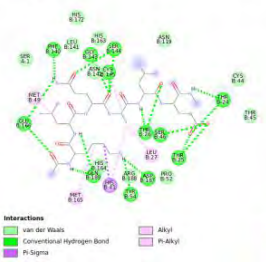
KLQALA



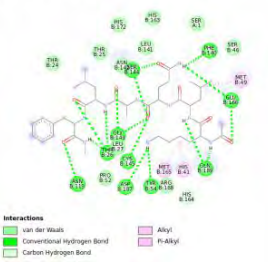
KLQALD



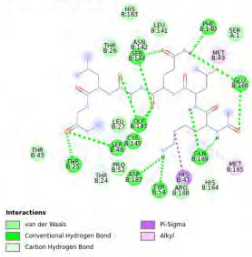
KLQALE



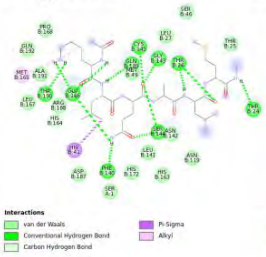
KLQALF



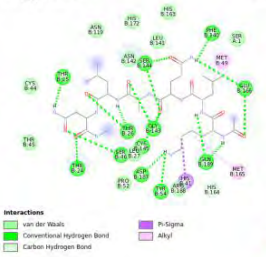
KLQALG



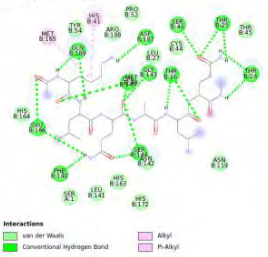
KLQALM



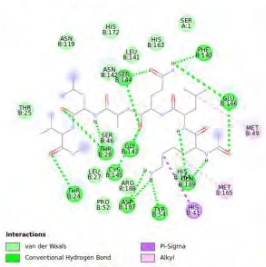
KLQALN



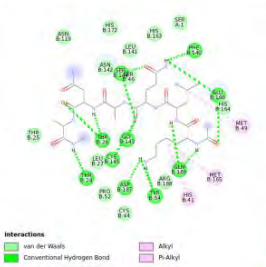
KLQALQ



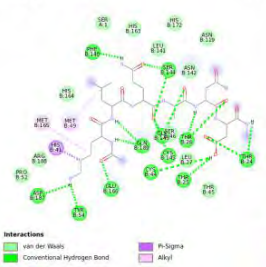
KLQALV



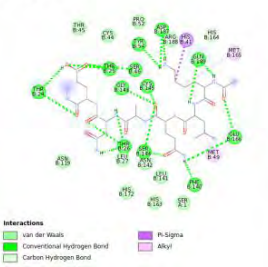
KLQANA



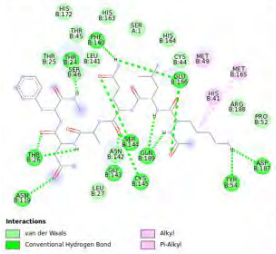
KLQAND



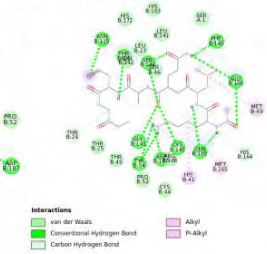
KLQANE



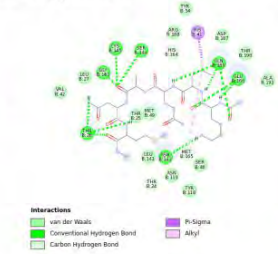
KLQANF



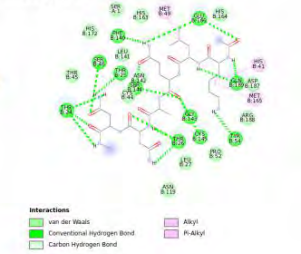
KLQANG



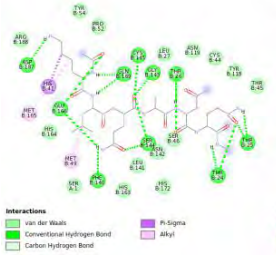
KLQANM



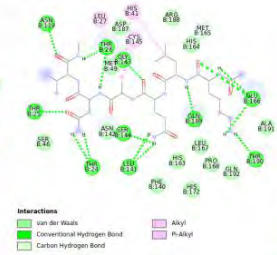
KLQANN



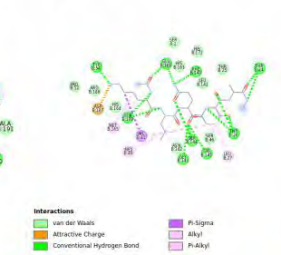
KLQANQ



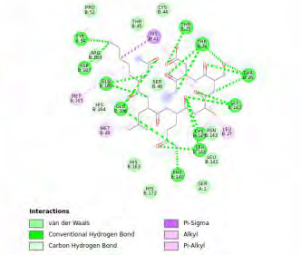
KLQANV



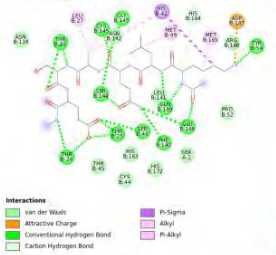
KLQASA



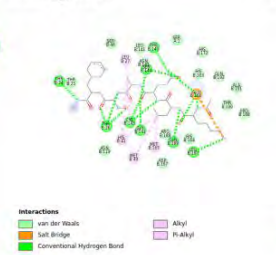
KLQASD



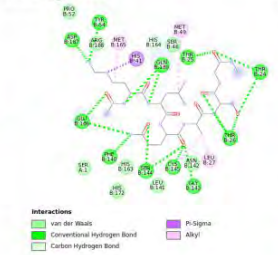
KLQASE



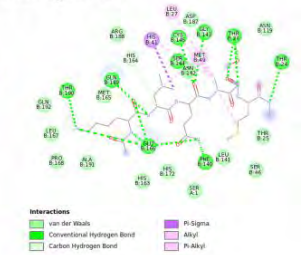
KLQASF



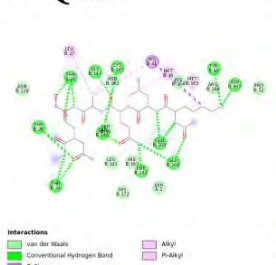
KLQASG



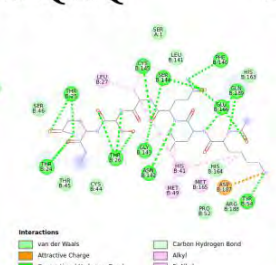
KLQASM



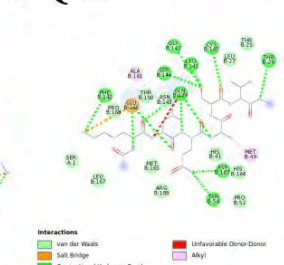
KLQASN



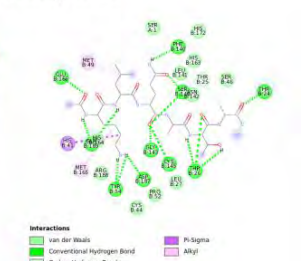
KLQASQ



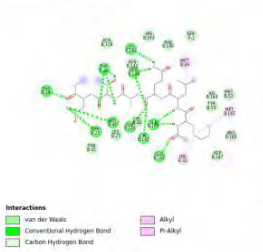
KLQASV



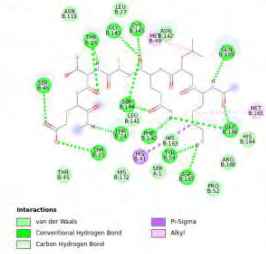
KLQATA



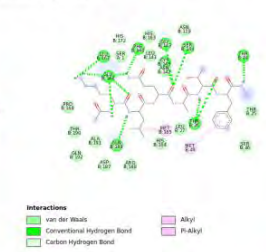
KLQATD



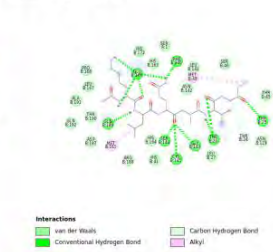
KLQATE



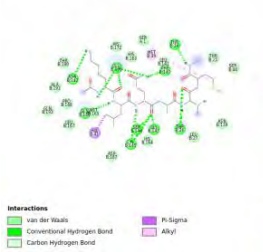
KLQATF



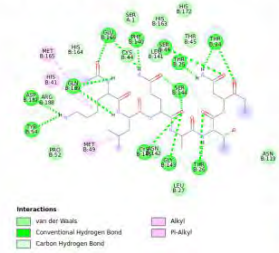
KLQATG



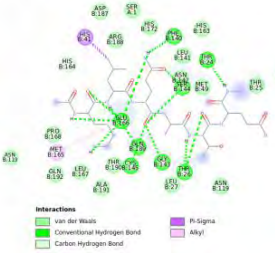
KLQATM



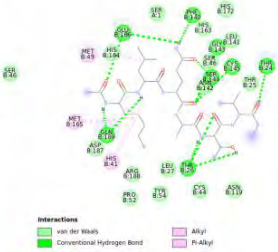
KLQATN



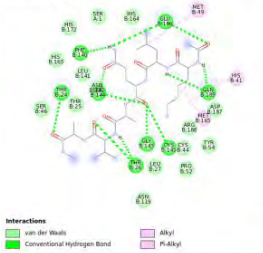
KLQATQ



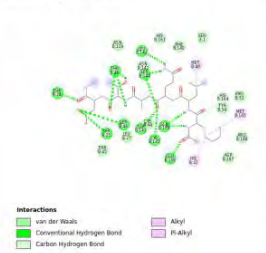
KLQATV



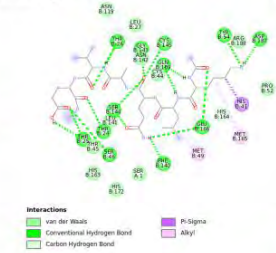
KLQAVA



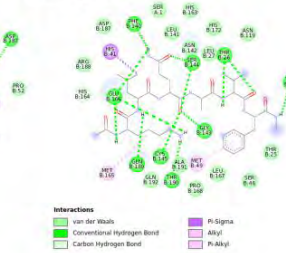
KLQAVD



