Cyclooxygenase-1 as an anti-stroke target: Potential inhibitor identification and non-synonymous single nucleotide polymorphism analysis.

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

Master of Science in Bioinformatics & Computational Molecular Biology
(Coursework and Thesis)
of

RHODES UNIVERSITY, SOUTH AFRICA
Research Unit in Bioinformatics (RUBi)
Department of Biochemistry & Microbiology
Faculty of Science

by
Tendai Muronzi
18M8622
February 2019

Abstract

Stroke is the third leading cause of death worldwide, with 87% of cases being ischemic stroke. The two primary therapeutic strategies to reduce post- ischemic brain damage are cellular and vascular approaches. The vascular strategy aims to rapidly re-open obstructed blood vessels, while the cellular approach aims to interfere with the signalling pathways that facilitate neuron damage and death. Unfortunately, popular vascular treatments have adverse side effects, necessitating the need for alternative chemotherapeutics. In this study, cyclooxygenase-1 (COX-1), which plays a significant role in the post- ischemic neuroinflammation and neuronal death, was targeted for identification of novel drug compounds and to assess the effect of nsSNPs on its structure and function. In a drug discovery part, ligands from the South African Natural Compounds Database (SANCDB-https://sancdb.rubi.ru.ac.za/) and ZINC database (http://zinc15.docking.org/) were used for high-throughput virtual screening (HVTS) against COX-1. Additionally, five nsSNPs were being investigated to assess their impact on protein structure and function. Three of these SNPs were in the COX-1 dimer interface. Molecular docking and molecular dynamics simulations revealed asymmetric nature of the protein. Several ligands, peculiar to each monomer, exhibited favourable binding energies in the respective active sites. SNP analysis indicated effects on inter-monomer interactions and protein stability.

Declaration of Authorship

I, **Tendai Muronzi**, declare that this thesis entitled, *Cyclooxygenase-1 as an anti-stroke target: Potential inhibitor identification and non-synonymous single nucleotide polymorphism analysis*, submitted to Rhodes University is solely my own research work. I have acknowledged all authors' concepts and referenced direct quotations from their works. I also declare that this thesis has never been submitted to any different institution for whatever degree.

Signature.....

Date.....

Acknowledgements

I would like to express my deep gratitude to my research supervisor Professor Özlem Tastan Bishop, for conceptualising this research work, and her constant guidance, contribution and encouragement.

Appreciation goes to Magambo Phillip Kimuda, for being a constant guiding hand; his advice and self-less assistance made it possible for this MSc thesis to be coherent, and successfully completed in time. Further gratitude goes to my coursework lecturers, who equipped me with the knowledge needed to tackle the research project.

I would also like to acknowledge our collaborators at the University of Maryland for providing the initial SNP dataset.

I am grateful for all the support from the entire Muronzi-Nyamatanga clan: To Taku, my sister, friend and confidant; thank you for always being a phone call away and believing in me when my confidence was failing.

To my aunt and uncle (Mr. and Mrs. Muvoti), thank you for being my proxy parents; a niece could not have felt better loved and supported.

To my parents,thank you for the unconditional love; and financial and emotional support. You have always believed in me and put all you could into making my dreams a reality.

I would also like to express my very great appreciation to my friends, Nicolette, Tanya, Vyla, Nyasha, Mpume, Linda, Denise and Amanda (of two decades). I could not have asked for greater support system of like-minded women.

Last but not the least, I want to express my gratitude to all members of the Research Unit in Bioinformatics (RUBi). To my officemates, Lorna, Margie and Phillip, thank you for the support and frequent laughs. I am also particularly grateful to Varaidzo for her invaluable support, love and friendship; sometimes all I needed was to vent in Shona. And to my MSc colleagues, I am so proud of us, we could not have made it without the constant support we gave each other.

Contents

List of Figures

List of Tables

List of Abbreviations

ACE2 angiotensin I converting enzyme 2.

AChE acetylcholinesterase.

ATB Automated Topology Builder.

BBB blood-brain barrier.

BC betweenness centrality.

BLAST Basic Local Alignment Search Tool.

CDD Conserved Domain Database.

COX cyclooxygenase.

EGF epidermal growth factor.

GBD Global Burden of Disease.

GROMACS GROningen MAchine for Chemical Simulations.

GWAS genome-wide association studies.

HMM Hidden Markov Model.

HTVS High-throughput virtual screening.

LINCS LINear Constraint Solver.

MAFFT Multiple Alignment using Fast Fourier Transform.

MBD membrane binding domain.

MM-GBSA molecular mechanics Generalized Born for solvent-accessible surface area.

MM-PBSA molecular mechanics Poisson-Boltzmann for solvent-accessible surface area.

MSA multiple sequence alignment.

MUSCLE MUltiple Sequence Comparison by Log-Expectation.

NCBI National Center for Biotechnology Information.

NMDAR N-Methyl-D-aspartic.

P2Y P2Y purinoceptor.

PAI-1 plasminogen activator inhibitor-1.

PAIN pan assay interference.

PANTHER Protein ANalysis Through Evolutionary Relationships.

PBC periodic boundary conditions.

PCA Principal Component Analysis.

PDB Protein Data Bank.

PGE2 prostaglandin E 2.

PGG2 hydroperoxide prostaglandin G 2.

PGH2 prostaglandin H 2.

PhD-SNP Predictor of Human Deleterious SNPs.

PME particle mesh Ewald.

PolyPhen POLYmorphism PHENotyping.

PRIMO PRotein Interactive MOdeling.

PROVEAN Protein Variation Effect ANalyzer.

PSD-95 postsynaptic density protein 95.

PSEP position-specific evolutionary preservation.

PSIC position-specific independent count.

PTGS1 prostaglandin endoperoxide synthase 1.

QMEAN Qualitative Model Energy ANalysis.

RMSD Root mean square deviation.

RMSF Root mean square fluctuation.

SANCDB South African natural compound database.

SASA solvent accessible surface area.

SCOP Structural Classification of Proteins.

SIFT Sorting Intolerant From Tolerant.

SNAP Screening for NonAcceptable Polymorphisms.

SNP single nucleotide polymorphism.

SNV single nucleotide variation.

SVM support vector machine.

TBX2 thromboxane A 2.

tPA recombinant tissue plasminogen activator.

VDW van der Waals.

WHO World Health Organisation.

ZINC ZINC Is Not Commerical

List of web-servers used

Web-server/ application	Web address
ATB	https://atb.uq.edu.au/
BBB filter	https://www.cbligand.org/BBB/
BLASTp	https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins
ConSurf	http://consurf.tau.ac.il/2016/index_proteins.php
HHPred	http://toolkit.tuebingen.mpg.de/#/
MAFFT	https://www.ebi.ac.uk/Tools/msa/mafft/
MUSCLE	https://www.ebi.ac.uk/Tools/msa/muscle/
PAINS filter	https://www.cbligand.org/PAINS/
PANTHER	http://www.pantherdb.org/tools/
PhDSNP	http://snps.biofold.org/phd-snp/phd-snp.html
PolyPhen	http://genetics.bwh.harvard.edu/pph2/
PredictSNP	https://loschmidt.chemi.muni.cz/predictsnp/
PRIMO	https://primo.rubi.ru.ac.za/
ProSA-	https://prosa.services.came.sbg.ac.at/prosa.php
PROVEAN	http://provean.jcvi.org/index.php
SIFT	https://sift.bii.a-star.edu.sg/www/SIFT_seq_submit2.html
RCSB PDB	https://www.rcsb.org/
QMEAN	https://swissmodel.expasy.org/qmean/
VAPOR	https://huma.rubi.ru.ac.za/#vapor
Verify3D	http://servicesn.mbi.ucla.edu/Verify3D/

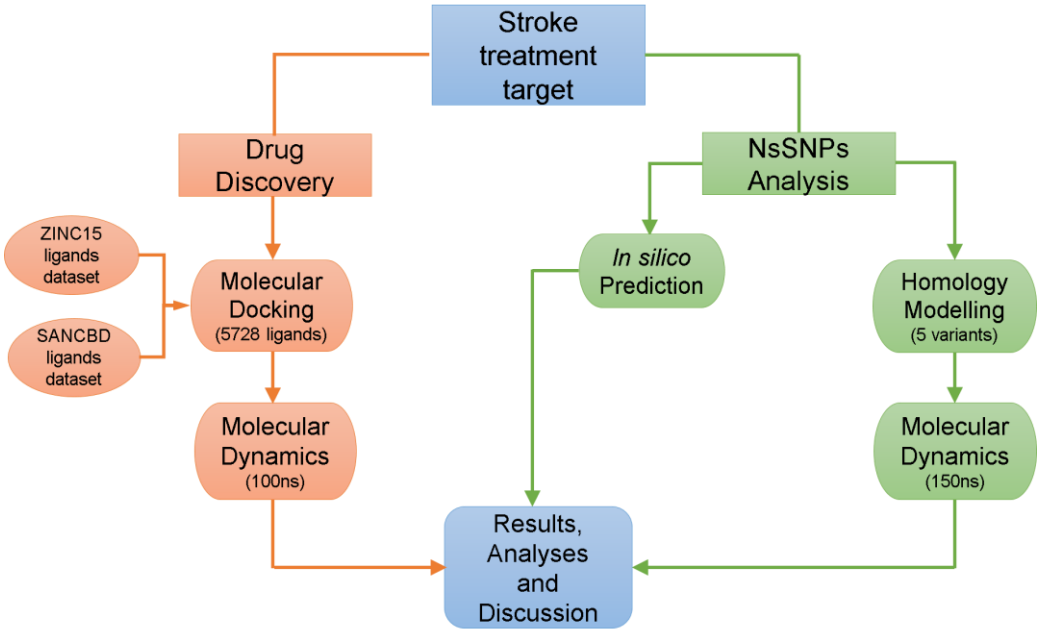
Table of Amino Acids

Name	Three letter code	One letter code
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamine	Gln	Q
Glutamic acid	Glu	E
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

Research Overview

An overview of research project is summarized in a flow diagram in the Figure below.

The research is divided into two parts, analysis of target protein variants and the search for inhibitors for the protein wild-type. Presence of SNPs can affect protein function and ultimately protein-drug interaction, as such, it will be worth investigating potential drug compounds discovered in the drug discovery part of the research on the variants.



1 Introduction

1.1 Literature Review

1.1.1 Stroke

Stroke is a disease that occurs when there is an interruption in blood supply to the brain, resulting in oxygen–deprivation and subsequent death of neurons in areas fed by the affected arteries. There are two main types of stroke, ischemic and haemorrhagic; due to obstructed or ruptured arteries, respectively. Transient ischemic stroke, often referred to as a mini stroke is due to a temporary blockage.

According to the World Health Organisation (WHO) in 2016 [1], stroke was the second leading cause of death worldwide, after heart disease, and the leading cause of disability. The American Heart Association [2] cites that strokes occur in the US once every 40 seconds, causing death every four minutes; while the UK Stroke Association cites an occurrence of approximately one stroke every five minutes.

Despite these statistics, during 2005, 87% of stroke deaths and disability adjusted life years (DALYs) [3] from stroke worldwide occurred in low to middle-income countries, which is about seven time more than in developed countries. From this the WHO projected that the overall stroke mortality would accelerate in these countries [4]. True to this projection there has evidence of a general decline of stroke in high income countries and an increase in low income countries [5].The study reported the highest ischemic stroke mortality rates to have been observed in Russia and Kazakhstan, with the lowest in Western Europe, North and Central America, Turkmenistan and Papua New Guinea.

According to the INTERSTROKE study [6], ischemic stroke accounts for 34% of stroke in Africa versus the 9% in the developed world. A study in 2015 [7] cited a stroke prevalence of rate of up to 923/100000 in Egypt. , while a separate study showed stroke being the leading cause of elderly admission in Sudan, Tanzania and Nigeria [8]. In Sub-Saharan Africa stroke has been shown to affect a younger age group [6]. Overall, observed population-based stroke prevalence rate in Africa sits at nearly 387.9/100 000 population [9]. These occurrences have a major impact on the growth and development of African economies.

Africa has presented particularly susceptible to stroke due to population growth, evolving industrialisation and increased consumption of western diets, which had led to the rise of many modifiable vascular disease risk factors [10].

1.1.1.1 Risk factors

Risk factors are attributes or variables that associated with an increased likelihood of disease of infection. Stroke risk factors can be divided into modifiable and non-modifiable factors. Modifiable factors include history of high blood pressure, diabetes mellitus and heart disease, while non-modifiable factors include family history of cerebrovascular diseases, age, sex, and ethnicity [11]. High cholesterol, smoking, obesity [12], high blood pressure and diabetes are leading causes of stroke. As such a large percentage of stroke occurrences can be prevented by correcting modifiable risk factors through making healthy life choices [13].

1.1.1.2 Treatment of ischemic stroke

The two primary therapeutic strategies that exist to reduce brain damage after ischemic stroke are a cellular and a vascular approach [14]. The aim of the vascular is the rapid re-opening of obstructed blood vessels, while the cellular aims to interfere with the signalling pathways that facilitate neuron damaged and death [15]. Efforts in vascular therapy include mechanical removal of the blockage or administration of thrombolytic such as, recombinant tissue plasminogen activator (tPA), anticoagulants antiplatelet drugs within 4.5 hours [16] of exhibiting stroke symptoms. The cellular approach, which entails administration of neuroprotective agents [17] [18], calcium channel blockers and free radical scavengers has however proven more complicated to tackle [19]. Although they exist and are in use, drugs for stroke treatment continue to have side effects. This may be due to lack of clarity on the mechanism of interaction.

Additionally, use of genome-wide association studies (GWAS) has become a tool to assess susceptibility of genes for stroke [20] , to assist in treatment [21].

1.1.2 Anti-stroke target: Cyclooxygenase (COX)

Protein targets of drugs are found in diverse locations throughout the body; many are secreted or transmembrane proteins, while others are found in specific subcellular locations. Receptors make up the largest class of drug targets, followed by enzymes and transporter proteins [22].

To date, protein targets for stroke treatment play variety of biological functions in the fallout from ischemia. Worth noting are glutamate receptors such N-methyl-D-aspartate receptors (NMDAR) [23], acetylcholinesterase (AChE), angiotensin I converting enzyme 2 (ACE2), P2Y purinoceptor 12 (P2Y12), postsynaptic density protein 95 (PSD-95) , peroxisome proliferator-activated receptor gamma (PPARγ), plasminogen activator inhibitor-1 (PAI-1) and cyclooxygenase (COX) enzymes [24], affecting different processes from excitotoxicity neurotransmission and synaptic plasticity to thrombosis and atherosclerosis. Also of interest are non-coding RNA [25] which have been studied for their role in regulation of gene function at the transcriptional and post-transcriptional level, thus providing neuroprotection Studies continue to be conducted to expand this list .