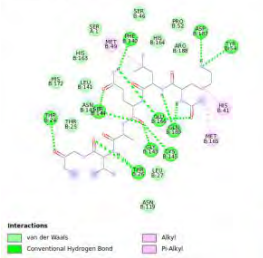
KLQAVE



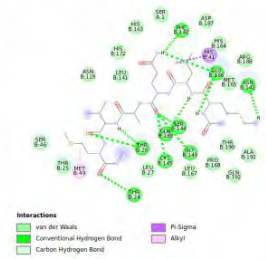
KLQAVF



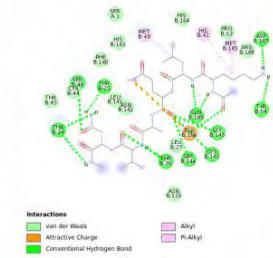
KLQAVG



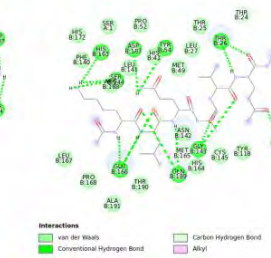
KLQAVM



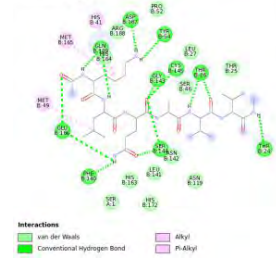
KLQAVN



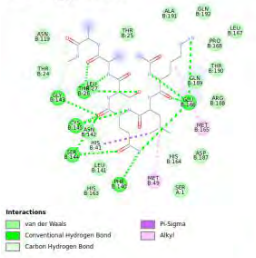
KLQAVQ



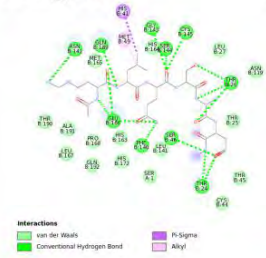
KLQAVV



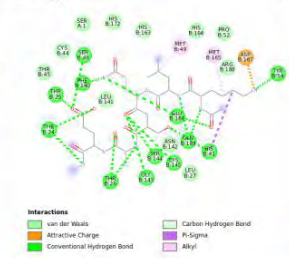
KLQSAA



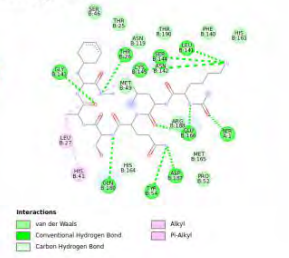
KLQSAD



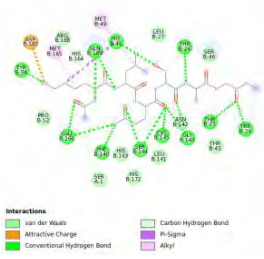
KLQSAE



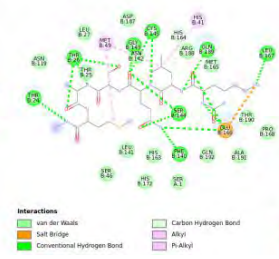
KLQSAF



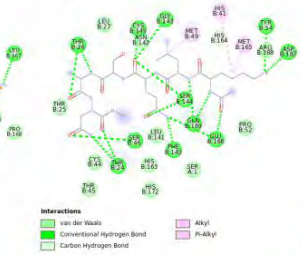
KLQSAG



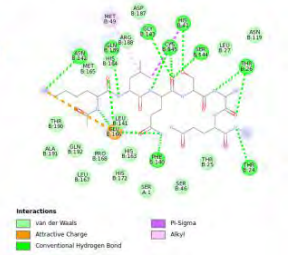
KLQSAM



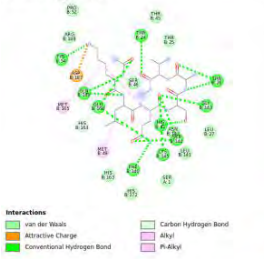
KLQSAN



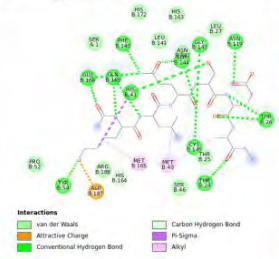
KLQSAQ



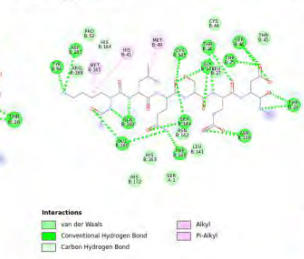
KLQSAV



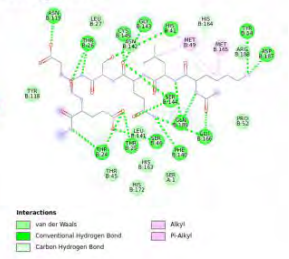
KLQSEA



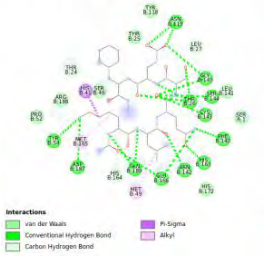
KLQSED



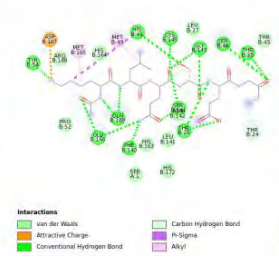
KLQSEE



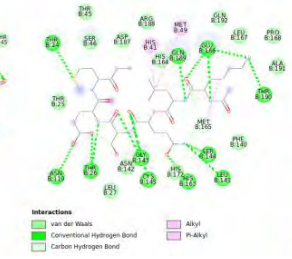
KLQSEF



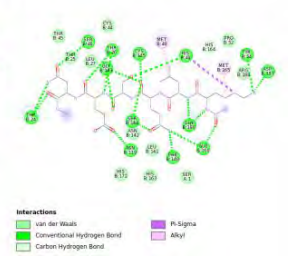
KLQSEG



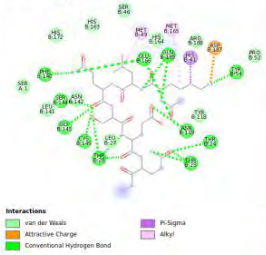
KLQSEM



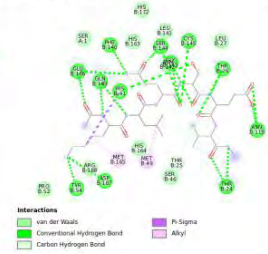
KLQSEN



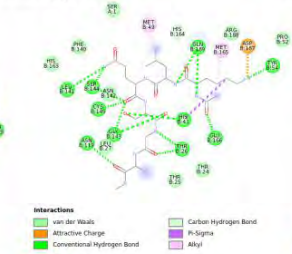
KLQSEQ



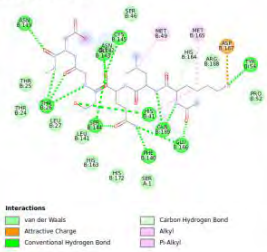
KLQSEV



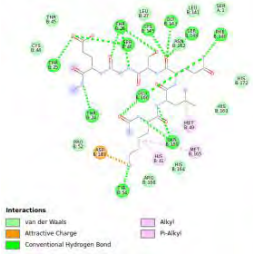
KLQSGA



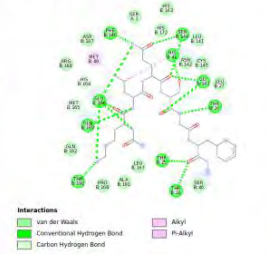
KLQSGD



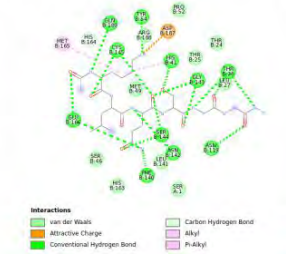
KLQSGE



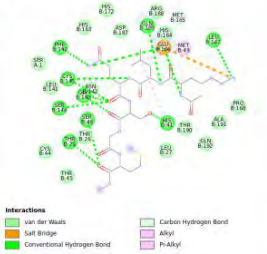
KLQSGF



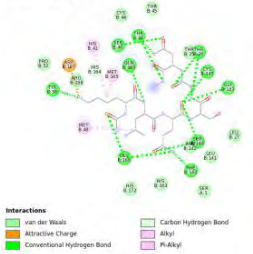
KLQSGG



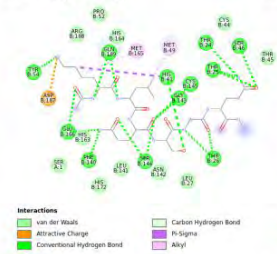
KLQSGM



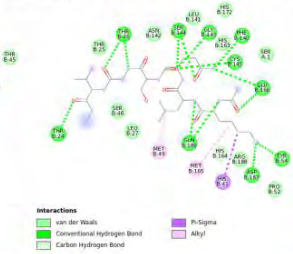
KLQSGN



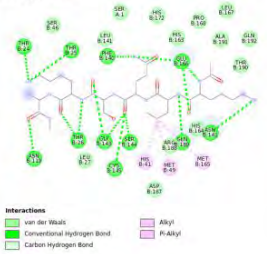
KLQSGQ



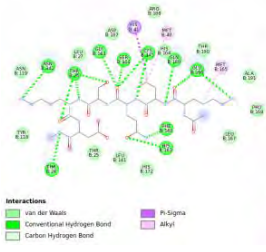
KLQSGV



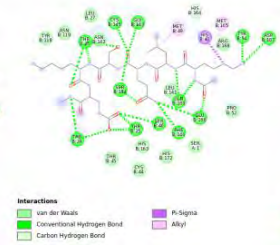
KLQSKA



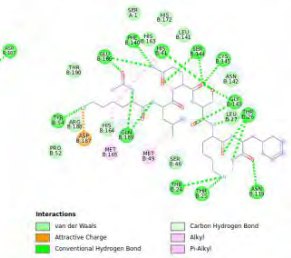
KLQSKD



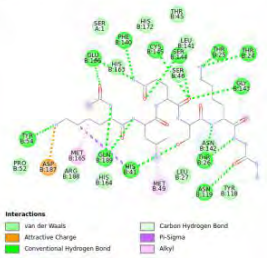
KLQSKE



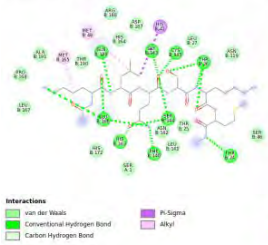
KLQSKF



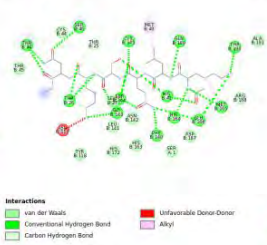
KLQSKG



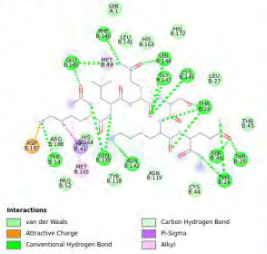
KLQSKM



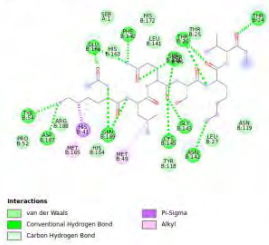
KLQSKN



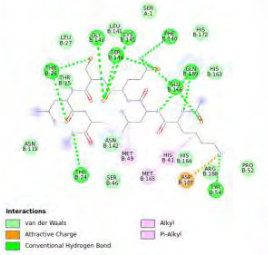
KLQSKQ



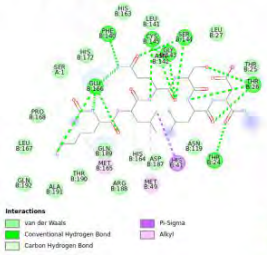
KLQSKV



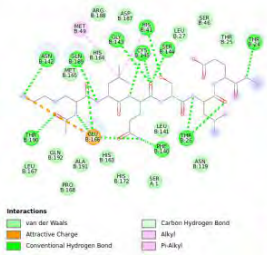
KLQSLA



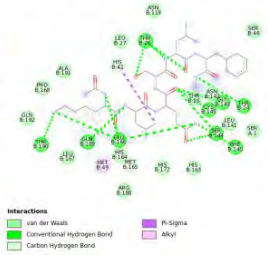
KLQSLD



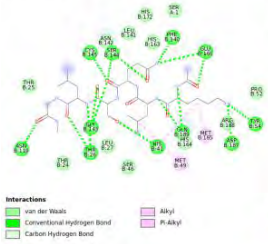
KLQSLE



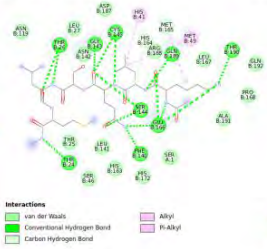
KLQSLF



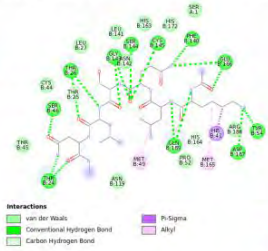
KLQSLG



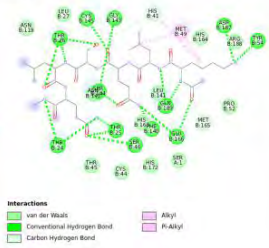
KLQSLM



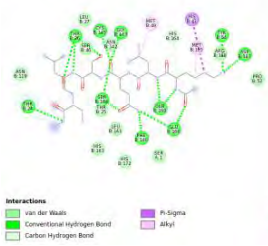
KLQSLN



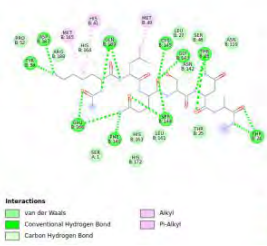
KLQSLQ



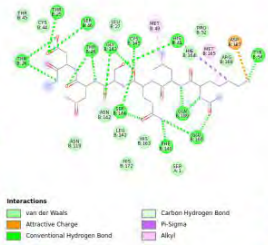
KLQSLV



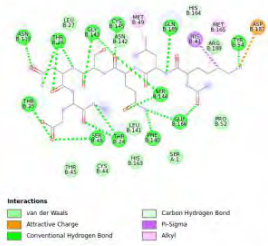
KLQSNA



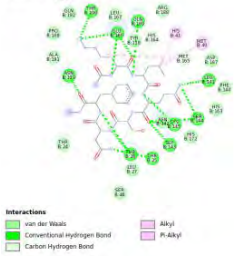
KLQSND



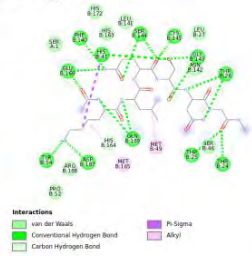
KLQSNE



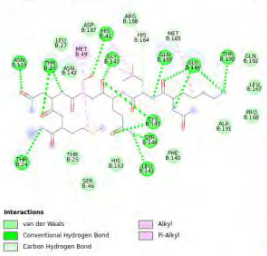
KLQSNF



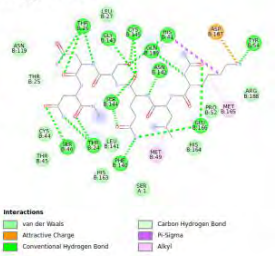
KLQSNG



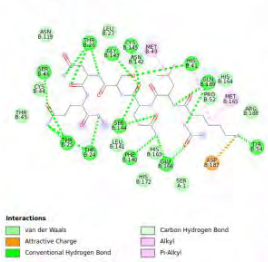
KLQSNM



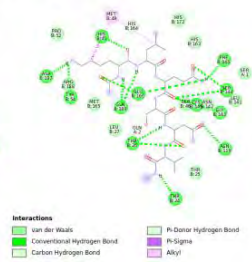
KLQSNN



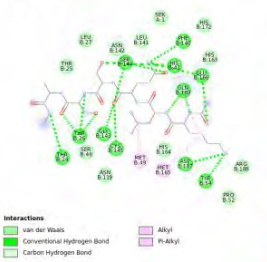
KLQSNQ



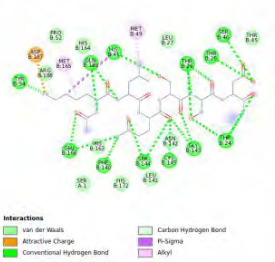
KLQSNV



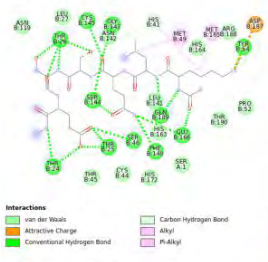
KLQSSA



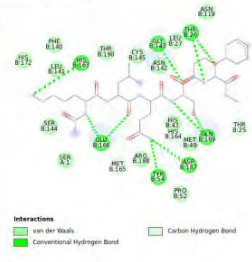
KLQSSD



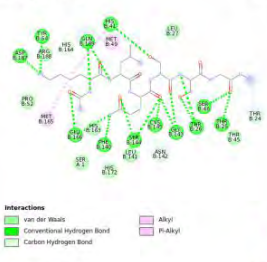
KLQSSE



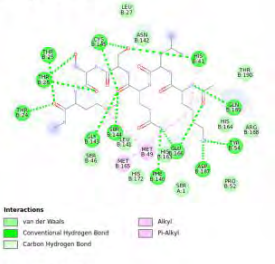
KLQSSF



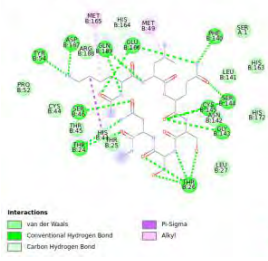
KLQSSG



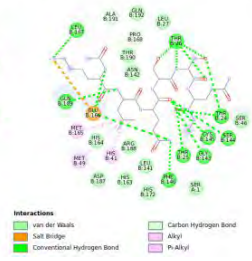
KLQSSM



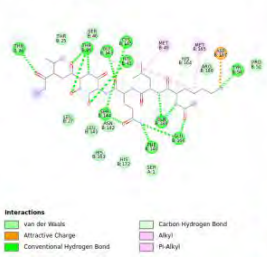
KLQSSN



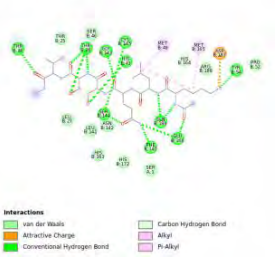
KLQSSQ

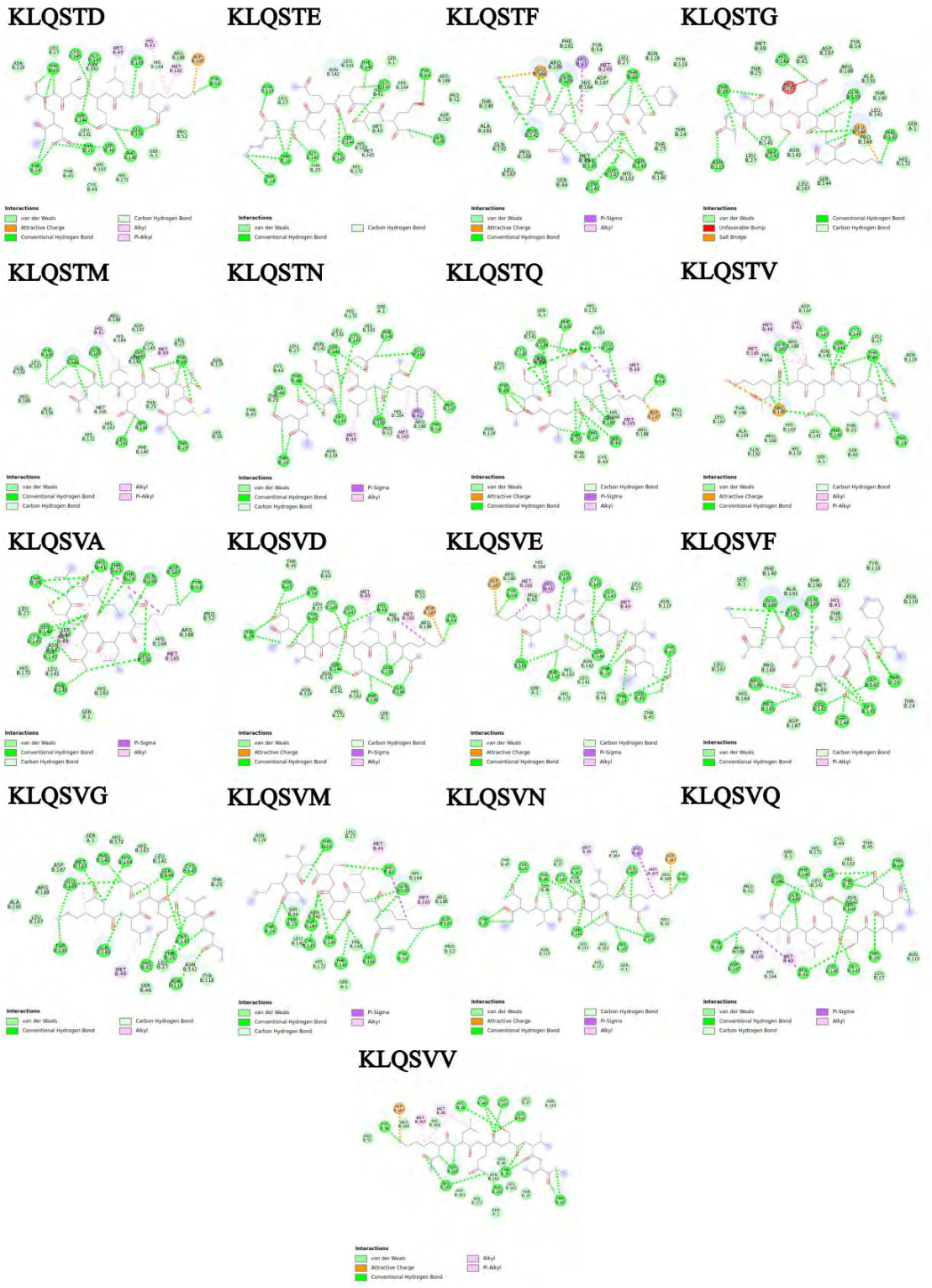


KLQSSV



KLQSTA





Supplementary figure 3.3. Resolution of intermolecular interactions between M^{pro} and substrates at active site. 2D representation of the protein-ligand interactions at active sites for M^{pro} complexed with KLQ hexapeptides. The images were generated on BIOVIA Discovery Studio 2020 Client.