1.1.2.1 COX-1 structure and function:

COX-1, also known as prostaglandin G/H synthase 1, prostaglandin endoperoxide synthase 1, is a monotopic membrane protein that, in humans, is encoded by the *ptgs1* gene. The protein is a bifunctional enzyme involved in the committed step of prostaglandin synthesis from arachidonic acid (AA). Prostaglandins are autocrine signalling molecules mediating physiological processes such as inflammation and platelet aggregation. Prostaglandins produced from AA metabolism are shown in Figure 1.2.

There are two main isozymes of COX (COX-1 and COX-2) that share nearly 60% identity [26] and the same secondary structure, existing as homodimers [27].The monomers share an interface that covers an area of over 2600 $Å^2$, consisting of 22 molecular inter-monomer interactions [28].

The precursor structures of COX-1 and -2 consist of 600 - 604 amino acids, before cleavage of the signal peptides at the N-terminal [29]. Each monomer consists of an epidermal growth factor-like (EGF) domain, a membrane binding domain (MBD) and a C-terminal catalytic domain which takes up 80% of the protein. The EGF-like domain is made up of two anti-parallel β-sheets linked by disulphide bonds. The domain leads into the four α-helical MBD, which anchors the protein into the lipid bilayer [30]. The catalytic domain, containing conserved α-helical structures, accommodates a cyclooxygenase active site, which is the binding site for a majority of nonsteroidal anti-inflammatory drugs (NSAIDs); and a peroxidase active site containing a heme cofactor. Although physically separate, these sites are functionally and structurally complementary.

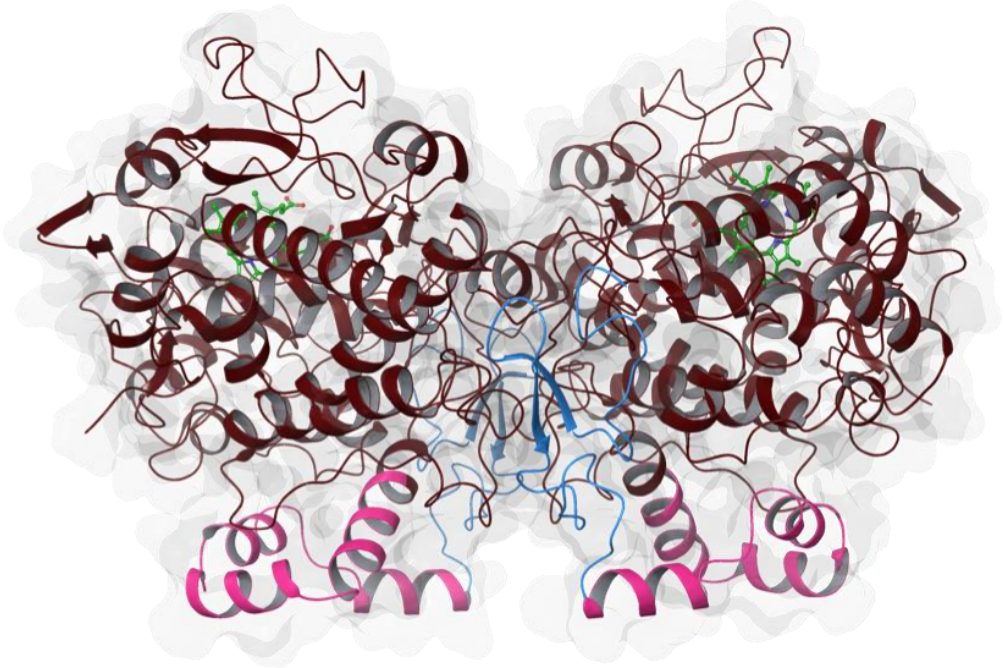


Figure 1.1: Structure of COX-1, where the EGF-like domain, the MDB and the C-terminal catalytic domain are shown in blue, pink and brown, respectively. The heme co-factors are shown in green.

AA is bis-oxygenated into unstable hydroperoxide prostaglandin G 2 (PGG2) in the cyclooxygenase site, then reduced to the prostaglandin H 2 (PGH2) in the peroxidase site. A long hydrophobic channel spans from an entrance at the MDB to the core of the catalytic domain, with Arg-119, Tyr-384 and Ser-529 being exceptions to the hydrophobicity. Residues Arg-119, Tyr-354 and Glu-523 [31] separate the cyclooxygenase active site from the entrance to the channel. The AA binding site occurs from Arg-119 to Tyr-384, where Arg-119 form a salt bridge or hydrogen bond with the substrate; Ser-529 coordinates it and Tyr-384 forms tyrosyl radical responsible for initiating the cyclooxygenase reaction. Ser-529, at the centre of this active site, is additionally the site of acetylation by aspirin which results in the irreversible and reversible inhibition of COX-1 and COX-2, respectively. The peroxidase site is essentially a cleft located furthest from the MBD, in which His-206 is a crucial proton donor and the heme co-factor iron metal is coordinated by His-387.

The COX active site channel is larger in COX-2, due to three amino acid differences between the isozymes. Two isoleucine/valine substitutions in COX- 2 are chief in this, one at position 522 that causes a structural modification that affords access to an additional side pocket, and another at position 433 that al- lows neighbouring Phe-517 to be flexible, further exposing said side pocket. Access to the side-pocket is the foundation of COX-2 drug selectivity. A third amino acid difference changes the chemical environment of the active site. COX-2 has an arginine in place of histidine in position 512 (His-512) of COX- 1, which can interact with polar compounds, further affecting the selectivity of possible inhibitors [32] [33] .COX-2 specific inhibitors that have been developed are often called coxibs.

1.1.2.2 Monomer asymmetry

Contrary to the assumption that both monomers are active simultaneously, recent studies suggest that substrate or inhibitor binding in the cyclooxygenase active site of one monomer impedes binding of another molecule in the other. COX isoenzymes are, therefore, sequence homodimers and conformational heterodimers [34] [35] [36].

The first evidence of possible asymmetric behaviour of the COX monomers was provided by Kulmacz and Lands [37] who found that NSAIDs inhibited COX- 1 at a stoichiometry of one NSAID per dimer, and that the COX-1 heme binding sites had slightly different affinities for heme [38]. Maximal cyclooxygenase catalysis occurs when the site with higher affinity was occupied by heme.

The monomer with the higher heme affinity performs catalytic operations and is designated E_{cat}. Its allosteric partner, E_{allo}, likely affords stability and plays an enabling role [39] [40].The catalytic efficiency of E_{cat} is governed by its interaction with E_{allo}, which differs per substrate or inhibitor binding to the allosteric monomer [36] [39] [41] [42]. Despite the conformational asymmetry, both monomers are necessary for function as monomeric species of the enzymes have proven to be inactive.

1.1.2.3 1 Role of COX-1 in ischemia:

Neuroinflammation is a defence response to pathogenic events or traumatic injury, such as a stroke; but is also recognized as a major contributor to neurological damage [43]. COX enzymes are the key and rate-limiting enzymes in the synthesis of prostaglandins, which are lipid metabolites that are involved in several physiological and pathological processes, including neuroinflammation. COX metabolites, such as thromboxane A 2 (TBX2) and prostaglandin E 2 (PGE2) [Figure 1.2] contribute to post-ischemic cerebral blood flow reduction, brain oedema, inflammation, and neuronal damage [44].

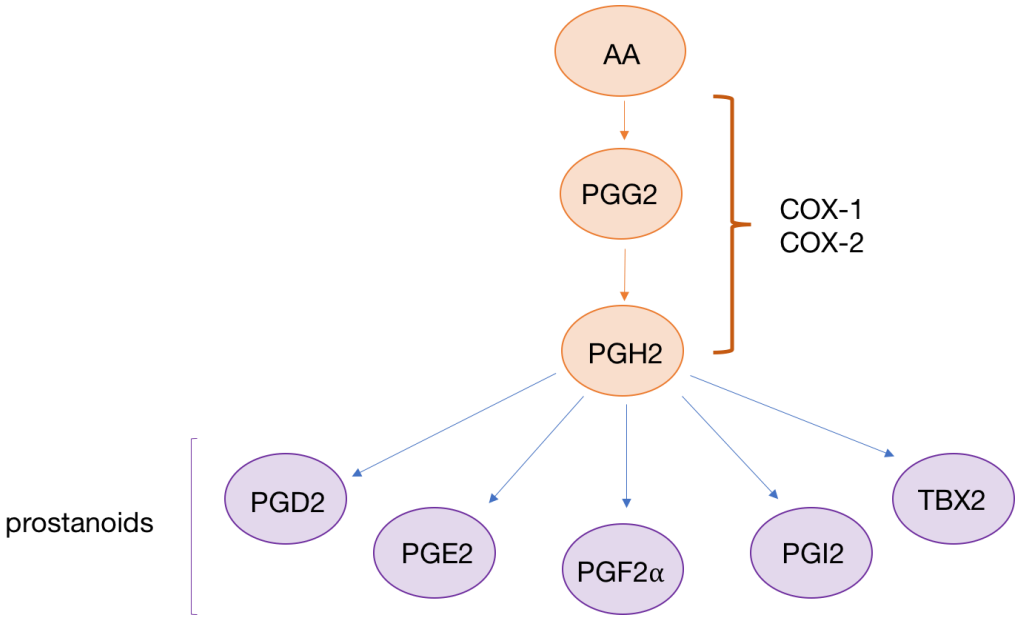


Figure 1.2: COX metabolic pathway of arachidonic acid (AA).Prostanoid products are thromboxane (TBX2), prostacyclin (PGI2),Prostaglandin F 2α (PGF2α), prostaglandin E2 (PGE2), prostaglandin D2 (PGD2)

COX-1 has been **classically** considered to be responsible for homeostatic prostaglandin synthesis [45] as it is constitutively expressed in most tissues. In the central nervous system, COX-1 is largely expressed in the microglial cells, which are chief protagonists in the cascade of events leading to tissue injury after cerebral damage [46]. This suggests a pro-inflammatory role of COX-1 in the post-ischemic fallout [47] [48] [49].

Although the actual mechanisms by which COX-1 is involved in neuroinflammation, its pharmacological inhibition and attenuation of microglial activation [47] reduces the inflammatory response and neuronal loss [50].This therefore qualifies COX-1 as a probable target for neuroprotection.

1.1.3 Related studies

Due to the high mortality brought about by stroke, several studies continue to be conducted to find novel drugs and protein targets.

The P2Y1 receptor (P2Y1R), which facilitates platelet aggregation presents potential anti-thrombotic drug target. A docking study was performed to screen for other ligands to act as nucleotide antagonists of P2Y1R in the Traditional Chinese Medicine Systems Pharmacology Database and Analysis Platform (TCMSP) to identify potential anti-thrombotic drugs from natural medicinal plants. The study yielded compounds and plant chemical constituents which were not previously widely used for anti-thrombosis treatment [51].

Additionally, various drugs anti-thrombotic, thrombolytic, neuroprotective, and anti-neuroinflammatory drugs are undergoing or have completed phase I and II trials for the treatment of acute ischemic stroke [52]. New targets in excitotoxicity, the immune system and the blood brain barrier are being exploited using novel NMDAR antagonists, calcium channel blockers, anti-inflammatory agents, neurotropic factors, monoclonal antibodies, as well as non-coding RNAs [53].

While not stroke treatment studies, studies have also been conducted into inhibition of COX-1, with some targeting parasite invasion and cancer [54] induced inflammation. Several of these use aspirin covalent inhibition [55] as a reference [31] [56] .Similar to the current study, Dewi [57] has screened for potential COX inhibitors based on natural products.

1.2 Problem Statement

Stroke is a global problem, which does not discriminate against any ethnic groups. In 2013 the global estimation of incidence was 10.3 million new strokes, with 67% [11] of these being ischemic strokes; making it a leading cause of disability worldwide.

Popular medication to treat ischemia, not only has a time-limit, but has side-effects. tPA is one such treatment with an associated risk of intra–cerebral haemorrhage. Studies show that every hour delay in treatment results in an almost 8% decrease in functional independence of victims [58]. Additionally, susceptibility to stroke varies according to ethnicity, sex and age, making it probable that medication may have varying efficacies across the different population groups.

For these reasons, there is a pressing need for more efficient, less toxic stroke therapeutic approaches, based on a cell apoptotic (and necrotic) processes and tissue repair; that factor in variations target protein structure.

Single nucleotide polymorphism (SNP) analysis of anti-stroke targets presents a method to improve drug design factoring in populations living with the target variants. This study attempts to analyse effects of selected non- synonymous SNPs on protein structure and subsequently, function. Compound screening is performed to identify possible target protein inhibitors for design and development of novel stroke drugs.

1.3 Research Aims

The first aim of the research is the analysis of structural variations of COX-1 due to non-synonymous single nucleotide polymorphisms (SNPs).The second is drug discovery, by screening compounds against the COX-1 wild-type, for possible inhibitor design. Analysis of variants will allow better understanding of the protein, and pave way for further studies into variant targeted inhibitors, guided by wild-type drug discovery study.

Objectives:

1. Identification and *in silico* analysis of nsSNPs to understand the effects on protein structure and function.
2. Homology modelling of COX-1 homologs for further *in silico* studies
    (a) Modelling of COX-1 wild-type for molecular docking and molecular dynamics analysis
    (b) Modelling nsSNP variants for molecular dynamics analysis to further understand effect on protein structure and function
3. Molecular docking of ligands against COX-1 wild-type for discovery and identification of potential drug compounds
    (a) SANCDB subset of natural compound ligands
    (b) ZINC subset of synthetic compound ligands
4. Molecular dynamics of homologs and protein-ligand complexes for analysis protein-protein and protein-ligand interactions
    (a) Wild-type molecular dynamics to analyse protein motions and residue interactions; and use as a reference for variants.
    (b) NsSNP variants molecular dynamics to analyse effect of the SNPs on protein behaviour relative to the wild-type.
    (c) Protein-ligand complexes molecular dynamics to analyse docked ligand stability and protein-ligand interactions.

2 In silico nsSNP Prediction

2.1 Introduction

A single point mutation in a gene can alter pre-mRNA splicing, amino acid sequence and ultimately protein structure [59]. Such mutations in nucleotide sequence are called single nucleotide variations (SNV). When a SNV is observed in a population with at frequency of at least 1% [60], it is then referred to as a single nucleotide polymorphism (SNP). SNVs alone account for almost 90% of genetic variations observed in humans [61].

SNPs which do not alter the amino acid sequence of resultant protein are referred to as silent or synonymous mutations, whereas those that do are referred to as non-synonymous mutations [62] [63]. Non-synonymous SNPs (nsSNPs) can be further divided into missense and nonsense mutations. In missense mutations, the nucleotide difference results in an amino acid substitution, while it results in the introduction of a premature stop codon in nonsense mutations, giving a truncated polypeptide sequence and ultimately non-functional proteins [64] [65] . Missense mutations can also immensely alter protein structure and function [66] in various ways, such as, disrupting folding and stability, altering binding site physiochemical properties and modifying structure flexibility.

Protein structure plays a major role in drug design and development [67], as such mutations that cause a change in structure can confer patient variability in drug response. Missense mutations occurring in a drug target protein may have a pharmacodynamics [64] effect, while those occurring in a protein involved in drug metabolism may alter drug pharmacokinetics [64] [65].