APPENDICES

CHAPTER TWO

Appendix A: The generation of the capped multi-conformer hexapeptide substrates using Python, SMILES and SMARTS

```
#!/usr/bin/env python
# coding: utf-8

# In[ ]:



from rdkit import Chem
from rdkit.Chem import AllChem
from rdkit.Chem import Draw

p3list = ["T","R","K","V","M"]
p2list = ["L"]
p1list = ["Q"]
p1plist = ["S","A"]
p2plist = ["A","G","E","T","K","L","N","S","V"]
p3plist = ["E","N","A","D","F","G","M","Q","V"]

peptidestrings, peptidemolecules = [],[]
for p3 in p3list:
        for p2 in p2list:
            for p1 in p1list:
                for p1p in p1plist:
                    for p2p in p2plist:
                        for p3p in p3plist:
                            sequence = p3+p2+p1+p1p+p2p+p3p
                            peptidestrings.append(sequence)
                            peptide=Chem.rdmolfiles.MolFromSequence(sequence)
                            methylcarbonyl=Chem.MolFromSmiles('NC(=O)C')
                            methylamine=Chem.MolFromSmiles('NC')

                            ctermpattern=Chem.MolFromSmarts('[$(OC(=O)CN)]')
                            new=AllChem.ReplaceSubstructs(peptide,ctermpattern,methylamine)
                            new[0]
                            ntermpattern=Chem.MolFromSmarts('[$(NCC(=O))]')
                            final=AllChem.ReplaceSubstructs(new[0],ntermpattern,methylcarbonyl)
```

```
                        final[0]
                        peptide = AllChem.RemoveHs(final[0])
                        peptide_h= AllChem.AddHs(peptide)
                        AllChem.EmbedMolecule(peptide_h)
                        AllChem.UFFOptimizeMolecule(peptide_h)
                        writer_pdb= AllChem.PDBWriter(sequence+".pdb")
                        writer_pdb.write(peptide_h)
                        peptide_confs= peptide_h
                        ids = AllChem.EmbedMultipleConfs(peptide_confs,numConfs=100)
                        writer= AllChem.SDWriter(sequence+".sdf")
                        for ido in ids:
                            writer.write(peptide_confs,confId=ido)
                        peptidemolecules.append(peptide_h)



# In[ ]:



Draw.MolsToGridImage(peptidemolecules)

# In[12]:

import nglview as nv
first_structure = nv.RdkitStructure(peptidemolecules[0])
first_view= nv.NGLWidget()
first_view.add_component(first_structure)
first_view
#view.add_sticks(first_view)

# In[ ]:
```

CHAPTER THREE

Appendix B: Geometry optimization of the conformers of hexapeptide substrates using OpenBabel and xtb software - Part 1

```
#!/usr/bin/python

import os

files=os.listdir("./")
for file in files:
    if "sdf" in file:
        directory=file[:-4]
```

```
        os.system("mkdir "+directory)
        os.system("babel -isdf "+file+" -oxyz "+directory+"/"+directory+".xyz -h -m")
        print("module add chpc/openbabel/2.3.1/cmake-3.7.1/gcc-4.9.0; cd /mnt/lustre/users/szabo/SARS_CoV_2/Ligands/"+directory+";/mnt/lustre/users/szabo/SARS_CoV_2/Ligands/sdf_xyz_pdb.py")
```

Appendix C: Geometry optimization of the conformers of hexapeptide substrates using OpenBabel and xtb software - Part 2 (sdf_xyz_pdb.py)

```
#!/usr/bin/python

import os

file_2 = os.listdir(".")


for file_name2 in file_2:
    if file_name2.endswith(".xyz"):
        pdbname = file_name2[:-4]+".pdb"
        if not os.path.isfile(pdbname):

            print("rm -f charges wbo xtbopt.log xtbopt.xyz xtbrestart")
            os.system("rm -f charges wbo xtbopt.log xtbopt.xyz xtbrestart")


            file_name_2 = file_name2[:-4]

            print("xtb {0}.xyz".format(file_name_2))
            os.system("/home/szabo/bin/xtb {0}.xyz -opt".format(file_name_2))

            print("babel -ixyz xtbopt.xyz -opdb {0}.pdb".format(file_name_2))
            os.system("babel -ixyz xtbopt.xyz -opdb {0}.pdb".format(file_name_2))

print("Done")
```

Appendix D: Automated preparation of the ligands (hexapeptide conformers) for molecular docking

```
#!/usr/bin/python
```

```
import os

files = os.listdir(".")

for filename in files:
    if filename.endswith(".pdb"):
        print("prepare_ligand4.py -l {0}".format(filename))
        os.system("prepare_ligand4.py -l {0}".format(filename))

print("Done!")
```

Appendix E: Automated creation of the Vina configuration files

```
#!/usr/bin/python

import os

directories = os.listdir(".")

rec = "conf_A.pdbqt"
protein = rec
prefix = rec[:-6]+"_"

x_value_A, x_value_B = "12.059","-18.444"
y_value_A, y_value_B = "8.933","-16.361"
z_value_A, z_value_B = "29.021","7.944"
exhaust = "480"
cpu = "24"

for directory in directories:
    if(directory == "receptors" or directory.endswith("py") or directory.endswith("pbs") or
directory.endswith("txt")):
        pass
    else:
        os.system("cp ./receptors/"+rec+" ./"+directory)
        files = os.listdir("./"+directory)
        for file1 in files:
            if file1.endswith(".pdbqt")and not "conf" in file1:
                filename= file1[:-6]
                print(filename)
                vina_A, vina_B = "./"+directory+"/"+prefix+filename+"_A.vina",
"./"+directory+"/"+prefix+filename+"_B.vina"

                vina = open(vina_A,"w")
```

```
                vina.write("receptor = "+protein)
                vina.write("\nligand = "+filename+".pdbqt")
                vina.write("\nlog = "+prefix+filename+"_A.log")
                vina.write("\nout = "+prefix+filename+"_A.all.pdbqt")
                vina.write("\ncenter_x = "+x_value_A+"\ncenter_y = "+y_value_A+"\ncenter_z =
"+z_value_A)
                vina.write("\nsize_x = 22\nsize_y = 22\nsize_z = 22")
                vina.write("\nenergy_range = 4\nexhaustiveness = "+exhaust+"\ncpu = "+cpu+"\n")
                vina.close()

                vina = open(vina_B,"w")
                vina.write("receptor = "+protein)
                vina.write("\nligand = "+filename+".pdbqt")
                vina.write("\nlog = "+prefix+filename+"_B.log")
                vina.write("\nout = "+prefix+filename+"_B.all.pdbqt")
                vina.write("\ncenter_x = "+x_value_B+"\ncenter_y = "+y_value_B+"\ncenter_z =
"+z_value_B)
                vina.write("\nsize_x = 22\nsize_y = 22\nsize_z = 22")
                vina.write("\nenergy_range = 4\nexhaustiveness = "+exhaust+"\ncpu = "+cpu+"\n")
                vina.close()

print("\n\nDONE!")
```

Appendix F: Typical example of the specifications in a Vina configuration file

```
receptor = conf_A.pdbqt
ligand = KLQAAE1.pdbqt
log = conf_A_KLQAAE1_A.log
out = conf_A_KLQAAE1_A.all.pdbqt
center_x = 12.059
center_y = 8.933
center_z = 29.021
size_x = 22
size_y = 22
size_z = 22
energy_range = 4
exhaustiveness = 480
cpu = 24

receptor = conf_A.pdbqt
ligand = KLQAAE1.pdbqt
log = conf_A_KLQAAE1_B.log
out = conf_A_KLQAAE1_B.all.pdbqt
center_x = -18.444
```

```
center_y = -16.361
center_z = 7.944
size_x = 22
size_y = 22
size_z = 22
energy_range = 4
exhaustiveness = 480
cpu = 24
```

Appendix G: Automated generation for all the Vina commands that execute Molecular Docking using AutoDock Vina

```
#!/usr/bin/python

import os

#makes commands.txt

vina_dir = os.popen('find . -name "conf_A*_B.vina" -type f').read()
vina_dir = vina_dir.split("\n")
vina_dir.pop(-1)

command = open("commands.txt","w")
for directory in vina_dir:
    folder = directory[2:8]
    vina_file = directory[9:]
    vina_num = directory[22:-7]
    pdbqt_name= directory[:-5]+".all.pdbqt"

    if not "folder" in directory:
        if not os.path.exists(pdbqt_name):
            if int(vina_num) <41 and int(vina_num) > 30:
                command.write("module add chpc/autodock_vina/1.1.2/gcc-6.1.0;cd
/mnt/lustre/users/szabo/SARS_CoV_2/pdb_ligands/Ligands/"+folder+"/;vina --config "+vina_file+"\n")
                print(pdbqt_name)
        else:
            if int(vina_num) < 41 and int(vina_num) > 30:
                print("existing ",pdbqt_name)
command.close()
```

Appendix H: Examples of the Vina commands

```
module add chpc/autodock_vina/1.1.2/gcc-6.1.0;cd
```

```
/mnt/lustre/users/szabo/SARS_CoV_2/pdb_ligands/Ligands/KLQASE/;vina --config
conf_A_KLQASE31_B.vina
module add chpc/autodock_vina/1.1.2/gcc-6.1.0;cd
/mnt/lustre/users/szabo/SARS_CoV_2/pdb_ligands/Ligands/KLQASE/;vina --config
conf_A_KLQASE32_B.vina
module add chpc/autodock_vina/1.1.2/gcc-6.1.0;cd
/mnt/lustre/users/szabo/SARS_CoV_2/pdb_ligands/Ligands/KLQASE/;vina --config
conf_A_KLQASE40_B.vina
module add chpc/autodock_vina/1.1.2/gcc-6.1.0;cd
/mnt/lustre/users/szabo/SARS_CoV_2/pdb_ligands/Ligands/KLQASE/;vina --config
conf_A_KLQASE38_B.vina
module add chpc/autodock_vina/1.1.2/gcc-6.1.0;cd
/mnt/lustre/users/szabo/SARS_CoV_2/pdb_ligands/Ligands/KLQASE/;vina --config
conf_A_KLQASE33_B.vina
module add chpc/autodock_vina/1.1.2/gcc-6.1.0;cd
/mnt/lustre/users/szabo/SARS_CoV_2/pdb_ligands/Ligands/KLQASE/;vina --config
conf_A_KLQASE37_B.vina
module add chpc/autodock_vina/1.1.2/gcc-6.1.0;cd
/mnt/lustre/users/szabo/SARS_CoV_2/pdb_ligands/Ligands/KLQASE/;vina --config
conf_A_KLQASE35_B.vina
```