It is assumed that SNPs occurring in highly conserved functional regions are more likely to directly disrupt protein activity [68]; however, those occurring in less conserved regions can still affect protein folding, dynamics, stability and activity. Thus, accurately predicting the effect of a SNP, without delving into protein structure is near impossible.

Apart from their influence and relevance on drug discovery, SNPs can also be used to determine the susceptibility or infer protection of individuals to certain diseases [69]. GWAS can be applied in the analysis of phenotype and mutation associations [70].SNPs are thus significant in clinical research and drug development. As such, the ability to distinguish between protein function-altering nsSNPs and functionally neutral ones is important for prioritising research. However, due to the prevalence of nsSNPs [71] per gene, it is not always financially and temporally feasible to carry out wet laboratory experiments to determine their biological significance. For this reason, bioinformatics SNP prediction tools are used by researchers for initial screening to identify potentially deleterious SNPs.

2.1.1 SNP prediction

Deleterious SNP prediction attempts to determine SNP effect using structure and sequence-based approaches. Protein sequence-based predictions identify important residues using homology and evolutionary conservation information [72] [73]. This approach has an advantage over structure-based as protein sequence information is more readily available for all proteins than three- dimensional structural information. The structure-based approach concentrates on the protein mechanistic attributes and stability, providing an understanding of phenotypical effect of the SNPs. SIFT [74], SNAP [75], MutPred [76] and PANTHER [77] are examples prediction tools that use the sequence-based approach, while ERIS [78] and CUPSAT [79] use the of structure-based approach. An integration of both approaches such as used in PROVEAN [80], PolyPhen [81] and MUpro [82] has the obvious advantage of cross-referencing the results from both approaches, thereby providing a wider coverage of the different aspects of SNP analysis.

Unfortunately, none of the methods are perfect, nor do they always return similar results. This is due to the different algorithms used per tool, and the varying datasets used to train them. The choice in training data can significantly affect SNP classification and estimated error rates [83].

To tackle this problem, the best approach is to get a consensus from several different prediction tools before proceeding with further SNP analysis.

Web servers such as VAPOR [84] which merges predictions from several tools into a table, and PredictSNP [85] which combines the predictions to generate a consensus score, can be used for comprehensive prediction results.

2.1.2 SNPs in COX-1

According to literature, COX-1 exhibits high variability [86], which may be responsible for the clinical evolution of several diseases and adverse drug reactions [87] [88] [89] [90]. The variants concentrated on in this study P126T, N143K, L237, R244W and I557T, are listed in Table 2.1 and positions of the SNPs in the protein highlighted in Figure 2.1. While these variations were selected based on their location on the protein to better assess monomer interaction as well as active site activity; some such as L237M, are already of wet-lab interest.

P126T lies in monomer crosstalk region. A study substituting residue in this position with a cysteine [41] exhibited cross-linking, but decrease in COX activity. Given these wet-lab observations it was of interest to see how an amino acid substitution in said position would influence change in protein motions. N143K on the other hand occurs at a well conserved site of glycosylation [91] that potentially plays a number of roles in the protein.

L237M has been studied for its effect on drug efficacy in COX-1 [92] [93], and possible susceptibility to stroke [94].Of interest is the role L237M may play in aspirin intolerance, in conjunction with other allele variants in some cases [92] [95]. Asian populations have been widely studied for polymorphism induced resistance to aspirin

[95] [96] which is at times used as anticoagulant in post-ischemic treatment.

To date no wet-lab work, or extensive literature exists on R244W and I557T. However, the two substitutions occurred in areas of interest to the study.

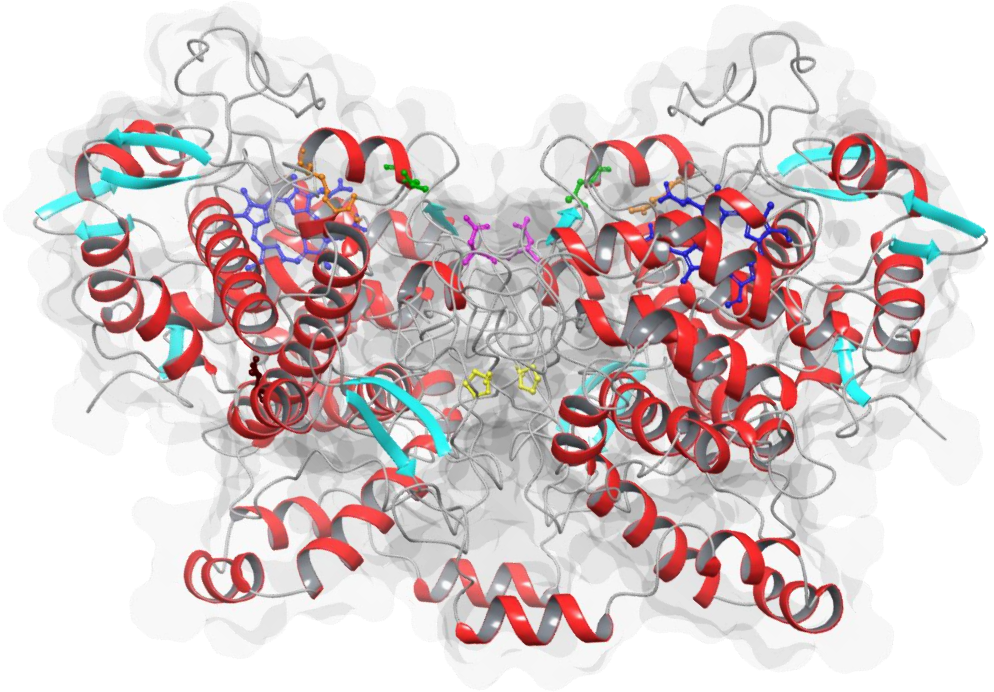


Figure 2.1: SNP positions mapped onto COX-1 wild-type; P126T in yellow, N143 in pink, .L237M in green, R244W in orange and I577T in maroon.

2.2 Methodology

2.2.1 Data retrieval and query submission

The SNP dataset used in the study is shown in Table 2.1 below. The initial dataset was provided by collaborators at the University of Maryland (UM), and additional SNPs were obtained from Ensembl [97] (release 93) based on proximity to dimer interface. SNPs chosen were located in either the catalytic domain or the dimer interface, to analyse effect on protein catalytic activity or dimer interactions, respectively. Most SNPs fell under both domains.

The query sequence and amino acid substitutions were given as input for the several prediction servers. Default parameters were used in all instances.

Table 2.1: Non-synonymous SNP dataset associated with stroke drug target, COX-1.

dbSNP rsNum	Allele	Position	Classification	AA Substitution	Domain	Source	Population	Frequency
rs147795437	C>A	chr9:122378798	missense	P126T	catalytic/ dimer interface	Ensembl	global	A=0.00001 (1/125568, TOPMED)
rs368871343	C>A	chr9:122378851	missense	N143K	catalytic/ dimer interface	Ensembl	global	A=0.000 (1/5008, 1000G)
rs5789	C>A	chr9:122381694	missense	L237M	catalytic/ dimer interface	UM collaborators	global	A=0.01841 (2312/125568, TOPMED)
rs201936137	C>T	chr9:122381694	missense	R244W	catalytic	UM collaborators	global	T=0.00039 (49/125568, TOPMED)
rs20101676	T>C	chr9:122392414	missense	I557T	catalytic/ dimer interface	Ensembl	global	C=0.00002 (4/246246, GnomAD)

2.2.2 Programs used to predict SNP effect

The varying programs used for SNP effect prediction and the algorithms they use are listed in Table 2.2.

2.2.2.1 PhD-SNP

The Predictor of human deleterious SNPs (PhD-SNP) method utilises support vector machine (SVM) classifiers [98] based on sequence information and conservation. Three slightly different algorithms exist, classifying variations from neutral to disease-related [98]. The tool output is a prediction of whether an amino acid substitution is a neutral or deleterious polymorphism.

2.2.2.2 PolyPhen-2

Polymorphism Phenotyping v2 (PolyPhen-2) predicts the consequence of amino acid substitutions on protein structure and function using a naïve Bayes classifier [81] [99].

This utilises multiple sequence alignments and protein structural properties to estimate the position-specific independent count (PSIC) score for every variant and calculates the score difference between them [100].

2.2.2.3 SIFT

The Sorting Intolerant from Tolerant (SIFT) algorithm [101] predicts the potential impact of amino acid substitutions on protein function. Sequence homology is used to assess the likelihood of adverse effects from amino acid substitutions. It is based on the idea that evolutionarily conserved regions are likely to be less tolerant of mutations; therefore, making amino acid substitutions, insertions or deletions in these regions are more probable to affect function.

The query sequence is aligned with homologous sequences, and a normalised probability score of observing a different amino acid in the positions of interest is computed.

2.2.2.4 PROVEAN

The Protein Variation Effect Analyzer (PROVEAN) algorithm predicts the functional impact for all classes of protein sequence variations. In addition to single amino acid substitutions, the algorithm is capable of handling in-frame insertion, deletions, and multiple substitutions [99] enabling a higher accuracy of prediction.

The algorithm entails calculation of a delta score from a set of homologous sequences. The PROVEAN score, which is an unbiased average delta score is then computed and used to make the predictions. The scores correlate with biological activity level and can therefore be used to gauge the degree of functional impact of a protein variation.

2.2.2.5 PANTHER

Protein analysis through evolutionary relationships (PANTHER) analyses protein sequence evolutionary information using multiple sequence alignments and Hidden Markov Model (HMM) libraries to predict SNP effect [102]. A position specific substitution score called subPSEC is used to estimate the functional impact of any particular nsSNP on the protein [102].

PANTHER additionally classifies proteins by function, further refining SNP prediction. The output generated by the tool is a probability of the queried variant being deleterious.

2.2.2.6 SNAP

Screening for nonacceptable polymorphisms (SNAP) uses neural networks and improved machine-learning methodologies to predict the functional effects of SNPs. Functional and structural annotations of the sequence, and biophysical and evolutionary characteristics are used for prediction of gain or loss in protein function [75]. A reliability index which measures the accuracy of the prediction is also included in the output, allowing further filtering of predictions.

2.2.2.7 PredictSNP

PredictSNP is a consensus classifier, that integrates tools such as MAPP [103], nsSNPAnalyzer [104], PANTHER, PhD-SNP, PolyPhen-1, PolyPhen-2, SIFT and SNAP. The six best performing of these tools are used to give a consensus prediction [85]. Confidence scores of the integrated tools incorporated into the final score which is finalised by calculation of the PredictSNP confidence score

2.2.2.8 VAPOR

The Variant Analysis Portal (VAPOR) [] tool is incorporated into the HUMA [105] web server from the Research Unit in Bioinformatics (RUBi).The tool compiles predictions from PROVEAN, PolyPhen-2, PhD-SNP for functional impact; and I-Mutant 2.0 [106] and MUpro [82] for change in protein stability. The two types of predictions are each presented in a separate table.

2.2.2.9 ConSurf

ConSurf [107], while not a SNP prediction tool, used for predicting functional regions in proteins based evolutionary conservation. The server automatically collects homologues based on the query sequence, generates an MSA and constructs a phylogenetic tree that reflects evolutionary relation. This information is used to estimate the evolutionary rates of each position in the protein, using a probabilistic framework.

Resulting conservation information can then be used to predict whether a substitution in any specific position would be tolerated.

Table 2.2: Methods implored in different *in silico* SNP prediction tools.

Prediction tool	Method
PhD-SNP	SVM based
PolyPhen-2	Naïve Bayesian based
SIFT	Alignment score – MSA based
PROVEAN	HMM based
PANTHER	Alignment score - HMM based
SNAP	Neural network based
PredictSNP	Consensus classifier
VAPOR: MUpro	SVM based

.

2.3 Results and Discussion

In this study five SNPs located in the catalytic domain or dimer interface of COX-1 were identified for analysis. The SNPs were subjected to *in silico* prediction using various prediction tools to provide comprehensive insight on the possible effect of each SNP on protein structure, function and stability. Results of the *in silico* analysis are shown in Table 2.3. Outputs and scoring systems of each of the tools differ, as such they were interpreted based on set thresholds to assign an overall prediction.

Table 2.3: Predictions and scores on effect of nsSNPs of interest on COX-1, from various programs with distinct algorithms.

Substitution	PANTHER _PSEP	PROVEAN Score	SIFT Score	PhD-SNP Prediction	PhD-SNP RI	pph2_score	SNAP Prediction	SNAP RI	PredictSNP Prediction	PredictSNP Confidence	MUpro ΔΔG
P126T	911	-7.663	0.0	Disease	4	0.995	Non-neutral	2	Deleterious	87%	-1.112580
N143K	750	-5.462	0.0	Disease	9	0.999	Non-neutral	2	Deleterious	87%	-1.561699
L237M	456	-1.259	0.12	Neutral	5	0.2	Non-neutral	0	Neutral	68%	-0.6630431
R244W	1628	-7.768	0.0	Disease	8	1	Non-neutral	5	Deleterious	87%	-0.6687723
I557T	176	-4.245	0.07	Disease	5	0.972	Non-neutral	1	Deleterious	87%	-1.805353

2.3.1 PhD-SNP

PhD-SNP does not output any score values, but instead prediction of whether the variant is disease-related or neutral, with an associated reliability index (RI) out of ten. In Table 2.3, PhD-SNP ranked all SNP, except L237M as disease causing. The RIs of N143K and R244W were high, ascertaining the accuracy of the predictions. P126T and I557T score intermediately in this respect, with P126T having the lowest RI, suggesting a lower impact of the protein.L237M was classified as neutral with a median RI, and thus a slight possibility of disease causing effect.

2.3.2 PolyPhen-2

The PolyPhen-2 score (pph2) ranges from 0 to 1, where the former is benign and the latter deleterious. All SNPs were predicted as deleterious, with R244W ranking highest, implying it is the most deleterious; and L237M the least damaging with its low ranking.

2.3.3 SIFT

The output scores of SIFT range from 0 to 1, with 0 being deleterious and 1 being neutral [108], allowing a quantitative comparison of the biological significance of SNPs. Additionally, a score between 0 and 0.05 is predicted to affect protein function. Based on the scoring system, all five SNPs of interest were predicted as deleterious,

scoring below 0.5. L237M and I557T however scored above 0.05 [Table: 2.3], suggesting no adverse effects on COX-1 function.

2.3.4 PROVEAN

The default threshold score is -2.5, where variants with a score above it are considered neutral and those equal to or below are deleterious. As seen in the PhD-SNP predictions in Table: 2.3, all the SNPs but L237M scored below –the threshold and were therefore classified as deleterious. L237M scored above -2.5, classifying as neutral.

2.3.5 PANTHER

PANTHER scoring is based on PSEP (position-specific evolutionary preservation) measured in million years (my).The thresholds used in the output are "probably damaging", "possibly damaging" and "probably benign" for a time greater than 450 my , between 450my and 200my, and less than 200my respectively.