Appendix I: PBS job file specifying the execution of Molecular Docking using the implemented multi-CPU parallelization algorithms on AutoDock Vina

```
#!/bin/bash
#PBS -e /mnt/lustre/users/szabo/SARS_CoV_2/pdb_ligands/Ligands/gnu_parallel.stderr.out
#PBS -o /mnt/lustre/users/szabo/SARS_CoV_2/pdb_ligands/Ligands/gnu_parallel.stdout.out
#PBS -V
#PBS -P CHEM0802
#PBS -M youremailaddress
#PBS -l select=20:ncpus=24
#PBS -W group_list=largeq
#PBS -l walltime=96:00:00
#PBS -q large
#PBS -m be
#PBS -r n
#PBS -mb

module add chpc/gnu/parallel-20160422

WORKING_DIR=/mnt/lustre/users/szabo/SARS_CoV_2/pdb_ligands/Ligands/

cd ${WORKING_DIR}
echo "start `date +%s`"
```

```
parallel -M --sshdelay 0.2 -j 1 -u --sshloginfile ${PBS_NODEFILE} < commands.txt
echo "end `date +%s`"
```

Appendix J: Automated separation of the best binding poses for each hexapeptide

```
import os

folders= os.popen("ls -d */").read()
folders= folders.split("\n")
folders.pop(-1)
best_dock,best_aff,affs=[],0,[]

for folder in folders:
    print(folder)
    affs, best_dock = [], []
    os.chdir("./"+folder)
    affinity = os.popen("head *pdbqt| grep 'VINA'").read()
    affinity= affinity.split("\n")
    affinity.pop(-1)

    for aff in affinity:
        num = aff[25:30].strip()
        affs.append(float(num))

    affs.sort()
    best_aff = [affs[0],affs[1],affs[2]]

    files= os.popen("ls *.pdbqt").read()
    files= files.split("\n")
    files.pop(-1)
    for aff in best_aff:
        for filename in files:
            content = open(filename,"r")
            text = content.readlines()
            content.close()
            if(str(aff) in text[0]):
                best_dock.append(filename)
            else:
                pass

    os.mkdir("./best_dock/")
    for best in best_dock:
        print(best)
        os.system("cp {0} ./best_dock/".format(best))
```

```
    os.chdir("..")
```

Appendix K: Automated summarisation of the docking results, detailing the number conformers redocked, the best binding conformer and its respective binding energy for each hexapeptide

```
import os

dirs = os.popen("ls -d */").read()
dirs = dirs.split("\n")
dirs.pop(-1)

docks = open("dock_summary.csv","w")
docks.write("Substrate;Docked Conformers;Best Conformer;Best Energy (kcal/mol)\n")

for dir1 in dirs:
    os.chdir("./"+dir1)
    best_files = os.listdir("./best_dock")
    best_files.sort()
    best = best_files[0]
    energy = os.popen("head -n1 ./best_dock/{0}".format(best)).read()
    energy = energy[25:29]
    conf = best[13:-21]
    others = os.popen("ls ./other_files/*.all.pdbqt | wc -l").read()
    others = others.strip("\n")
    lig = dir1[:6]
    docks.write("{0};{1};{2};{3}\n".format(lig,others,conf,energy))
    os.chdir("..")
```

CHAPTER FOUR

Appendix L: Automated restoration of the amino acid information in the best binding pose and superimposition of the best binding poses

```
import os

folders= os.popen("ls -d */").read()
folders= folders.split("\n")
folders.pop(-1)
```

```
for folder in folders:
    os.chdir(folder)
    best_docks = os.popen("ls ./best_dock").read()
    best_docks = best_docks.split("\n")
    best_docks.pop(-1)

    for dock in best_docks:
        print(folder)
        print(dock)
        num = dock[13:-21]
        f_dir= "./best_dock/"+dock
        print(f_dir)
        print(num)

        os.system("../sort.pl {0} > {1}.pdb".format(f_dir,num))

    counter,load_text,sticks_show= 0,"","" 
    while(counter < len(best_docks)):
        load_text+="load ./best_dock/{0}, LIG{1}\n".format(best_docks[counter],counter+1)
        sticks_show+="show sticks, LIG{0}\n".format(counter+1)
        counter+=1



    text1='''
##################
### Set Style ###
##################
hide everything
set bg_rgb, white
zoom\n'''
    text2='''
####################
### Save a copy ###
####################
set antialias, 2
set hash_max, 220
set ray_shadows,0
png {0}.png, width=25cm, height=25cm, dpi=300
quit'''.format(folder[:-1])
    final_text=load_text+text1+sticks_show+text2
    pymol_vis=open("pymol_vis.pml","w")
    pymol_vis.write(final_text)
    pymol_vis.close()
    os.system("pymol pymol_vis.pml")
```

```
    os.chdir("..")
```