Based on these thresholds, I557T fell in the probably benign range, denoting that the position is poorly preserved; while the remaining SNPs were comfortably classified as probably damaging. L237M narrowly classified as probably damaging, meaning the substitution in the position is likely more tolerable than the rest.

2.3.6 SNAP

While all SNPs of interest were classified as non-neutral, the RIs, which in this case are both a measure of accuracy and a reflection of the degree of functional effect of a substitution, differed. As in the pph2 [Table: 2.3] prediction, L237M and R244W were at opposite ends of the scoring range signifying L237M had the highest probability to affect protein functionality and R244W the lowest.

2.3.7 Predict SNP

The PredictSNP prediction for the classified all the SNPs but L237M as deleterious with a confidence score of 87% The L237M had a lower confidence score of 68%. The raw output from the PredictSNP, in Supplemental Table 1 showed a general low accuracy score for L237M predictions across all the tools in relation to the other SNPs. These attributed to the final accuracy score. Six out of eight of the tools used maintain that L237M was neutral. The SNAP and SIFT predictions were in agreement with the standalone analyses shown in Table 2.3 The consensus generated by PredictSNP was a fitting reflection of the prediction of all the tools.

2.3.8 VAPOR

As some of the tools integrated in VAPOR are already covered in previous sections, only predictions peculiar to VAPOR were analysed.

The tools concentrating on protein stability, MUpro and I-Mutant 2.0, were concentrated on. SNPs can cause changes in the internal energy of proteins, affecting structure and stability. As a result change in Gibbs free energy ($\Delta\Delta G$) can be used to measure effect of a SNP on protein stability [109]. Both MUpro and I-Mutant 2.0predict change in $\Delta\Delta G$ using SVMs and state of whether the change will result in increase or decrease of stability.

MUpro predicted decrease in $\Delta\Delta G$ for all the SNPs, while I-Mutant 2.0 returned partial results which were therefore discarded. Although the degree of energy change varied for each SNP, all were predicted to destabilise the protein; this is due to the direction of energy change being more relevant than the magnitude [110]**.**

2.3.9 ConSurf

Much like the PANTHER PSEP predictions, the ConSurf results shown in Figure 2.2, are based on evolutionary conservation. Evolutionary conservation of any amino acid in a protein implies intolerance to change due to the influence of the residue in maintaining structural integrity and function [111]**.** As such, SNPs located in conserved regions are likely to have more deleterious effects than those falling in regions tolerant of variation.

Residue positions in the ConSurf output amino acid sequence were classified from variable through to conserved, to varying degrees, using a nine-colour pallet. Additional information on residue solvent accessibility and functionality is provided, differentiating between buried region in the core of the protein and those exposed on the surface. Buried residues tend to be structurally important, while the exposed are functionally important [112]**.** The combination of position conservation and solvent accessibility information provides a more exhaustive view of importance of said position in the protein.

ConSurf results assigned positions 126, 143 and 244 to conserved regions, predicting them to be exposed and therefore functionally important. Positions 237 and 557 were predicted to be buried, with 557 being the least conserved followed by 237.

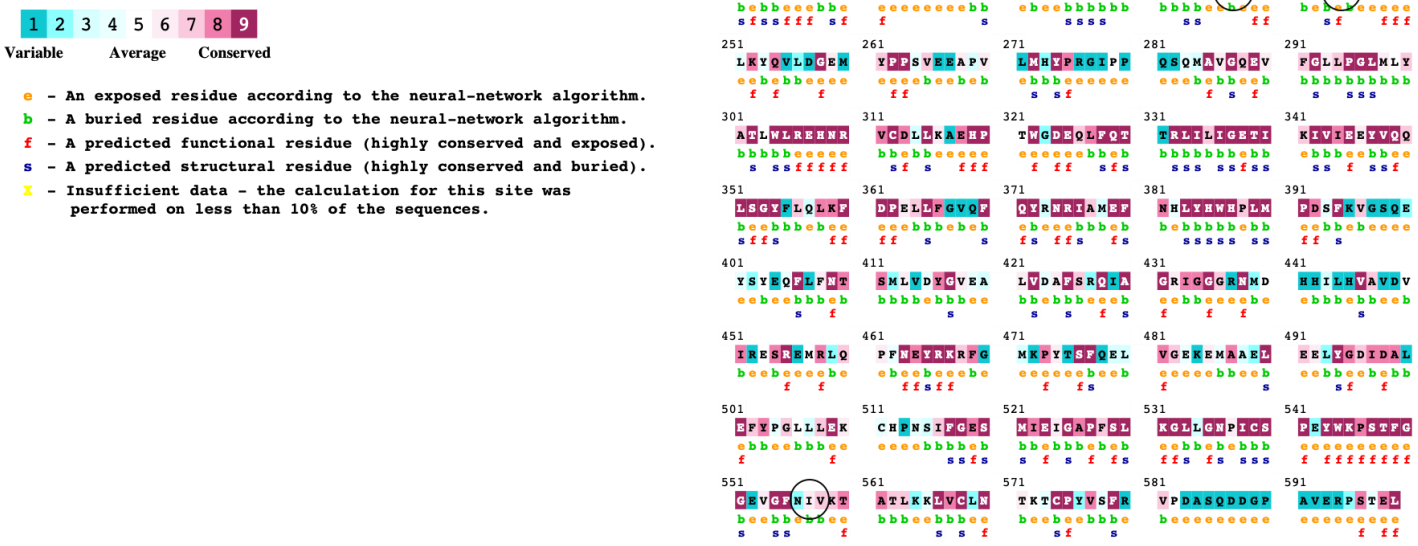


Figure 2.2: ConSurf results showing varying degrees of conservation for each position in the amino acid sequence, using a colour code ranging from 1-9. Positions of SNPs in the experimental dataset are highlighted by a black circle

2.4 Conclusion

The aim of this chapter was to assess the potential effect of SNPs in COX-1, P126T, N143K, L237M, R244W and I557T on protein structure, function and stability using *in silico* methods. A summary of the predictions is presented in Figure 2.3 as a histogram.

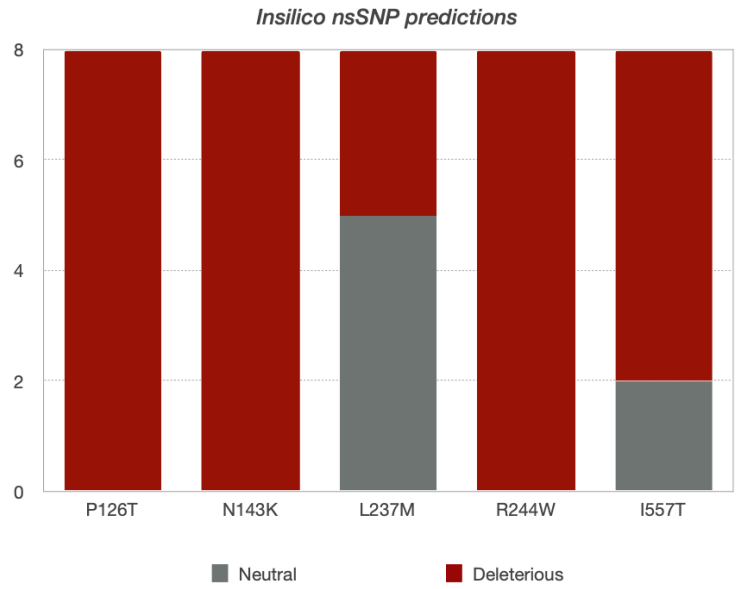


Figure 2.3: Compilation of results from SNP prediction tools PhD-SNP, PolyPhen-2, SIFT, PROVEAN, PANTHER, SNAP, PredictSNP and MUpro.

Prediction tools were in unanimous agreement for three out of the five nsSNPs analysed. ConSurf conservation results [Figure 2.2] conform to the prediction tools, where consistently deleterious SNPs, P126T, N143K and R244W, lie in highly conserved regions; while L237M and I557T, do not. Predictions for the latter are contradictory.L237M is cited in literature [87] [94] [95], to possibly have adverse effects on COX-1 function, which is not necessarily corroborated by all the prediction tools.

Since prediction tools are not infallible, beyond getting a consensus, further analysis needs to be conducted. This can be conducted in a wet-lab point mutation experiment followed by functional assays [113] or *in silico* by introducing the SNPs to the structure and subjecting said structure to a molecular dynamics (MD) simulation [114].

3 Homology Modelling

3.1 Introduction

Structure prediction can be achieved using either ab initio or template-based approaches. The template-based approach is further divided into threading and homology modelling, where the latter is most accurate approach [115]. The main aim of protein structure modelling is calculating a model comparable to what would be achieved experimentally. *In silico* structure prediction modelling generates protein models for varying purposes such as structure-based drug design [116], protein function analysis, antigenic behaviour analysis, and rational design of proteins with increased function or stability. It provides an alternative solution in situations where experimental techniques like NMR analysis or X-ray crystallography have failed.

3.1.1 Steps in homology modelling

Homology modelling is based on the notion that evolutionary related proteins are highly likely to share conserved structural features, despite differences in primary amino acid sequence [117] [118]. A study by Illergård [119] and colleagues of has shown that protein structure can be ten times more conserved than sequence.
The process of homology modelling consists of four main steps, template selection, target and template alignment, model construction and model assessment. These steps can be repeated until a satisfactory model is built.

Template selection The very basis of homology modelling is dependent availability of homologous proteins, as such accuracy of a model generated is reliant on the template used. Sequence identity, which is the percentage of amino acids that match between a target and template sequences [120], is one of the major considerations when picking a template for homology modelling. It has generally been established that proteins of similar length, that share a sequence identity above 40%, tend to adopt similar structure [121], though exceptions to this rule exist [122] [123]. This region above 40% identity is dubbed the homology modelling "safe zone". According to Rost [124], sequence identities between 20-35% lie in what is called the “twilight zone” where homology between proteins is precarious. Here significant attention needs to be given to sequence coverage and length. Recent studies have shown that homology models can be generated on a sequence identity of as low as 20% [125]. Search tools, such as BLAST and HHPred, which use pairwise alignment to match homologous sequences, can be used for template identification.

Evaluation of template quality is also crucial to the template selection process. Features such as R-factor and resolution, as well as presence of co-crystallised ligands need to be considered [126]. To aid with this evaluation, the Protein Data Bank (PDB)

provides a summary of structure quality with its entries.

Sequence alignment Once a suitable template has been identified, the next step is mapping target amino acid sequence onto the template structure. This is done by aligning the target and template sequences. While a pairwise alignment can be used in cases of high sequence identity, the method is not always reliable [127]. As such it is necessary to conduct generate an MSA as it is more sensitive in detecting evolutionary relationships among proteins and genes [128] .MSAs are useful in homology modelling, to place deletions or insertions in regions where the sequences are highly divergent, and map hydro-phobic and -philic regions.

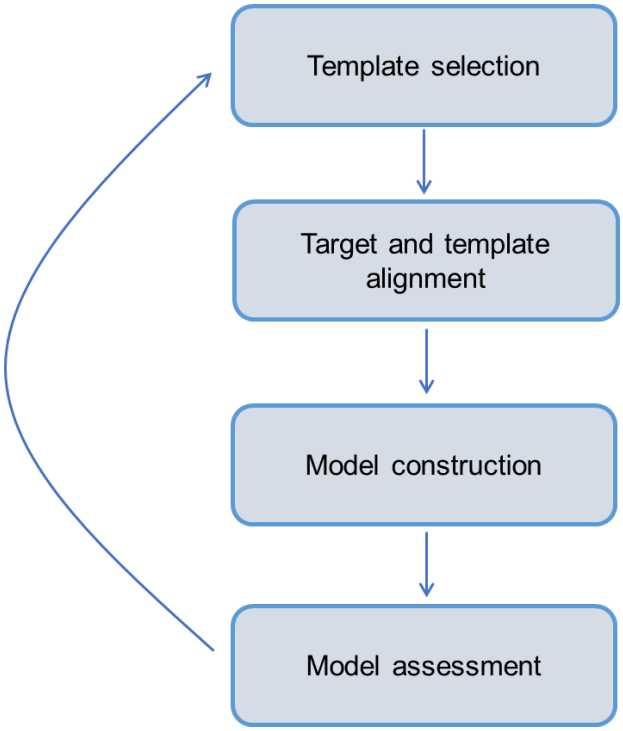


Figure 3.1: Flowchart of the four main steps in homology modelling: template selection, alignment, model construction and model assessment. The steps are reiterated if the model produced is not acceptable.

Alignment errors tend to be the main cause of deviations in homology modelling even when the correct template is chosen, hence it is at times necessary to generate several MSAs and manually edit them.

Numerous programs exist to date to construct MSAs, using different algorithms such as MAFFT [129], Clustalω [130], MUSCLE [131] and T-Coffee [130]. To generate accurate MSAs these program algorithms employ either dynamic or heuristic programming, with most shying towards heuristic .Heuristic methods, that are classified as either be progressive or iterative, are used to achieve this.

Model construction Model generation can be further broken down into back- bone

generation, loop modelling, side-chain modelling and model optimisation [126]. Backbone generation essentially copies template residue coordinates where they match the target and backbone coordinates where they do not. It is on this skeleton were the rest of the model is built. Gaps and insertions are filled in during loop modelling.

Loops are considered as the most variable regions of a protein and often determine the functional specificity of a protein structure [132] .Loop modelling is, therefore, a paramount to determining the usefulness of homology models. Protein side chains exist in energy conformations called rotamers. In side chain modelling, side chains are fixed onto the backbone coordinates and rotamers are selected based on target protein sequence and the given backbone coordinates, by using defined energy functions and search strategies [133] [134].

Even with all the pieces in place, the model produced may have still be conformationally inaccurate, with errors in bond lengths and angles. Model refinement aims to resolves these issues by optimising the model through an energy minimisation procedure [135]; or molecular dynamics, thus ridding the model of steric clashes. Care needs to be taken not to distort the model during the minimisation procedure [132].

While building a protein homology model is complicated, several programs and servers are available that can build a complete model from query sequences relatively easily, using different methods. MODELLER [136], ROBETTA [137], SwissModel [138] and PRIMO [139] are examples of such, where MODELLER exists as a stand-alone program, ROBETTA and SwissModel as web-servers, and PRIMO as an interactive pipeline. Additionally, specialised servers for specific steps of the process exist [127], such as ModLoop [140] for loop construction, and SCWRL [141] for side-chain prediction and modelling.

3.1.1.1 MODELLER

Using a template and target sequence alignment as input, MODELLER uses spatial restraints [142] to calculate and model a structure containing all non-hydrogen atoms. Restraints included are from the template structure and a representative set of known structures. Energy terms from the CHARMM22 [143] force-field are used to ensure the proper stereochemistry is combined with the spatial restraints. The output is assessed using a Discrete Optimized Protein Energy (DOPE) [144] score that analyses energy of the model, and a GA341 [145] composite fold assessment score. A normalized DOPE (z-DOPE) score can be derived from the statistics of raw DOPE scores, where a positive score is likely to be a poor model, and any lower than -1 is likely to be more native-like.

3.1.2 Model Validation

Each stage of homology modelling is reliant on the previous, meaning any errors introduced propagate downstream into the final model. Model validation and assessment is thus necessary to interpret and pick out any errors. Validation programs can be roughly into two categories, one that checks for proper protein stereochemistry and another that checks the fitness of sequence to structure and scores the fitting of

each residue to its current environment [127]. Some examples of validation programs which assess different aspects of the model include ProSA-web [146], Verify3D [147], PROCHECK [148], QMEAN [149] and WHAT-IF [150].

3.1.2.1 ProSA-web

ProSA uses C-α atoms to calculate model quality by comparing it to empirical energy potentials observed from a database of all known protein structures. ProSA output is a z-score of the overall model. The z-score measures the deviation of the total energy of the structure with respect to an energy distribution derived from random conformations [151]. A z-score outside the range characteristic for native proteins of similar size indicates an erroneous structure. A plot highlighting problematic regions on the model is also provided, thus showing local model quality.

3.1.2.2 Verify3D

The program assesses compatibility of a model to the 3D profile predicted for its sequence. The assessment is based on the location and environment of each residue position and compares these to databases of known high quality structures. The DSSP [152], which is a database of secondary structure assignments for all protein entries in the Protein Data Bank (PDB), is used for this.

Verify3D then scores each residue on a scale of -1 to +1, where a score greater than 0.2 suggests that the residue is in a structurally favourable environment. Regions scoring below 0.2 are assumed to be less reliable.

3.1.3 COX-1 3D structure

To date, no experimental atomic resolution 3D structure is available for the human COX-1 enzyme. The enzyme therefore needs to be modelled for *in silico* work. Fortunately, the enzyme shares 60% identity with COX-2 which has been widely studied and crystallised, as mentioned in Chapter 1. Additionally, COX sequence is highly conserved among mammals [29], with important domains, such as the heme-binding motif, being conserved in other peroxidases beyond the kingdom [27]; this is shown in the MSA in Figure3.2.

3.2 Methodology

3.2.1 Template Selection

The human COX-1 amino acid sequence (accession number P23219) was obtained from UniProt [153] and was submitted to NCBI BLAST [154], HHPred [155] and PRIMO to attain a suitable templates for modelling.

HHPred detects and predicts homology among distantly related proteins by implementing pairwise comparison of profile HMMs. A wide number of databases, such as the PDB, SCOP [156], Pfam [157] and CDD [158] are available for the search process. The PDB was searched for template identification.

Alternatively, PRIMO avails both protein BLAST (BLASTp) or HHsearch .BLAST is set as the default search option as it runs faster than HHsearch, identifying any closely related templates in the PDB.

3.2.2 Homology Modelling

3.2.2.1 Wild-type

MUSCLE and MAFFT were used to generate the initial template-target alignment and then further generate an MSA using default parameters. Sequences used to generate the MSA are listed in Table 3.1. Both MUSCLE and MAFFT use progressive-iterative alignment, with one of main the differences being the use of Fast Fourier Transform (FFT) in sequence alignment in MAFFT [159] [160]. The resultant MSAs were compared

A suitable template was identified, based on template coverage, identity and overall template quality. This template was used to create an alignment (PIR) file with the target sequence, which included the heme co-factors. This alignment file was used as input for MODELLER.

Model construction was conducted using MODELLER (v.9.16), run on a local cluster. Instructions on the modelling run, such as the PIR file, refinement settings, number of models to generated and the assessment method to be used were specified in an in-house Python script that called on the program.

Using a refinement setting of "slow", a 100 models were generated and assessed using the z-DOPE score. The scores were used to rank and select the top three models for further validation.

3.2.2.2 Variants

The SNPs were introduced into the target sequence for each of the variants. The resulting mutated sequences were used for model generation in MODELLER as described in section 3.2

3.2.3 Model Validation

The high-ranking models were validated by submission of to the ProSA [146] and Verify3D [147] web-servers. Default parameters were used in all instances.

Table 3.1: Protein sequences used to generate the MSA. COX-1 (PGHS1) and COX-2 (PGHS2) sequences of various organisms were used Sequences in the table in the order they appear in the MSA in Figure 3.2.

NCBI Accession ID	Organism	Protein
P23219	*Homo sapiens*	PGHS1
NP_035328.2	*Mus musculus*	PGHS2
NP_000954.1	*Homo sapiens*	PGHS2
NP_001009476.1	*Ovis aries*	PGHS1
XP_006939501.1	*Felis catus*	PGHS1
XP_026369696.1	*Ursus arctos horribilis*	PGHS1
XP_020028681.1	*Castor canadensis*	PGHS1
XP_425326.4	*Gallus gallus*	PGHS1
XP_00601955.1	*Alligator sinenis*	PGHS1
XP_007423361.1	*Python bivittatus*	PGHS1
XP_022620892.1	*Seriola dumerili*	PGHS1
XP_008334142.1	*Cynoglossus semilaevis*	PGHS1
XP_027326714.1	*Anas platyrhynchos*	PGHS1
XP_022328617.1	*Crassostrea virginica*	PGHS2
XP_014768109.1	*Octopus bimaculoides*	PGHS2
WP_113066573.1	*Nitrosospira multiformis*	PGHS1
WP_090657908.1	*Nitrosomonas marina*	PGHS1
XP_014289700.1	*Halyomorpha halys*	PGHS1

3.3 Results and Discussion

In this study, homology models of human COX-1 wild-type, as well as five of its variants were built as no experimentally resolved structure currently exists

The models were built and scored using MODELLER, and further validated using external servers.

3.3.1 Template selection and sequence alignment

Recurring hits in template selection were, ovine COX-1 (PDB ID: 1Q4G) and murine COX-2 (ID: 3NTI), which are shown in Table 3.1. Murine COX-2 was chosen as a suitable template, due to better overall metrics as seen in Figure 3.2 and a structure resolution of 1.73Å, versus the ovine 2Å.

Table 3.2: Qualities of potential templates identified from the PDB using HHpred and NCBI BLASTp.

Template PDB ID	Uniprot Acession ID	Organism name	Identity	Coverage	Resolution (Å)
1Q4G	P05979	*Ovis aries*	91%	92%	2.00
3NT1	Q05769	*Mus musculus*	63%	92%	1.73

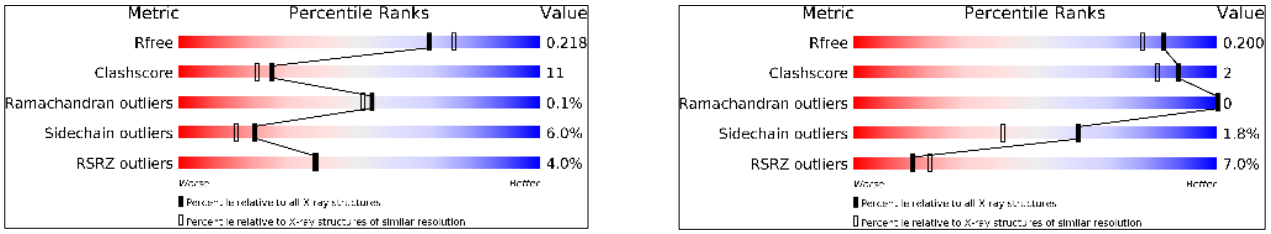


Figure 3.2: Structure metrics summary of ovine COX-1, PDB ID:1QAG (left) and murine COX-2, PDB ID:3NT1(right) as found in RCSB PBD.

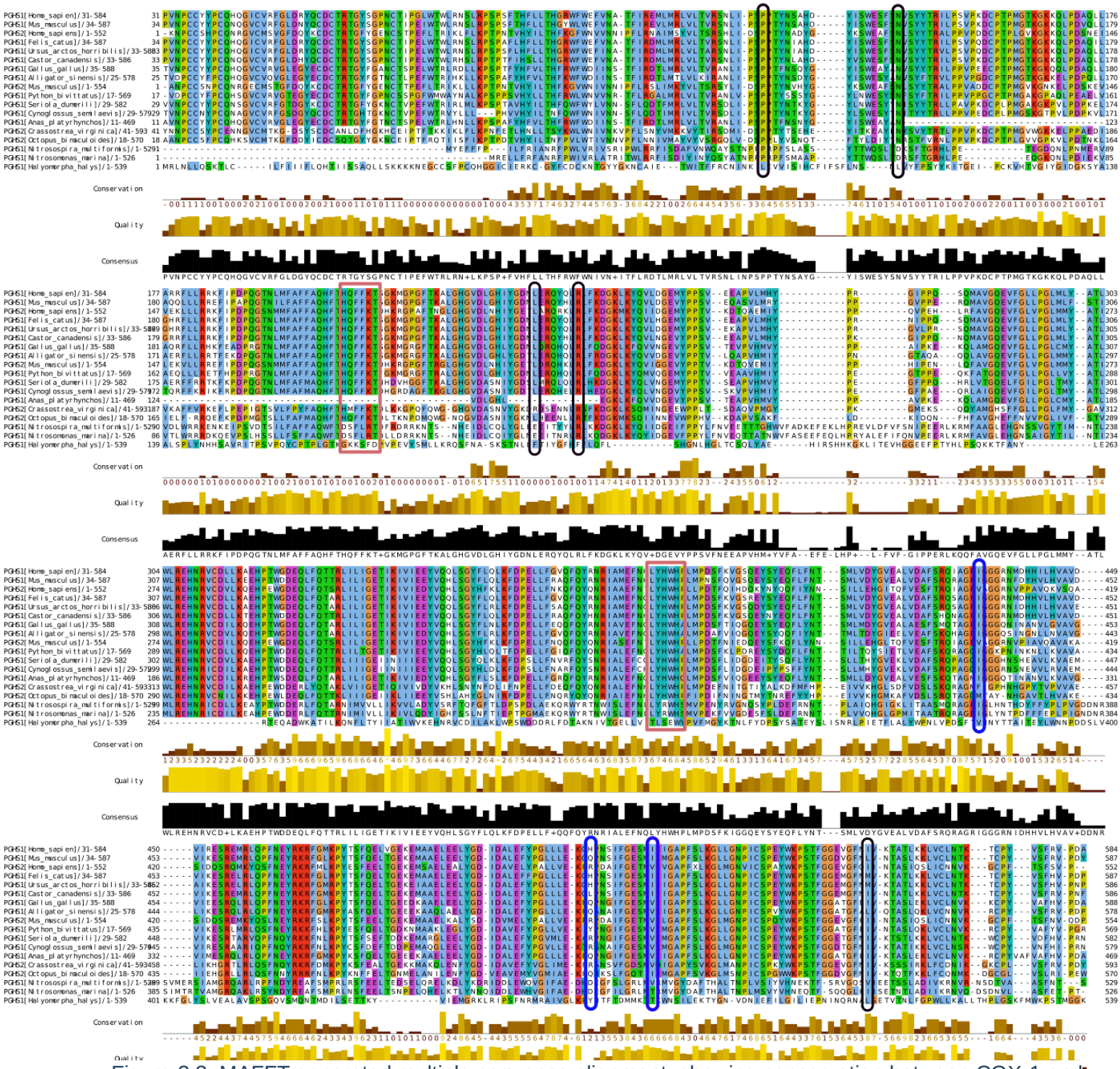


Figure 3.3: MAFFT generated multiple sequence alignment, showing conservation between COX-1 and COX-2 with species and across taxonomic classes. Highlighted are the heme-binding motif (encased in red), that is conserved among heme peroxidases such as the COXs and myeloperoxidase; sites of active site residue differences between COX-1 and COX-2 are (in blue), and sites of SNPs focused on in this study (in black)

3.3.2 Model generation and validation

While the structure-based residue numbering scheme generally applied to both COX isoforms in literature corresponds with ovine COX-1, from which the first COX 3D structure was resolved [27]; the residue numbering of the models generated in this study is based on the human COX-1 amino acid sequence. As a result, some amino acid references from literature will be one position off.

MODELLER in-program scoring of COX-1 models for the wild-type and variants was a z-DOPE score ≤ -1.00, signifying the models were good. Additional model validation corroborated these scores as shown in Table: 3.2. The ProSA generated Z-score for the template 3NT1 was -8.97, which is comparable to the Z-scores attained for the models. Overall model quality for the wild- type matched X-ray solved proteins structures of similar size as seen in Figure 3.3, with a score of -8.73, indicating similarity of the model to native structures. The local model quality plot shows per residue energies, where negative values indicate a good modelling and positive values problematic regions.

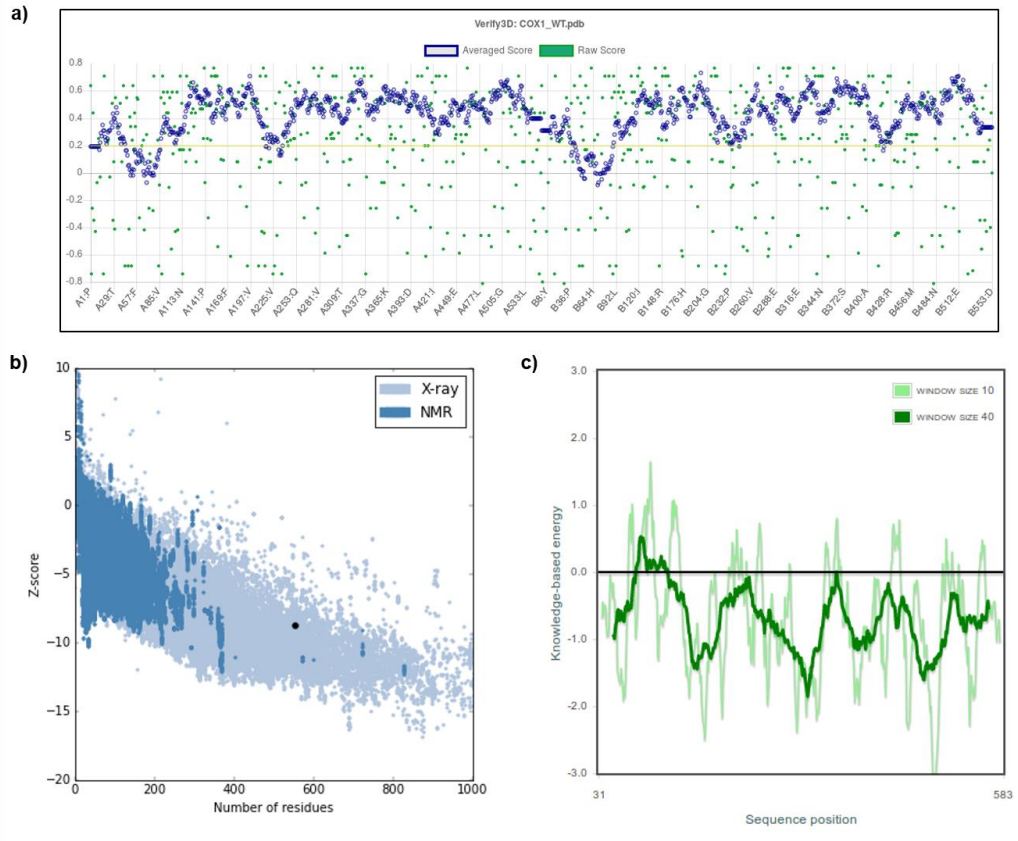


Figure 3.4: COX-1 wild-type homology model validation results, from Verify3D (a), and ProSA-web overall (b) and local (c) model quality

Based on the window size 40 assessment, most of the model residues had negative scores, further agreeing with the overall assessment.