Appendix M: The custom Perl script to restore amino acid information of the best binding poses.

```
#!/usr/bin/perl
#
open(PDBIN,"< $ARGV[0]");

my @atommatrix;
my $an=0;

sub distance
{
  my $a1=shift;
  my $a2=shift;

  my $x1=$atommatrix[$a1][1];
  my $x2=$atommatrix[$a2][1];
  my $y1=$atommatrix[$a1][2];
  my $y2=$atommatrix[$a2][2];
  my $z1=$atommatrix[$a1][3];
  my $z2=$atommatrix[$a2][3];
  my $dist=sqrt( (($x1-$x2)*($x1-$x2))+(($y1-$y2)*($y1-$y2))+(($z1-$z2)*($z1-$z2)) );
  #print "distance $dist\n";
  return $dist;
}


sub allorder
{
  my $at=shift;
  my $or=0;
  for(my $i=0;$i<$an;$i++)
   {
     if(($atommatrix[$i][0] =~ m/C/)and(distance($at,$i)<1.7))
     {
       $or++;
     }
     if(($atommatrix[$i][0] =~ m/N/)and(distance($at,$i)<1.7))
     {
       $or++;
```

```
      }
    }
  return $or;
}



sub corder
{
  my $at=shift;
  my $or=0;
  for(my $i=0;$i<$an;$i++)
    {
      if(($atommatrix[$i][0] =~ m/C/)and(distance($at,$i)<1.7))
      {
        $or++;
      }
    }
  return $or;
}

sub iscarbonyl
{
    my $at=shift;
    my $co=-1;
    for(my $i=0;$i<$an;$i++)
    {
      if(($atommatrix[$i][0] =~ m/O/)and(distance($at,$i)<1.3))
      {
        $co=$i;
      }
    }
return $co;
}

sub findcs
{
  my $at=shift;
  my @cs;
  for(my $i=0;$i<$an;$i++)
  {
      if(($atommatrix[$i][0] =~ m/C/)and(distance($at,$i)<1.7))
      {
        push @cs, $i;
```

```
        #print "$i\n";
      }
   }
   return @cs;
}

sub findterminalNMe
{
   my $n=-1,$me=-1;
   for(my $i=0;$i<$an;$i++)
   {
      if($atommatrix[$i][0] =~ m/N/)
      {
            #print "$i is N\n";
            my @cs=findcs($i);
            foreach my $c (@cs)
            {
              my $o=iscarbonyl($c);
              my $or=corder($c);
              #print "$c $o order $or\n";
              if($or eq 1)
              {
                $atommatrix[$i][4]=8;
                $atommatrix[$c][4]=8;
                #print "found NME $i $c\n";
                $n=$i;$me=$c;
              }
            }
      }
   }
 return($n,$me)
}

sub findterminalcco
{
   my $me=-1,$co=-1,$oc=-1;
   for(my $i=0;$i<$an;$i++)
   {
      if($atommatrix[$i][0] =~ m/C/)
      {
        my $or=allorder($i);
        my $ca=iscarbonyl($i);
        if(($or eq 2) and ($ca eq -1))
        {
```

```
        ##print "candidate $i\n";
         my @cs=findcs($i);
            foreach my $c (@cs)
            {
               my $o=iscarbonyl($c);
               if($o ge 0)
               {
                   $atommatrix[$i][4]=1;
                   $atommatrix[$c][4]=1;
                   $atommatrix[$o][4]=1;
                   #print "found CCO $i $c $o\n";
                   $me=$i;$co=$c;$oc=$o;
               }
            }
      }
     }
  }
 return($me,$co,$oc);
}

sub findn
{
  my $at=shift;
  my $n=-1;
  for(my $i=0;$i<$an;$i++)
  {
     if($atommatrix[$i][0] =~ m/N/)
     {
       my $d=distance($at,$i);
       #print("N $i distance to $at = $d\n");
       if($d<1.4)
       {
         $n=$i
       }
     }
  }
return $n;
}


sub findco
{
  my $at=shift;
  my $co=-1;
```

```
  my $oc=-1;
  for(my $i=0;$i<$an;$i++)
  {
     if($atommatrix[$i][0] =~ m/C/)
     {
       if(distance($at,$i)<1.55)
       {
         my $o=iscarbonyl($i);
         if($o ge 0){$co=$i;$oc=$o}
       }
     }
  }
return ($co,$oc);
}

sub findca
{
  my $at=shift;
  my $ca=-1;
  for(my $i=0;$i<$an;$i++)
  {
     if($atommatrix[$i][0] =~ m/C/)
     {
       my $d=distance($at,$i);
       #print "C $i distance to N $at is $d\n";
       if($d<1.5)
       {
         my $o=iscarbonyl($i);
         #print "o is $o\n";
         if($o eq -1){$ca=$i;}
       }
     }
  }
return ($ca);
}

sub findnextbackbone
{
  my $co1=shift;
  my $n=findn($co1);
  my $ca=findca($n);
  my ($co,$oc)=findco($ca);

  return($n,$ca,$co,$oc);
```

```
}

sub fillsidechain
{
  my $at=shift;
  #print "working with atom $at\n";
  my $amino=$atommatrix[$at][4];
  my $level=$atommatrix[$at][5]+1;
  for(my $i=0;$i<$an;$i++)
  {
     my $d=distance($at,$i);
     my $tolerance=1.7;
     if(($atommatrix[$i][0] =~ m/S/)or($atommatrix[$at][0] =~ m/S/)){$tolerance=1.9}
     if(($d<$tolerance)and($atommatrix[$i][4] eq 0) and ($i ne $at))
     {
        $atommatrix[$i][4]=$amino;
        $atommatrix[$i][5]=$level;
        if($atommatrix[$i][0]  =~  m/H/){$atommatrix[$i][5]=$at};#identify  parent  of  Hydrogens
rather than level.
        fillsidechain($i);
     }

  }
}



sub identifyamino
{
  my $at=shift;
  my @O=[0,0,0,0,0,0],@C=[0,0,0,0,0,0],@N=[0,0,0,0,0,0],@S=[0,0,0,0,0,0];
  my $totalc=0;my $totalo=0; my $totaln=0; my $totals=0;



  for(my $i=0;$i<$an;$i++)
  {
     my $level=$atommatrix[$i][5];

     if(($level>1)and($atommatrix[$i][4] eq $at))
     {
     if($atommatrix[$i][0] =~ m/C/){$C[$level]++;$totalc++}
     if($atommatrix[$i][0] =~ m/N/){$N[$level]++;$totaln++}
     if($atommatrix[$i][0] =~ m/O/){$O[$level]++;$totalo++}
     if($atommatrix[$i][0] =~ m/S/){$S[$level]++;$totals++}
```

```
    }
  }

#print "o $totalo n $totaln s $totals c $totalc\n";

my $res="XXX";
if(($totalc eq 0) and ($totalo eq 0) and ($totaln eq 0) and ($totals eq 0)){$res="GLY"}
if(($totalc eq 1) and ($totalo eq 0) and ($totaln eq 0) and ($totals eq 0)){$res="ALA"}
if(($totalc eq 1) and ($totalo eq 0) and ($totaln eq 0) and ($totals eq 1)){$res="CYS"}
if(($totalc eq 1) and ($totalo eq 1) and ($totaln eq 0) and ($totals eq 0)){$res="SER"}
if(($totalc eq 2) and ($totalo eq 1) and ($totaln eq 0) and ($totals eq 0)){$res="THR"}
if(($totalc eq 2) and ($totalo eq 2) and ($totaln eq 0) and ($totals eq 0)){$res="ASP"}
if(($totalc eq 2) and ($totalo eq 1) and ($totaln eq 1) and ($totals eq 0)){$res="ASN"}
if(($totalc eq 3) and ($totalo eq 0) and ($totaln eq 0) and ($totals eq 0)){$res="VAL"}
if(($totalc eq 3) and ($totalo eq 0) and ($totaln eq 0) and ($totals eq 1)){$res="MET"}
if(($totalc eq 3) and ($totalo eq 2) and ($totaln eq 0) and ($totals eq 0)){$res="GLU"}
if(($totalc eq 3) and ($totalo eq 1) and ($totaln eq 1) and ($totals eq 0)){$res="GLN"}
if(($totalc eq 3) and ($totalo eq 0) and ($totaln eq 1) and ($totals eq 0)){$res="PRO"}
if(($totalc eq 4) and ($totalo eq 0) and ($totaln eq 1) and ($totals eq 0)){$res="LYS"}
if(($totalc eq 7) and ($totalo eq 0) and ($totaln eq 0) and ($totals eq 0)){$res="PHE"}
if(($totalc eq 7) and ($totalo eq 1) and ($totaln eq 0) and ($totals eq 0)){$res="TYR"}
if(($totalc eq 9) and ($totalo eq 0) and ($totaln eq 1) and ($totals eq 0)){$res="TRP"}
if(($totalc eq 4) and ($totalo eq 0) and ($totaln eq 3) and ($totals eq 0)){$res="ARG"}

if(($totalc eq 4) and ($totalo eq 0) and ($totaln eq 0) and ($totals eq 0) and ($C[3] eq
1)){$res="LEU"}
if(($totalc eq 4) and ($totalo eq 0) and ($totaln eq 0) and ($totals eq 0) and ($C[3] eq
2)){$res="ILE"}

return $res;
}

sub printamino
{
  my $amino=shift;
  my $code=shift;
  my $atomnumber=shift;

#N
  for(my $i=0;$i<$an;$i++)
  {
    if(($atommatrix[$i][4] eq $amino) and ($atommatrix[$i][5] eq 1) and ($atommatrix[$i][0]=~
m/N/))
    {
```

```
        printf "ATOM      %4d  N   $code  %4d    %8.3f%8.3f%8.3f  1.00  1.00
$atommatrix[$i][0]\n",$atomnumber,$amino,$atommatrix[$i][1],$atommatrix[$i][2],$atommatrix[$i]
[3];
        $atomnumber++
    }
  }

#CA
 for(my $i=0;$i<$an;$i++)
  {
    my $ic=-1;
    if(($atommatrix[$i][0]=~ m/C/) and ($atommatrix[$i][5] eq 1) and ($atommatrix[$i][4] eq
$amino)){$ic=iscarbonyl($i)}
    if(($ic eq -1) and ($atommatrix[$i][0]=~ m/C/) and ($atommatrix[$i][5] eq 1) and
($atommatrix[$i][4] eq $amino))
    {
        printf "ATOM      %4d  CA  $code  %4d    %8.3f%8.3f%8.3f  1.00  1.00
$atommatrix[$i][0]\n",$atomnumber,$amino,$atommatrix[$i][1],$atommatrix[$i][2],$atommatrix[$i]
[3];
        $atomnumber++
    }
  }

#C
#
 for(my $i=0;$i<$an;$i++)
  {
    my $ic=-1;
    if(($atommatrix[$i][0]=~ m/C/) and ($atommatrix[$i][5] eq 1)){$ic=iscarbonyl($i)}
    if(($ic gt -1) and ($atommatrix[$i][0]=~ m/C/) and ($atommatrix[$i][5] eq 1) and
($atommatrix[$i][4] eq $amino))
    {
        printf "ATOM      %4d  C   $code  %4d    %8.3f%8.3f%8.3f  1.00  1.00
$atommatrix[$i][0]\n",$atomnumber,$amino,$atommatrix[$i][1],$atommatrix[$i][2],$atommatrix[$i]
[3];
        $atomnumber++;
       printf "ATOM      %4d  O   $code  %4d    %8.3f%8.3f%8.3f  1.00  1.00
$atommatrix[$ic][0]\n",$atomnumber,$amino,$atommatrix[$ic][1],$atommatrix[$ic][2],$atommatrix
[$ic][3];
        $atomnumber++;



    }
```

```
  }

for(my $lev=2;$lev<8;$lev++)
{
  my @lcs;$lcs[2]="B";$lcs[3]="G";$lcs[4]="D";$lcs[5]="E";$lcs[6]="Z";$lcs[7]="H";
#BETA
 for(my $i=0;$i<$an;$i++)
  {
    if((not ($atommatrix[$i][0]=~ m/H/)) and ($atommatrix[$i][5] eq $lev) and ($atommatrix[$i][4]
eq $amino))
    {
       printf "ATOM     %4d    $atommatrix[$i][0]$lcs[$lev]    $code    %4d      %8.3f%8.3f%8.3f    1.00
1.00
$atommatrix[$i][0]\n",$atomnumber,$amino,$atommatrix[$i][1],$atommatrix[$i][2],$atommatrix[$i]
[3];
       $atomnumber++;
       for(my $j=0;$j<$an;$j++)
       {
         if(($atommatrix[$i][0]=~ m/H/) and ($atommatrix[$i][5] eq $i) and ($atommatrix[$i][4] eq
$amino))
          {
#                        printf "ATOM     %4d    H$atommatrix[$i][0]$lcs[$lev]
$code    %4d      %8.3f%8.3f%8.3f    1.00    1.00
$atommatrix[$i][0]\n",$atomnumber,$amino,$atommatrix[$i][1],$atommatrix[$i][2],$atommatrix[$i][3
];
                       $atomnumber++;
          }

       }
    }
  }

}

return $atomnumber;
}

while (my $line = <PDBIN>)
{
   if(($line =~ m/HETATM/)or($line =~ m/ATOM/))
   {
     my $atomname=substr $line,13,4;$atomname =~ s/\s+//g;
     my $x=substr $line,31,8;$x =~ s/\s+//g;
     my $y=substr $line,39,8;$y =~ s/\s+//g;
```

```
      my $z=substr $line,47,8;$z =~ s/\s+//g;
      #print $line;
      #print "$atomname,$x,$y,$z\n";
      $atommatrix[$an][0]=substr $atomname, 0 ,1;
      #print "*$atomname*\n";
      $atommatrix[$an][1]=$x;
      $atommatrix[$an][2]=$y;
      $atommatrix[$an][3]=$z;
      $atommatrix[$an][4]=0;
      $an++;
   }
}
close PDBIN;

my($cme,$cco,$coc)=findterminalcco();
print("backbone 1: $cme,$cco,$coc\n");
$atommatrix[$cme][4]=1;$atommatrix[$cco][4]=1;$atommatrix[$coc][4]=1;
$atommatrix[$cme][5]=1;$atommatrix[$cco][5]=1;$atommatrix[$coc][5]=1;
my($n1,$ca1,$co1,$oc1)=findnextbackbone($cco);
print("backbone 2: $n1,$ca1,$co1,$oc1\n");
$atommatrix[$n1][4]=2;$atommatrix[$ca1][4]=2;$atommatrix[$co1][4]=2;$atommatrix[$oc1][4]=
2;
$atommatrix[$n1][5]=1;$atommatrix[$ca1][5]=1;$atommatrix[$co1][5]=1;$atommatrix[$oc1][5]=
1;
my($n2,$ca2,$co2,$oc2)=findnextbackbone($co1);
print("backbone 3: $n2,$ca2,$co2,$oc2\n");
$atommatrix[$n2][4]=3;$atommatrix[$ca2][4]=3;$atommatrix[$co2][4]=3;$atommatrix[$oc2][4]=
3;
$atommatrix[$n2][5]=1;$atommatrix[$ca2][5]=1;$atommatrix[$co2][5]=1;$atommatrix[$oc2][5]=
1;
my($n3,$ca3,$co3,$oc3)=findnextbackbone($co2);
print("backbone 4: $n3,$ca3,$co3,$oc3\n");
$atommatrix[$n3][4]=4;$atommatrix[$ca3][4]=4;$atommatrix[$co3][4]=4;$atommatrix[$oc3][4]=
4;
$atommatrix[$n3][5]=1;$atommatrix[$ca3][5]=1;$atommatrix[$co3][5]=1;$atommatrix[$oc3][5]=
1;
my($n4,$ca4,$co4,$oc4)=findnextbackbone($co3);
print("backbone 5: $n4,$ca4,$co4,$oc4\n");
$atommatrix[$n4][4]=5;$atommatrix[$ca4][4]=5;$atommatrix[$co4][4]=5;$atommatrix[$oc4][4]=
5;
$atommatrix[$n4][5]=1;$atommatrix[$ca4][5]=1;$atommatrix[$co4][5]=1;$atommatrix[$oc4][5]=
1;
my($n5,$ca5,$co5,$oc5)=findnextbackbone($co4);
print("backbone 6: $n5,$ca5,$co5,$oc5\n");
```

```
$atommatrix[$n5][4]=6;$atommatrix[$ca5][4]=6;$atommatrix[$co5][4]=6;$atommatrix[$oc5][4]=
6;
$atommatrix[$n5][5]=1;$atommatrix[$ca5][5]=1;$atommatrix[$co5][5]=1;$atommatrix[$oc5][5]=
1;
my($n6,$ca6,$co6,$oc6)=findnextbackbone($co5);
print("backbone 7: $n6,$ca6,$co6,$oc6\n");
$atommatrix[$n6][4]=7;$atommatrix[$ca6][4]=7;$atommatrix[$co6][4]=7;$atommatrix[$oc6][4]=
7;
$atommatrix[$n6][5]=1;$atommatrix[$ca6][5]=1;$atommatrix[$co6][5]=1;$atommatrix[$oc6][5]=
1;
my($nme,$men)=findterminalNMe();
$atommatrix[$nme][4]=8;$atommatrix[$men][4]=8;
$atommatrix[$nme][5]=1;$atommatrix[$men][5]=1;
print("backbone 8: $nme,$men\n");

#print "chain $cme, $cco, $n1, $ca1, $co1, $oc1, $n2, $ca2, $co2, $oc2\n";
print "point A\n";
fillsidechain($ca1);
print "point B\n";
fillsidechain($ca2);
print "point C\n";
fillsidechain($ca3);
print "point D\n";
fillsidechain($ca4);
print "point E\n";
fillsidechain($ca5);
print "point F\n";
fillsidechain($ca6);
print "point G\n";

#print "In this we have $an atoms\n";
for(my $i=0;$i<$an;$i++)
{
  #print "filling from atom $i\n";
  fillsidechain($i);
}

my $amin1="ACE";
my $amin2=identifyamino(2);
my $amin3=identifyamino(3);
my $amin4=identifyamino(4);
my $amin5=identifyamino(5);
my $amin6=identifyamino(6);
my $amin7=identifyamino(7);
```

```
my $amin8="NME";

print "REMARK    $amin1 $amin2 $amin3 $amin4 $amin5 $amin6 $amin7 $amin8\n";

my $nextno=printamino(1,$amin1,1);
$nextno=printamino(2,$amin2,$nextno);
$nextno=printamino(3,$amin3,$nextno);
$nextno=printamino(4,$amin4,$nextno);
$nextno=printamino(5,$amin5,$nextno);
$nextno=printamino(6,$amin6,$nextno);
$nextno=printamino(7,$amin7,$nextno);
$nextno=printamino(8,$amin8,$nextno);

for(my $j=1;$j<9;$j++)
{
  for(my $i=0;$i<$an;$i++)
  {
    if($atommatrix[$i][4] eq $j)
    {
      $k=$i+1;
      print "$k $atommatrix[$i][0] $atommatrix[$i][4] (level $atommatrix[$i][5])\n";
    }
  }
}
```