ProSA pegged problematic areas and those scoring below 0.2 from Verify3D [S. Figure 2], coincided with loop regions in the structure, which are notorious for being difficult to crystallise. Fortunately, these regions did not cover SNPs to be investigated or catalytically important regions. COX-1 variant models followed a similar trend [Table: 3.2 and S. Figures 3-7], as such the models were deemed suitable to proceed with the study.

Table 3.3: Evaluation scores of homology models generated through MODELLER and further validated by Verify3D and ProSA programs. The protein sequence was mutated manually at all the respective SNP position.

Model	z-Dope score	ProSA Z-score	Verify3D
COX-1 WT	-1.02	-8.73	Pass - 89.60% score >= 0.2
COX-1 P126T	-1.07	-8.68	Pass - 86.44% score >= 0.2
COX-1 N143K	-1.01	-8.75	Pass - 88.88% score >= 0.2
COX-1 L237M	-1.00	-8.67	Pass - 86.17% score >= 0.2
COX-1 R244W	-1.02	-8.70	Pass - 89.97% score >= 0.2
COX-1 I557T	-0.99	-8.59	Pass - 89.42% score >= 0.2

3.4 Conclusion

The aim of this chapter was to build homology models of COX-1 for both the drug discovery and SNP analysis parts of this study. The models generated were of a high quality, with low scoring regions on the structures lying in regions inconsequential to this study.

The COX-1 wild-type model was to be used for molecular docking, in the drug discovery part of the study in Chapter 4, while, both the COX-1 wild-type and variant models were to be used to analyse SNP impact on the protein. MD simulation was to be used to analyse the consequence of the SNPs on the protein structure and function, with the wild-type as a reference in Chapter 5. The homology models were, therefore, of a suitable quality for both these purposes.

It is worth noting that MD simulations have an added benefit of refining stability the models [161] through the structure minimisation performed during said simulations that further optimises residue geometries. As such the models were further refined in these downstream processes.

4 Molecular Docking

4.1 Introduction

High-throughput virtual screening (HTVS) has become an increasingly important tool for drug discovery [67]. Compared to traditional experimental screening, virtual screening has the advantage of low cost and effective screening [162].

HTVS can be classified into ligand- and structure-based methods. Molecular docking, which encompasses both these methods is the most common form of HTVS. The main aim of molecular docking is to model a ligand-receptor complex structure, showing interaction between the two at an atomic level. This allows characterisation of ligand behaviour in the binding site of target proteins as well as elucidation of fundamental biochemical processes [163]. The docking process involves the prediction of ligand conformation, position and orientation within the binding site and assessment of the binding affinity using a scoring function. Ideally, pose prediction algorithms should be able to reproduce the experimental binding pose and the scoring function should rank said pose highest among all generated conformations.

Knowledge of the binding site location, which can be obtained through literature of similar proteins or co-crystallized protein-ligand structures significantly improves the docking process. In cases, where the docking is focused on a known binding site, the process is called targeted docking; while docking without any assumption about the binding site is referred to as blind docking. Blind docking therefore has unbiased, and mimics experimental screening [164]. Co-crystallised structures can additionally be used to validate docking parameters through re-docking to recreate the co-crystallised pose.

Based on the "induced-fit" theory” [165] the ligand and receptor should be treated as flexible during docking due to conformational changes that occur [166]. However due to the computational expense that would entail, docking is popularly performed with a flexible ligand and a rigid receptor as a trade-off between accuracy and computational time [167].Docking programs such as AutoDock [168] and FlexX [169] have adopted this methodology. Different search algorithms are used to implement these methods, such as systematic, molecular dynamics and genetic [170].

Scoring functions aim to separate correct poses from the incorrect in reasonable computation time; as a result, they involve estimation rather than calculation of the binding affinity between the protein and ligand. Scoring functions can be divided into force-field (physics), empirical and knowledge-based scoring functions [171]. Extensions of force-field based scoring functions, such as in AutoDock, consider hydrogen bonds, solvations and entropy contributions.

4.1.1 Steps in molecular docking

The docking procedure essentially contains four steps, that is, the ligand and receptor preparations, the docking process and pose scoring.

Receptor preparation: The protein is usually pre-processed by adding the appropriate number of hydrogen atoms, particularly, the protons required to define ionisation and tautomeric states of the amino acids in the protein. Further processing includes removal or inclusion of water molecules, cofactors, metals, as well as consideration of missing residues or atoms, according to the parameters available. After the preparation of protein, the binding site should be assigned. Some receptors possess more than one active site, hence the one of interest should be specified.

Ligand preparation: Preparation of the ligand entails assigning the correct atom types based on the appropriate ionisation states, chiralities and tautomeric states of the ligand. Protonation states and tautomeric forms of a ligand are of importance as they influence ionic and hydrogen bonding abilities. Various parameters such as desired pH, pKa, structure optimisation and partial charge calculations can be set using semi-empirical quantum chemical methods; and structure minimisation using a molecular mechanics force-field. Chemical property filters such Lipinski rule of 5, pan assay interference (PAINS) assays [172] and blood-brain barrier (BBB) [173] [174] can additionally be used to assist in discerning amongst non-drug like and drug like ligand candidates. The amount of pre-processing required tends to depend on database ligands are retrieved from.

Docking: The docking methodology used mainly depends on the search algorithm in play and the subsequent scoring function. The scoring function is applied in accordance to the same force field atoms of the protein and the ligand have been set up in. In addition to the aforementioned binding energy scoring, docking results can be further assessed via visual analysis of the ligand-protein interactions using software such as LigPlot+ [175] or BIOIVIA Discovery Studio (DS) [176].

4.1.2 Ligand/compound databases

Existing compound libraries usually contain 2D structures generated from molecular formulas or SMILES strings. The ZINC Database [177] maintained at University of California San Francisco (UCSF) provides commercially available compounds for structure based virtual screening. The compounds are grouped in different property subsets that can be downloaded in various structure file formats for virtual screening. The PubChem database [178] maintained at NCBI provides information on biological activities of small molecules, storing the vital information from a multitude of depositors. Other important databases include the directory of useful decoys (DUD) database [179] derived from ZINC, and ChEMBL [180].

Several natural products databases also exist to assist with in silico drug discovery, comprising of compounds extracted from various plant and marine sources. Natural product scaffolds are crucial to drug discovery, with a numerous existing drug being derived from or inspired by them [181]. The Traditional Chinese Medicine (TCM) Database@Taiwan [182], ConMedNP [183] and the South African natural compound database (SANCDB) [184] are examples of such, all containing region-specific compounds.

4.1.3 COX-1 and aspirin

Aspirin (acetylsalicylic acid) is unique among NSAIDs due to its covalent modification of COX-1 and COX-2 via the acetylation of the hydroxyl group of Ser-529 [31]. The reaction irreversibly inhibits COX-1 activity and subsequently the production of prostaglandins While Ser-529 on its own is not a catalytic residue, its acetylation blocks the cyclooxygenase channel, thus preventing productive binding of AA within the channel.

Experiments on point-mutated COX-2 have revealed that additional active site residues are Arg-119, Tyr-347 and Tyr-384 [185]. Findings suggest Arg-119 ensures the proper orientation of aspirin, rather than to directly participate in the reaction; while Tyr-384 plays a crucial role in the cyclooxygenase reaction.

Due to its irreversible inhibition of COX-1, aspirin is used as a reference for docking in this study.

4.2 Methodology

4.2.1 Receptor and ligand preparation

Minimised structures from SANCDB and the ZINC database Drugs Now subset were used for docking. The SANCDB subset consisted of 623 ligands and the ZINC subset of 5105 ligands.

All molecular structures were optimised by using MGLTools Python scripts from AutoDockTools (v1.5.6) software to add charges, assign hydrogen atoms, and set up rotatable bonds. The RESP charges and parameters for the protein heme group were attained from Henriques, *et al* [186]. The docking grid dimensions and grid centre Cartesian coordinates to assign a binding cavity were calculated using DS, based on size of the largest ligand and the centre of the protein, respectively. Using these calculations, the grid box size was set to 101^3Å.

Blind docking was initially performed on one monomer (chain A) of the protein using aforementioned grid box and centre. A targeted docking was then conducted on the remaining monomer, using only ligands that successfully docked into the chain A active site, a smaller grid box of 25^3Å and a grid centre based on Ser-529 Cα.

4.2.2 Docking

The molecular docking was conducted using Autodock Vina (v1.1.2) [187] with GNU parallel, on a total of 240 CPU cores. Vina configuration files were prepared to specify the ligand and receptor; and the binding site coordinates and box size. Additional parameters set in the all configuration files were an exhaustiveness of 120, the CPU usage of four cores per job and an energy range of 4.

4.2.2.1 Validation docking

Vina was employed for validation of docking parameters, by re-docking salicylic acid into a previously separated aspirin-huCOX-2 co-crystallised structure (PDB ID: 5F1A). Targeted socking was used for this validation.

4.2.3 Results screening

Docking results were screened based on proximity to active site, Vina-based energy scoring, analysis of protein-ligand interactions and an MD simulation with subsequent re-scoring. Results from docking in chain A were used in analysis and paired with whatever the corresponding chain B results were.

Proximity to the active site was calculated using Euclidean distance from the Cα of an important residue, by applying the equation shown below. The cut-off distance was set at 10Å. Binding energy was used to screened results with a cut-off of ≤ –

8kcal/mol the peroxidase active site; and ≤ -7.5 and ≤ -9 kcal/mol for the SANCDB and ZINC 15 subsets respectively, in the cyclooxygenase active site

$$d\ (P_1, P_2) = \sqrt{(x_2 - x_1) + (y_2 - y_1) + (x_2 - x_1)}$$

Ligands that met both the criteria were then visually analysed to assess protein-ligand interactions. The interactions were visualised using DS (v.14.1), LigPlot+ (v.1.4.5) and PyMOL [188] (v.1.7) to determine bonds and interactions formed, including hydrogen, hydrophobic and van der Waals (VDW). PDB 5F1A was used as a reference.

Additionally, molecular dynamics was used to assess stability of the remaining ligands. MD simulations of 10ns were run on favourable candidates and snapshots of the complexes captured at the end were re-scored using a Vina re-score function. These results will be discussed in Chapter 5.

4.3 Results and Discussion

In this chapter, select ligands from SANCDB and ZINC databases were docked against the COX-1 wild-type to identify potential inhibitors. As COX-1 is a dimer, blind docking was conducted on chain A, followed by targeted docking of the high-ranking ligands in chain B. A validation docking was conducted to validate docking parameters prior to docking of the ligands.

The resulting posed ligands were initially screened based on proximity to a crucial active site residue and docking binding energies, which are shown in Table 4.1. Additional screening was based on ligand interactions with active site residues.

4.3.1 Validation docking

While not resulting in a hydrogen bond Ser-529, docking parameters used were able to reproduce the pose from the co-crystallised structure, 5F1A, as shown in Figure 4.1.As such the parameters were deemed suitable to use for the rest of the docking study.

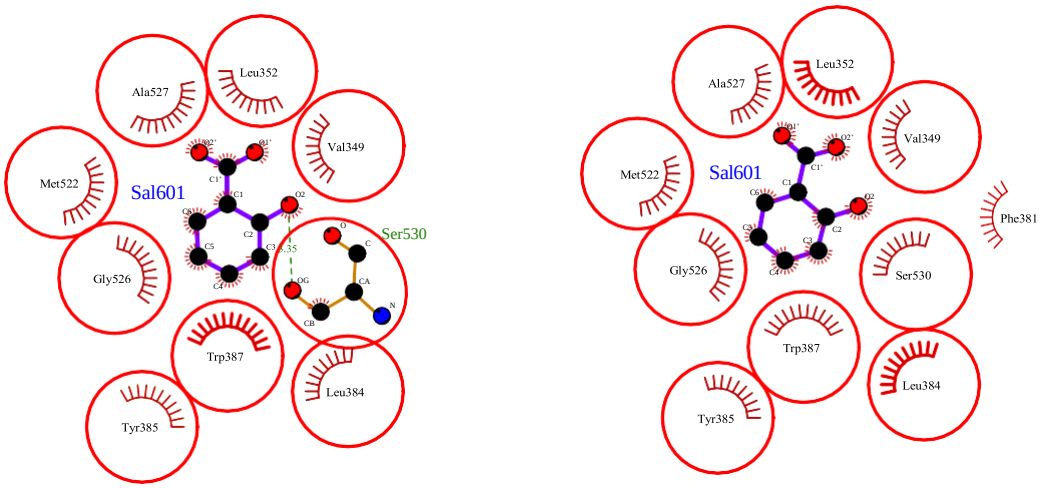


Figure 4.1: :Validation docking, showing re-docking co-crystallised salicylic acid into COX-2, using 2D Ligplot+ interaction diagram.

4.3.2 SANCDB and ZINC 15 subset docking

Euclidean distance was used to calculate distance of docked ligands in the protein. Additional screening for any ligand docking in the peroxidase active site using His-206.

From the initial screening, based on distance and binding energy, the cyclooxygenase site had 31 hits from the SANCDB subset as shown in Figure 4.2 and 123 hits from the ZINC15 subset shown in Figure.4.3b. The peroxidase active site returned no hits from the SANCDB and 17 from ZINC15 [Figure 4.3a].

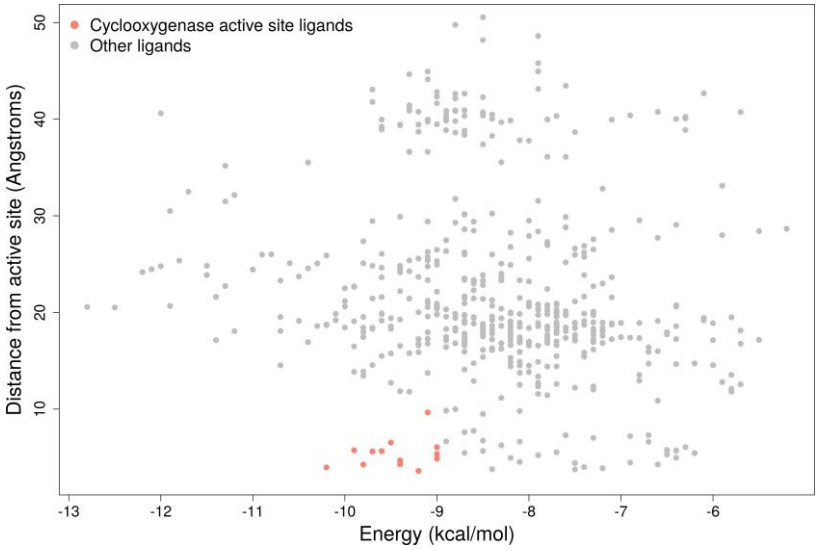


Figure 4.2: Scatter plot showing distribution of SANCDB ligands in relation to COX active site post-docking. Ligands meeting desirable criteria are highlighted in pink.

The ZINC subset due to its sheer size compared to the SANCDB naturally returned more hits. The Drugs Now subset consists of synthetic ligands that tend to be of a smaller molecular weight compared to the natural product ligands [189] from SANCDB. Based on the relatively small molecular weight (180.159g/mol) and complexity of aspirin, it is probable that ligands of "aspirin-like" properties are likely to interact with COX-1 as it does [190].

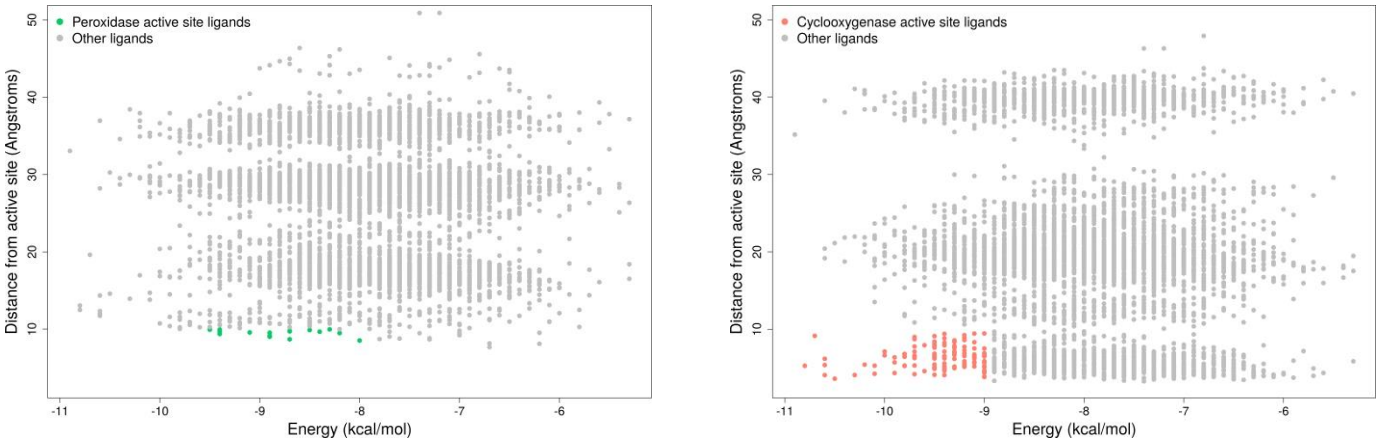


Figure 4.3: Scatter plot showing distribution of ZINC15 ligands in relation to peroxidase (left) and cyclooxygenase (right) active sites post-docking. Ligands meeting desirable criteria are highlighted in green and pink, respectively.

Inter-molecular interactions were used for further screening of docking hits. Hydrogen bonds are considered to be the most important of electrostatic interactions, conferring stability to ligand binding [191]. As such ligands exhibiting VDW interactions and at least two hydrogen interactions with relevant active site residues were selected. Good hits in chain A monomer of the protein were not necessarily mirrored in chain B, as reflected in Table 4.1.

Table 4.1: Ligand post-docking binding energies.

Active Site	Ligand		Docking Binding Energy (kcal/mol)	
	Database ID	Codename	Chain A	Chain B
Cyclooxygenase	SANC00239	SANC239	-7.7	-6.4
	SANC00521	SANC521	-8.7	-7.8
	SANC00627	SANC627	-8.1	-6.0
	SANC00721	SANC721	-9.4	-4.3
	ZINC04649897	ZINC925	-10.5	-6.4
	ZINC00624371	ZINC3113	-10.3	-7.5
	ZINC04525816	ZINC4587	-9.0	-8.2
	ZINC01018315	ZINC4671	-10.7	-6.6
	ZINC04523993	ZINC4749	-10.6	-8.8
Peroxidase	ZINC00674446	ZINC4394	-9.4	0.6
	ZINC02134322	ZINC4591	-9.5	-7.9

4.3.2.1 ZINC 15 ligands

Of the 13 ligands that docked favourably into COX-1 chain A, 7 were retrieved from ZINC15. The structures of the ZINC15 ligands are shown in Figure 4.4

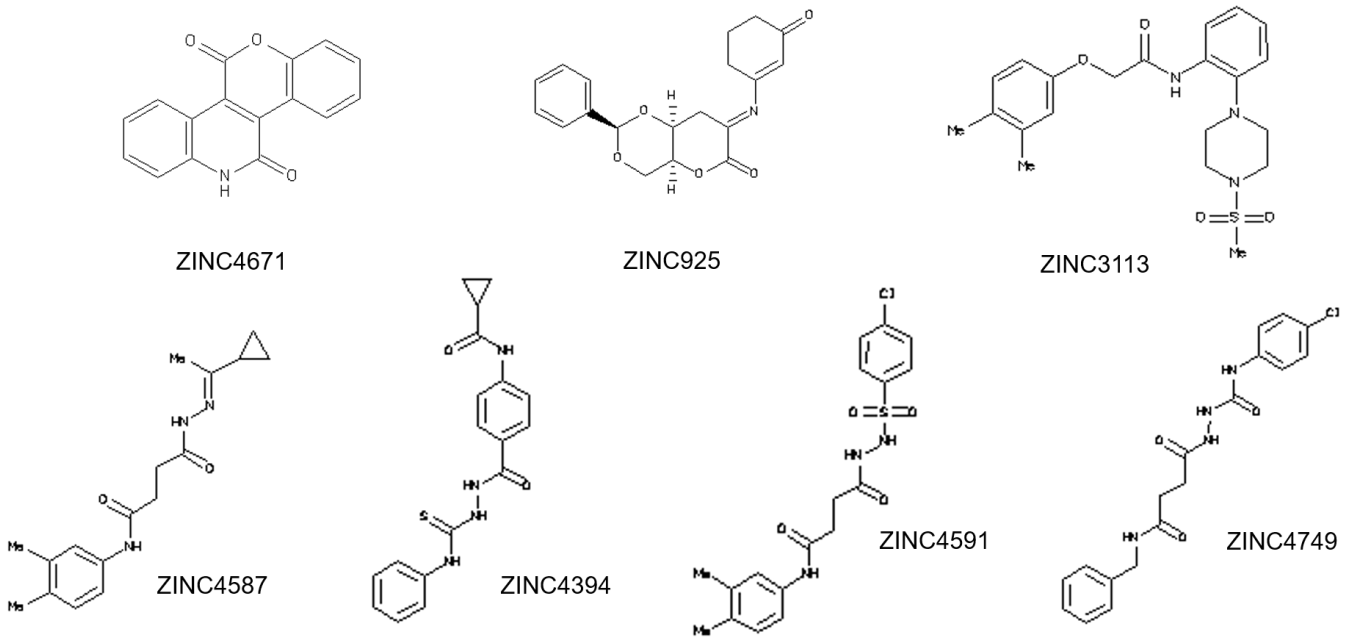


Figure 4.4: Structures of the ZINC15 subset synthetic ligands that docked favourably into COX-1.

The chemical and physical properties of the ligands were extracted from the ZINC15 database. These are summarised in the table below. Unfortunately, records for some of the ligands were absent at the time of retrieval, as such, these were not included in the table.

Table 4.2: Physical and chemical properties of the documented ZINC15 ligands at pH 7.

Ligand	Popular name	H-bond donors	H-bond acceptors	Molecular weight (g/mol)	Rotatable bonds	Bioactivity assays
ZINC925	4-[(3-oxo-1-cyclohexenyl)amino]-8-phenyl-2,7,9-trioxabicyclo[4.4.0]dec-4-en-3-one	0	6	341.363	2	-
ZINC3113	2-(3,4-dimethylphenoxy)-N-{2-[4-(methylsulfonyl)-1-piperazinyl]phenyl}acetamide	1	7	417.531	6	2
ZINC4671	12H-chromeno[4,3-c]quinoline-5,11-quinone	1	4	263.252	0	1*

While no molecule records were listed for ligands ZINC4394, ZINC4587, ZINC4591 and ZINC4749, the ligands were retained as they presented as potential drug compounds based on their binding energies with COX-1, shown in Table 4.1, and passing the other screening methods mentioned in Section 4.2.3. Ligands ZINC925, ZINC3113 and ZINC 4671 all met the three of Lipinski's rules which define necessary parameter ranges for viable oral drug compounds, which are, MWT ≤ 500, Log P ≤ 5, H-bond donors ≤ 5, and H-bond acceptor ≤ 10 [192].

Additionally, ZINC3113 and ZINC4671 were listed in bioactivity assays. ZINC3113 bioactivity was assayed and deemed inactive in two Alpha Screen-based biochemical high throughput studies, one as an activator of E3 ligase (PubChem AID 1259310) and the other as an inhibitor microphthalmia-associated transcription factor (MITF) (PubChem AID 1259310). Ligand ZINC4671 was synthesised for a study on *Plasmodium falciparum* cyclin-dependent protein kinase (Pfmrk), but its IC_{50} value could not be determined due to solubility problems [193].

4.3.2.2 SANCDB ligands

After screening the SANCDB ligands docked into COX-1, based on binding energies, proximity to active site and intermolecular interactions with active site residues, four ligands were selected for further analysis. The structures of the four ligands, SANC239, SANC521, SANC627 and SANC721 are shown in Figure 4.5, while the physical and chemical properties are listed in Table 4.3.

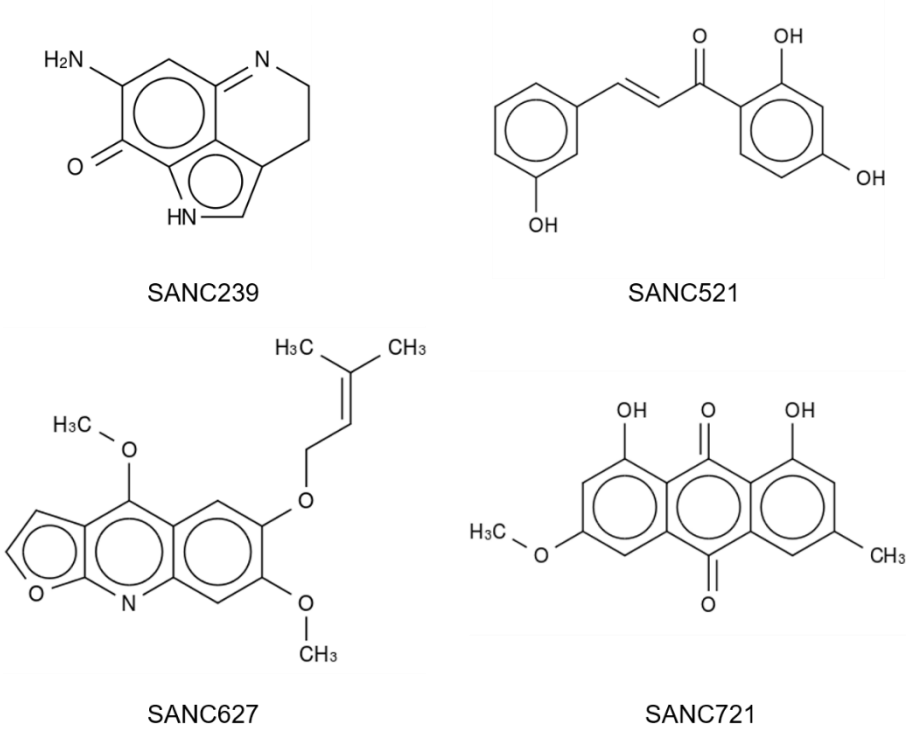


Figure 4.5: Structures of the SANCDB subset of natural product ligands that docked favourably into COX-1.

As shown in Table 4.3, all the four ligands passed the three aforementioned Lipinski rules, suggesting they too were viable drug compounds. SANC239, SANC521 and SANC721 have already been investigated for bioactivity in other studies. SANC239 has been cited to have anticancer properties in a study by Whibley and colleagues [194]; while SANC721 is recorded to have antimicrobial and antifungal properties [195]. SANC521 is listed to show moderate anti-tuberculosis properties in an anti-mycobacterial study [196].

Table 4.3 Physical and chemical properties of SANCDB ligands selected after screening process

Ligand	SANCDB entry name	Classification	Source organism	Molecular weight (g/mol)	H-bond donors	H-bond acceptors	Compound activity
SANC239	Makaluvamine	Alkaloid	*Strongylodesma aliwaliensis*	187.198	3	4	Anticancer activity
SANC521	(E)-3,2',4'-Trihydroxychalcone	Flavonoid	*Galenia Africana*	256.253	3	4	-
SANC627	Tecleanatalensine B	Furoquinoline alkaloid	*Teclea natalensis*.	313.348	0	5	-
SANC721	Physcion	Aromatic polyketide	*Eurotium rubrum*.	284.263	2	5	Antibacterial Antifungal Sun-screen agent

4.4 Conclusion

The aim of this chapter was to use molecular docking to screen ligands from varying databases (SANCDB and ZINC15) against the COX-1 wild-type in order to identify potential drug compounds.

After the molecular docking and results filtering, 11 ligands were identified that docked favourably into the cyclooxygenase as well as the peroxidase active site of the protein. Of these selected ligands, seven exhibited some molecular properties desirable of potential drug compounds, based on Lipinski's rule of five. This suggested said ligands could be putative lead compounds.

This final set of eleven ligands was set aside for further analysis in Chapter 5 using MD. Due to the rigid nature of the receptor in molecular docking, pose predictions are not sufficiently accurate. MD simulations provide a more apt discrimination between stable and unstable docked poses [197] [198] as both the receptor and ligand are flexible.

5 Molecular Dynamics

5.1 Introduction

Molecular dynamics (MD) is the computational simulation of the natural motions of atoms and molecules in a molecular system in a time-dependent manner [199]. MD can be used to supplement and complement conventional wet-lab experimentation. Due to the ability to manipulate numerous aspects of a simulation, it is possible to capture and analyse data impossible to do so in wet-lab. Additionally, computer simulated experimentation is faster and less expensive than the alternative.

Some biological applications of MD simulations include, molecular docking and drug design [200], through the analysis of protein-drug interactions; refinement of structure prediction [201]; protein motion and functional conformation analysis [202] [114], and protein folding analysis [203].

Several simulation engines exist to conduct MD, the most popular being, AMBER [204], CHARMM [205], GROMACS [206], and NAMD [207]. In these programs, MD simulations are conducted through determination of molecular and atomic trajectories by solving Newton's classical laws of motion for a system of interacting particles, via integration. Inter-particle forces and potential energy for these calculations are defined by molecular mechanics force-fields [208].