Appendix N: Supplementary edition of the atom types in restored amino acid information of the best binding poses.

```
import os

pdb_files = os.popen('''find . -name "*.pdb" -type f''').read()
pdb_files= pdb_files.split("\n")
pdb_files.pop(-1)
pdb_files.sort()

for pdb_file in pdb_files:
    pdb_content = open(pdb_file,"r")
    pdb_lines = pdb_content.readlines()
    pdb_content.close()

    if("XXX" in pdb_lines[0]):
        os.system("rm {0}".format(pdb_file))
        pass
    elif("complex" in pdb_file or "6xhm" in pdb_file):
```

```
        pass
    else:
        print(pdb_file)
        print(pdb_lines[0])
        counter =1
        while(counter < len(pdb_lines)):
            line = pdb_lines[counter]
            prot = line[17:20]
            atom = line[13:16]

            if(prot == "ACE" and   atom == "CA "):
                edit_line = line[:13]+"CH3"+line[16:]
                pdb_lines[counter]= edit_line
            elif(prot == "NME" and   atom == "CA "):
                edit_line = line[:13]+"CH3"+line[16:]
                pdb_lines[counter]= edit_line
            elif(prot == "GLN"):
                if(atom == "OE "):
                    edit_line = line[:13]+"OE1"+line[16:]
                    pdb_lines[counter]= edit_line
                elif(atom == "NE "):
                    edit_line = line[:13]+"NE2"+line[16:]
                    pdb_lines[counter]= edit_line
            elif(prot == "LEU"):
                if(atom == "CD "):
                    edit_line = line[:13]+"CD1"+line[16:]
                    line2 =   pdb_lines[counter+1]
                    edit_line2= line2[:13]+"CD2"+line2[16:]
                    pdb_lines[counter]= edit_line
                    pdb_lines[counter+1]= edit_line2
                    counter+=1
            elif(prot == "THR"):
                if(atom == "OG "):
                    edit_line = line[:13]+"OG1"+line[16:]
                    pdb_lines[counter]= edit_line
                elif(atom == "CG "):
                    edit_line =line[:13]+"CG2"+line[16:]
                    pdb_lines[counter]= edit_line
            elif(prot == "GLU"):
                if(atom == "OE "):
                    edit_line = line[:13]+"OE1"+line[16:]
                    line2 =   pdb_lines[counter+1]
                    edit_line2= line2[:13]+"OE2"+line2[16:]
                    pdb_lines[counter]= edit_line
```

```
                pdb_lines[counter+1]= edit_line2
                counter+=1
            elif(prot == "ASP"):
                if(atom == "OD "):
                    edit_line = line[:13]+"OD1"+line[16:]
                    line2 =   pdb_lines[counter+1]
                    edit_line2= line2[:13]+"OD2"+line2[16:]
                    pdb_lines[counter]= edit_line
                    pdb_lines[counter+1]= edit_line2
                    counter+=1
            elif(prot == "VAL"):
                if(atom == "CG "):
                    edit_line = line[:13]+"CG1"+line[16:]
                    line2 =   pdb_lines[counter+1]
                    edit_line2= line2[:13]+"CG2"+line2[16:]
                    pdb_lines[counter]= edit_line
                    pdb_lines[counter+1]= edit_line2
                    counter+=1
            elif(prot == "ARG"):
                if(atom == "NH "):
                    edit_line = line[:13]+"NH1"+line[16:]
                    line2 =   pdb_lines[counter+1]
                    edit_line2= line2[:13]+"NH2"+line2[16:]
                    pdb_lines[counter]= edit_line
                    pdb_lines[counter+1]= edit_line2
                    counter+=1
            elif(prot == "PHE"):
                if(atom == "CE "):
                    edit_line = line[:13]+"CE1"+line[16:]
                    line2 =   pdb_lines[counter+2]
                    edit_line2= line2[:13]+"CE2"+line2[16:]
                    pdb_lines[counter]= edit_line
                    pdb_lines[counter+2]= edit_line2
                    counter+=2
            elif(prot == "ASN"):
                if(atom == "OD "):
                    edit_line = line[:13]+"OD1"+line[16:]
                    pdb_lines[counter]= edit_line
                elif(atom == "ND "):
                    edit_line = line[:13]+"ND2"+line[16:]
                    pdb_lines[counter]= edit_line

            counter+=1
            pdb_final   = open(pdb_file,"w")
```

```
            for lines in pdb_lines:
                pdb_final.write(lines)
            pdb_final.close()
```

Appendix O: Automated addition of the hexapeptide PDB information onto the receptor PDB as a third subunit

```
#!/usr/bin/python
import os

prot_name="6xhm_apo.pdb"
folders= os.popen("ls -d */").read()
folders= folders.split("\n")
folders.pop(-1)

for folder in folders:
    pdb_list= os.listdir(folder)
    if(len(pdb_list) == 0):
        pass
    else:
        print(folder)
        lig_name=folder+pdb_list[0]
        print(lig_name)
        complex_name=folder+folder[:-1]+"_complex.pdb"

        ##read in protein
        proteinfile=open(prot_name,"r")
        protein=proteinfile.readlines()
        proteinfile.close()

        ##open output file
        complexfile=open(complex_name,"w")

        atomnumber=1
        resnumber=0
        for line in protein:
            if((line.startswith("ATOM")) or (line.startswith("HETATM"))):
                complexfile.write(line)
                atomnumber += 1

            if ("REMARK"in line):
                complexfile.write(line)
```

```
        ##sort the ligand
        ligfile=open(lig_name,"r")
        ligand=ligfile.readlines()
        ligfile.close()

        for line in ligand:
            if((line.startswith("ATOM")) or (line.startswith("HETATM"))):
                atomnumber += 1
                #delete characters (4) where the atom number should be columns 7-11
                numberstring="%4d" % atomnumber
                altered1line=line[:7]+numberstring+line[11:]
                #insert the atom number as formated integer with width 4
                #replace a character with chain c column 22
                #ligresnumber=int(line[22:26])
                #actresnumber=resnumber+ligresnumber
                altered2line=altered1line[:21]+"C"+altered1line[22:]
                #then write the line
                complexfile.write(altered2line)
        complexfile.close()
```

Appendix P: The assignment of letter code as means to uniquely identify the position of the PCA in the particular time interval

```
#!/opt/chemistry/anaconda3/bin/python

import os
import sys

files = os.popen("ls *xvg").read()
files = files.split("\n")
files.pop(-1)

def xpos(minx,maxx,x):
    deltax=(maxx-minx)/5.0
    position=x-minx
    numberofdeltas=position/deltax
    thepos=int(numberofdeltas+1)
    return thepos

def ypos(miny,maxy,y):
    deltay=(maxy-miny)/5.0
    position=y-miny
    numberofdeltas=position/deltay
    thepos=int(numberofdeltas+1)
```

```
    return thepos


def postoletter(minx,maxx,miny,maxy,x,y):
    thex=xpos(minx,maxx,x)
    they=ypos(miny,maxy,y)
    letter=""
    if they==1:
        if thex==1:
            letter="a"
        if thex==2:
            letter="b"
        if thex==3:
            letter="c"
        if thex==4:
            letter="d"
        if thex==5:
            letter="e"
    if they==2:
        if thex==1:
            letter="f"
        if thex==2:
            letter="g"
        if thex==3:
            letter="h"
        if thex==4:
            letter="i"
        if thex==5:
            letter="j"
    if they==3:
        if thex==1:
            letter="k"
        if thex==2:
            letter="l"
        if thex==3:
            letter="m"
        if thex==4:
            letter="n"
        if thex==5:
            letter="o"
    if they==4:
        if thex==1:
            letter="p"
        if thex==2:
```

```
            letter="q"
        if thex==3:
            letter="r"
        if thex==4:
            letter="s"
        if thex==5:
            letter="t"
    if they==5:
        if thex==1:
            letter="u"
        if thex==2:
            letter="v"
        if thex==3:
            letter="w"
        if thex==4:
            letter="x"
        if thex==5:
            letter="y"

    return letter



for i in files:
    code=""
    text = os.popen("cat "+i+" | tail -n2001| awk '{print $1,$2}'").read()
    lines=text.splitlines()
    maxx=-10
    maxy=-10
    minx=10
    miny=10
    for line in lines:
      x,y=line.split()
      if maxx<float(x):
          maxx=float(x)
      if maxy<float(y):
          maxy=float(y)
      if minx>float(x):
          minx=float(x)
      if miny>float(y):
          miny=float(y)
    maxx=maxx+0.01
    maxy=maxy+0.01
    counter=0
```

```
    totalx=0
    totaly=0
    for line in lines:
        x,y=line.split()
        if counter==200:
            counter=0
            averagex=totalx/200
            averagey=totaly/200
            letter=postoletter(minx,maxx,miny,maxy,averagex,averagey);
            code=code+letter
            totalx=0
            totaly=0
        totalx=totalx+float(x)
        totaly=totaly+float(y)
        counter=counter+1
    print(i,":",code)
```

Appendix Q: A preview of the letter codes that uniquely identify the PCA progression in the time intervals

```
APO_2dproj_ev_1_2.xvg : innlqpqrxx
KLQAAA_2dproj_ev_1_2.xvg : plhhhinonn
KLQAAD_2dproj_ev_1_2.xvg : qhcisnsssr
KLQAAE_2dproj_ev_1_2.xvg : qqlmmhhhgh
KLQAAG_2dproj_ev_1_2.xvg : qkhiinntsm
KLQAAM_2dproj_ev_1_2.xvg : ijnnrrrrlg
KLQAAN_2dproj_ev_1_2.xvg : hgqlmnoooo
KLQAAQ_2dproj_ev_1_2.xvg : srtniihlll
KLQAAV_2dproj_ev_1_2.xvg : gkllrrnnoo
```

Appendix R: The calculation of the differences in the letter codes between simulations and the construction of a pairwise comparison across all the systems

```
#!/opt/chemistry/anaconda3/bin/python

import os
import sys
import math

res = open("results.txt","r")
lines=res.readlines()
names=[]
codes=[]
```

```
number=0


def lettertox(l):
    x=0
    if (l=='a' or l=='f' or l=='k' or l=='p' or l=='u'):
        x=0
    if (l=='b' or l=='g' or l=='l' or l=='q' or l=='v'):
        x=1
    if (l=='c' or l=='h' or l=='m' or l=='r' or l=='w'):
        x=2
    if (l=='d' or l=='i' or l=='n' or l=='s' or l=='x'):
        x=3
    if (l=='e' or l=='j' or l=='o' or l=='t' or l=='y'):
        x=4
    return x
def lettertoy(l):
    y=0
    if (l=='a' or l=='b' or l=='c' or l=='d' or l=='e'):
        y=0
    if (l=='f' or l=='g' or l=='h' or l=='i' or l=='j'):
        y=1
    if (l=='k' or l=='l' or l=='m' or l=='n' or l=='o'):
        y=2
    if (l=='p' or l=='q' or l=='r' or l=='s' or l=='t'):
        y=3
    if (l=='u' or l=='v' or l=='w' or l=='x' or l=='y'):
        y=4
    return y


def difference(a,b):
    lena=len(a)
    lenb=len(b)
    if lena != lenb:
        return 100000000
    total=0

    for j in range (0,lena):
        deltax=lettertox(a[j])-lettertox(b[j])
        deltay=lettertoy(a[j])-lettertoy(b[j])
        difference=(deltax*deltax)+(deltay*deltay)
        total=total+difference
    return math.sqrt(total)
```

```
for line in lines:
    words=line.split()
    sub = words[0]
    sub_index = sub.rindex("_")
    lig_name = sub[:sub_index]
    names.append(lig_name)
    codes.append(words[2])
    number=number+1
print("name,",end="")

for i in range (0,number):
    print(names[i],end=",")
print("")

for i in range (0,number):
    print(names[i],end=",")
    for j in range (0,number):
        result=difference(codes[i],codes[j])
        print(result,end=",")
    print("")
```

Appendix S: A preview of the pairwise comparison matrix including all the systems

```
name,APO,KLQAAA,KLQAAD,KLQAAE,KLQAAG,KLQAAM,KLQAAN,KLQAAQ,KLQAAV,KLQAEA,KLQAED,KLQAEE,
KLQAEG,KLQAEM,KLQAEQ,KLQAEV,KLQAGA,KLQAGD,KLQAGE,KLQAGG,KLQAGM,KLQAGQ,KLQAGV,KLQAKA,K
LQAKD,KLQAKE,KLQAKG,KLQAKM,KLQAKN,KLQAKQ,KLQAKV,KLQALA,KLQALD,KLQALE,KLQALG,KLQALM,KLQA
LN,KLQALQ,KLQALV,KLQANA,KLQAND,KLQANE,KLQANG,KLQANM,KLQANQ,KLQANV,KLQASA,KLQASD,KLQASE,
KLQASG,KLQASM,KLQASN,KLQASQ,KLQASV,KLQATA,KLQATD,KLQATE,KLQATG,KLQATM,KLQATN,KLQATQ,KLQA
TV,KLQAVA,KLQAVD,KLQAVE,KLQAVG,KLQAVM,KLQAVN,KLQAVQ,KLQAVV,KLQSAA,KLQSAD,KLQSAE,KLQSAG,
KLQSAM,KLQSAN,KLQSAQ,KLQSAV,KLQSEA,KLQSEE,KLQSEG,KLQSEM,KLQSEN,KLQSEQ,KLQSGA,KLQSGD,KLQ
SGE,KLQSGN,KLQSGQ,KLQSGV,KLQSKA,KLQSKD,KLQSKG,KLQSKM,KLQSKN,KLQSKQ,KLQSKV,KLQSLA,KLQSLD
,KLQSLE,KLQSLG,KLQSLM,KLQSLN,KLQSLQ,KLQSLV,KLQSNA,KLQSND,KLQSNG,KLQSNM,KLQSNN,KLQSNQ,KL
QSSD,KLQSSE,KLQSSG,KLQSSN,KLQSSQ,KLQSSV,KLQSTA,KLQSTD,KLQSTE,KLQSTG,KLQSTM,KLQSTN,KLQSTQ
,KLQSVA,KLQSVD,KLQSVE,KLQSVG,KLQSVM,KLQSVN,KLQSVQ,KLQSVV
APO,0,7.54983443527075,6.48074069840786,7.74596669241483,7.54983443527075,5.74456264653803,6
.92820323027551,7.48331477354788,6.2449979983984,6.92820323027551,7.41619848709566,8.3066238
6291808,6.40312423743285,6.70820393249937,9.53939201416946,6.08276253029822,7.2801098892805
2,5.65685424949238,5.3851648071345,6.40312423743285,8,9.21954445729289,7.48331477354788,5.47
722557505166,7,4.79583152331272,5.47722557505166,6.92820323027551,6.92820323027551,6.480740
69840786,5.74456264653803,6,7.48331477354788,6.6332495807108,5.65685424949238,7.87400787401
181,5.8309518948453,6.48074069840786,6.08276253029822,6.40312423743285,8.30662386291808,5.83
09518948453,8.48528137423857,7.81024967590665,8.42614977317636,4.24264068711929,6.928203230
27551,5.29150262212918,6.40312423743285,5.65685424949238,6.557438524302,7,8.42614977317636,6
.70820393249937,6.85565460040104,6.70820393249937,5.74456264653803,6.16441400296898,5.744562
```

64653803,7.07106781186548,6.557438524302,6.85565460040104,7,5,6.557438524302,5.4772255750516
6,8,8.06225774829855,6.557438524302,8.06225774829855,6.78232998312527,6,7.68114574786861,5.74
456264653803,7,9.05538513813742,7.81024967590665,6,7.68114574786861,7.54983443527075,3.60555
127546399,6.92820323027551,5.8309518948453,6.92820323027551,7.61577310586391,5.916079783099
62,6,7,7.54983443527075,7,6.70820393249937,6.78232998312527,7.34846922834953,6.6332495807108,
6.2449979983984,7.21110255092798,8.30662386291808,7.61577310586391,7.61577310586391,8.717797
88708135,6.2449979983984,7.93725393319377,7.48331477354788,5.65685424949238,6.2449979983984
,7.41619848709566,6.92820323027551,8.30662386291808,7.74596669241483,4.12310562561766,6.6332
495807108,6.48074069840786,5.29150262212918,7.07106781186548,5.29150262212918,6.32455532033
676,7.21110255092798,9.1104335791443,5.47722557505166,6.32455532033676,6.85565460040104,6.92
820323027551,6.78232998312527,7.07106781186548,6.08276253029822,7.48331477354788,6.24499799
8398 3,7.74596669241483,6.16441400296898,7.21110255092798,10.6301458127347,6.85565460040104
KLQAAA,7.54983443527075,0,4.12310562561766,4.58257569495584,2.82842712474619,7.211102550927
98,4.58257569495584,6.40312423743285,4.69041575982343,7.93725393319377,8.83176086632785,5.09
901951359278,5.09901951359278,5.65685424949238,9.16515138991168,5.65685424949238,5.47722557
505166,8.06225774829855,7.48331477354788,3.74165738677394,3.87298334620742,5.65685424949238
,3,7.54983443527075,2.82842712474619,6.78232998312527,7.14142842854285,4.58257569495584,2.64
575131106459,8.30662386291808,8.36660026534076,7.54983443527075,2.64575131106459,8.66025403
784439,7.54983443527075,6.557438524302,4.58257569495584,3,4.47213595499958,2.44948974278318,
8.60232526704263,5.3851648071345,4.58257569495584,2.44948974278318,4.89897948556636,7.416198
48709566,8.66025403784439,5.56776436283002,4.89897948556636,5,3.74165738677394,7.2111025509
2798,7.34846922834953,3.46410161513775,3.46410161513775,8,8.83176086632785,5,8.7177978870813
5,3.87298334620742,4.47213595499958,6.78232998312527,3.74165738677394,6.92820323027551,5.291
50262212918,6.85565460040104,5.56776436283002,3.16227766016838,4.47213595499958,7.483314773
54788,8.42614977317636,7,4.24264068711929,7.07106781186548,8.60232526704263,7.2801098892805
2,3.74165738677394,6.70820393249937,2,4.47213595499958,6.92820323027551,3.87298334620742,4.7
9583152331272,6.08276253029822,7,7.87400787401181,6.557438524302,8.36660026534076,4,3.464101
61513775,8.48528137423857,8.18535277187245,6.08276253029822,4.79583152331272,9.591663046625
44,4.58257569495584,2.82842712474619,3.60555127546399,4.58257569495584,6.08276253029822,8.48
528137423857,3.74165738677394,3.60555127546399,7.54983443527075,9.05538513813742,3.16227766
016838,4.12310562561766,4.89897948556636,3.87298334620742,6.32455532033676,6.557438524302,8.
06225774829855,8.06225774829855,4.58257569495584,7.54983443527075,4.79583152331272,8.544003
74531753,6.78232998312527,7.54983443527075,7,7.48331477354788,3.3166247903554,6.40312423743
285,5.74456264653803,6.16441400296898,5.74456264653803,7.34846922834953,4.12310562561766,4.5
8257569495584,4.35889894354067,9.48683298050514,3.16227766016838