5.1.1 MD force-fields

Force-fields are essentially complex equations, where potential energy is deduced from the molecular structure. Force-field representation of molecular features include springs for bond length and angles, periodic functions for bond rotations and Lennard-Jones potentials, and the Coulomb's law for VDW and electrostatic interactions, which assure that energy and force calculations are fast.

Force-fields additionally exist in united-atom and all-atom models. In united atom force-fields several atoms, usually methyl groups, are joined into a single interaction site, whereas all atom force-fields explicitly represent aliphatic hydrogen atoms. While the molecular representation is the same, various force fields currently used in MD simulations differ in how they are parameterised. The determination of force field parameters requires significant empirical and quantum mechanical calculations. Though force-fields tend to have unique parameters and molecule types, but the resulting simulations are normally comparable [209]. Commonly used families of force fields are AMBER [210], CHARMM [211], GROMOS [212] and OLPS [213].

5.1.2 Running an MD simulation

The basic properties needed to run an MD simulation are topological, structural, energetic and thermodynamic properties. Each of these describes a different aspect of the system, the topological describes connectivity of atoms; the structural provides atomic position and conformation; the energetic describes forces acting on the system and the thermodynamic assigns experimental conditions. Application of these properties takes place over four general steps:

Setting up the system An initial model of the system can be obtained from either experimental structures or homology modelling data. The system can be represented by an atomistic representation for reproduction of the actual systems or coarse-grained representations for large systems. The topology is derived from this structure.

The standard procedure to set up a system once involves defining the size of the entire system; fixing structural errors; ionisation of titratable amino acids; addition of structural water molecules, counter-ions, and solvent; and energy minimisation.

A simulation box defining boundaries and size is set up to contain the system. To combat surface interactions with the boundary and approximate an infinite system, periodic boundary conditions (PBC) are most favourably used. Solvation of the system is crucial as many biological processes occur in aqueous solution; determining molecular conformation, electrostatic properties and binding energies [214]. Solvent representation is important to this end, with the most effective being the explicit molecule representation [214] [215] [216], which unavoidably increases the size of the systems [217]. Its advantage lies in the ability of the solvent molecules to maintain most of the solvation effects of real solvent including entropic ones like, hydrophobic effects.

The system then needs to be neutralised with the intent to avoid polarisation or set to desired pH; at a defined salt concentration. Energy minimisation, though it does not include the temperature, is used to compute an equilibrium configuration of the molecules in the system, to ensure there are no steric clashes. Starting from a non-equilibrium geometry the mathematical procedure reorients and moves atoms to find the lowest energy configuration and ideally the global minima on the potential energy surface.

Due to the vast number of possible atom types, not all ligands and co- factors, are not catered for in existing force-fields. As a result, ligand topologies and parameters are derived externally based on the force-field of choice, using quantum mechanical (for OPLS, AMBER and CHARMM) and semi-empirical (for GROMOS) calculations. While parameters can be fitted, automated tools such as ACPYPE [218], PRODRG [219], MKTOP [220] and Automated Topology Builder (ATB) [221] tend be used to generate the necessary force-field compatible parameters.

Equilibration To avoid a collapse of the system post minimisation, equilibration, of the solvent and ions around the protein is required at a desired temperature, pressure and density [222].

Integration of Newton's equations of motion allows a constant energy surface of the system, but not the temperature and pressure. Thermodynamic ensembles are therefore used for equilibration of the system, where the correct ensemble distribution

for specified temperature and pressure allows interpretation of the trajectory in a conventional way. Common ensembles are the constant-energy, constant-volume ensemble (NVE); the constant-temperature, constant-volume ensemble (NVT) and the constant-temperature, constant-pressure ensemble (NPT). NVE, also known as the microcanonical ensemble, obtained by solving Newton's equation without any temperature and pressure control conserves energy in the system [223]. The NVT ensemble, also referred to as the canonical ensemble is obtained by controlling the temperature with the volume kept constant throughout the run. The amount of substance (N), volume (V) and temperature (T) are conserved; and the energy of any endothermic and exothermic processes is regulated by a thermostat. The temperature of the system should reach a plateau at the desired value signifying the temperature has stabilized. The NPT ensemble, called the isothermal-isobaric ensemble, allows control over both the temperature and pressure, where the pressure is adjusted by adjusting the volume. This ensemble tends to be used during equilibration to achieve the equilibrium density corresponding to the desired pressure and temperature. Amount of substance (N), pressure (P) and temperature (T) are conserved and a thermostat [224] [225] and a barostat [226] are needed. While choice in the thermostat and barostat is not critical, it is important that an equilibrated state is reached.

Production Once an equilibrium attained at the desired temperature and pressure, position restrains can be released and a production run can be conducted. Ideally, the simulation time should long be enough to allow the protein to explore all the possible configurations.

Assessment The simulated trajectory must be analysed for data collection and to extract the desired properties. Assessment of protein motions can be conducted at a global and local level, based on access atomic positions, velocities and even forces as a function of time.

5.1.3 Analyses of MD

5.1.3.1 Global motions

RMSD Root mean square deviation (RMSD) is a measure of average distance between the atoms of superimposed proteins or other macromolecules. In proteins, RMSD will be calculated from Cα atoms fitting to protein backbones. RMSD provides a quantitative measure of several experimental properties. The analysis can be used to measure of similarity protein structures during homology modelling [215], conformational change over the course of a MD simulation [227] or protein and ligand stability in molecular docking assessment.

Principal component analysis Principal Component Analysis (PCA) of MD simulations is a popular statistical technique used to analyse essential dynamics of a system by applying a decomposition process that filters observed motions from the greatest to the smallest spatial scales.

Two types of PCA exist for analysis in MD simulations, internal and Cartesian coordinate PCA; where Cartesian PCA shows dominant overall motion of the protein, using mass-weighted Cartesian coordinates. The variance in simulation data is analysed over the length of the simulation and presented as a plot. Each point on the PCA plot represents the conformation of the protein at a specific time frame during the simulation. Due to the ordering of observed motions decreasingly, a large part of a system's conformational fluctuations can be described by the first few principal components (PCs) [228], meaning these PCs are most relevant.

5.1.3.2 Energy analyses

Free binding energy, in its most basic form is the difference between the energy protein-ligand complex and the sum of the individual energies of the protein and ligand as shown in the equation below. Several methods exist to estimate free binding energy such as MM-PBSA and MM-GBSA. The MM stands for molecular mechanics, PB and GB for Poisson-Boltzmann and Generalized Born, respectively, and SA for solvent-accessible surface area.

$$\Delta G_{binding} = G_{complex} - (G_{protein} + G_{ligand})$$

MM-PBSA, which was developed by Kollman, *et* al [229], combines three energy terms to account for the change in the free binding energy. The first one accounts for potential energy change in a vacuum (E_{mm}), the other for the desolvation of the different species ($G_{solvation}$) and the last for the entropy associated with complex formation. E_{mm} includes bond, angle, and dihedral energies as well as VDW and electrostatic interactions. $G_{solvation}$ can be further broken down into G_{polar} and G_{apolar}.

5.1.3.3 Visual analyses

Snapshots of MD simulation at specific time frames can be visualised using various software such as PyMOL, LigPlot+, DS or Schrodinger Maestro.

These snapshots capture intermolecular interactions within the system. Additionally, all frames of a simulation can be visualised as video in software such as Visual Molecular Dynamics (VMD) [230] without the periodic boundary conditions box.

5.1.4 Limitations of MD

Despite its obvious convenience, MD simulation has several limitations. Some of these include an inability to model quantum effects, the quality of force- fields, and time and size limitations. While research continues to be conducted to improve these aspects [231] [232] [233] [234] the computational intensity of MD simulations has been alleviated through use of high-performance computing (HPC).

5.2 Methodology

5.2.1 Setting up MD run

5.2.1.1 Wild-type and variants

The energy minimisation, equilibration, and dynamic simulations were performed using GROMACS (v.2016.1-dev), using a modified version of the GROMOS 54a7 forced-field that includes ATB specific atom types. GROMOS 54a7 is a united atom force field [235] that includes parameters for standalone and Histidine coordinated heme groups.

The system was immersed in a cubic solvent box of dimensions $1nm^3$, using the SPC (simple point charge) water model with periodic boundary conditions applied. The systems were neutralized at a concentration of 0.15mM NaCl to mimic human physiology. To ensure conformational stability, the whole system was minimised using steepest descent algorithm, with a maximum of 50000 steps and an F_{max} not exceeding 1000.0 kJ/mol/nm. For equilibration, the system was brought to a temperature of 300 K and then relaxed to an appropriate density by bring it to a pressure of 1 bar using the NVT and NPT ensembles, respectively. The Particle mesh Ewald (PME) method was used in all simulations to treat electrostatic interactions and the LINCS method to constrain bond lengths. Production runs of 150ns were then performed on 24CPU cores of the Lengau cluster at the Centre for High Performance Computing (CHPC), in Cape Town, with a time step of 2 fs.

5.2.1.2 Protein-ligand complexes

Protein-ligand complex simulations performed were using the same parameters and resources as the in section 5.2.1.2, with additional steps to account for presence of the ligands docked into the protein.

Ligand input files compatible with the GROMOS 54a7 force field were generated using the ATB server (v.3). The existing protein MD input files were updated to include the ligand information generated, resulting in a protein-ligand system.

Post energy minimisation restraints were applied to the ligands; and temperature coupling of ligands to the protein; and ions to the solvent was conducted using the Berendsen thermostat.

The systems were then equilibrated, and production runs of 10ns, which were later extended to 100ns after pose re-scoring, were conducted for initial filtering of-poor docking poses.

5.2.2 Trajectory analysis

Conformational change and stability over the course of the trajectory was analysed using RMSD, RMSF and RG, generated using GROMACS and Cartesian coordinate PCA.

MD-TASK [236], was employed for dynamic residue network (DRN) analysis DRN, was used to calculate betweeness centrality (Δ BC), average shortest distance (Δ L), and additionally generate residue contact maps. The residue contact maps were used to monitor interactions between SNPs and residues interacting with them.

Free binding energy of the protein-ligand complexes based on a 10ns excerpt was calculated using the of GROMACS compatible g-mmpbsa suite (v.1.6), using the solvent accessible surface area (SASA) model [237].

Additionally, visual analysis was conducted using VMD (v.1.9.2)

5.3 Results and Discussion

In this chapter MD simulations were conducted for a) the apo/wild-type protein, b) the protein nsSNP variants and c) the protein ligand complexes. Analyses were conducted as outlined by Brown *et al., 2017* [84].

MD simulations of the protein variants included five cases, P126T, N143K, L237M, R244W and I557T. Protein-ligand complex MD simulations for all 11 protein-ligand complexes were performed though only those for ligands SANC239 and ZINC4671 were analysed. All were analysed for global motions (RMSD, RG and PCA) and local motions (RMSF, *Average BC*, *Average L* and Contact Maps). MM-GBSA analysis was additionally performed on the protein-ligand complexes to calculate free binding energy.

5.3.1 Apo/Wild-type Analysis

The COX-1 wild-type MD simulations were conducted in duplicate.

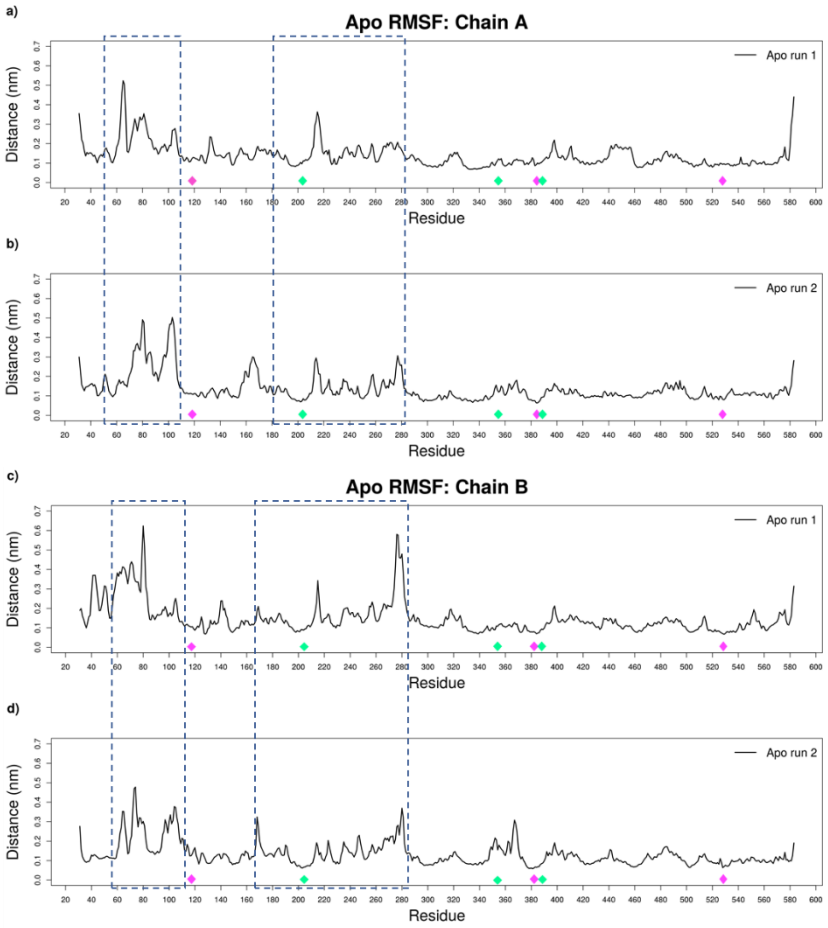


Figure 5.1: RMSF of COX-1 wild-type in duplicate, showing asymmetry between chain A (top) and chain B (bottom).Catalytic and important active residues for the cyclooxygenase (pink) and peroxidase (green) active sites are shown.

5.3.1.1 RMSF, RMSD and RG

RMSF analysis of the monomers in Figure 5.1 in both MD runs exhibited dissimilar behaviour. This suggests that COX-1 is indeed as conformational heterodimer, with one monomer conducting catalytic operations and the other being allosteric. The question of which monomer is catalytic cannot be answered using RMSF. None of the important active site residues [30] [238] [239] show any exceptional behaviour in both chains. Asymmetry in the monomers was further seen in the model B-factors, shown in Supplemental Figure (S.Figure) 1. Subtle differences a can be seen, with chain B-factor values.

While fluctuating regions differed between the chains, [Figure. 5.2] the helices of the membrane binding domain (between residues 60-100) exhibited more flexibility. This may due to these being the helices that anchors to and lie parallel to the membrane, as suggested by Fowler, *et al* [30]. It is possible that the helices that would normally be stabilised by the membrane, exhibits more free play due to lack of membrane in the MD simulation. Additionally, residues 60-70 make up the loop bridging EGF to first helix of MDB. Residues 260-300 of chain B constitute a loop, whose behaviour was not mirrored in chain A, further supporting the idea heterodimeric behaviour [].

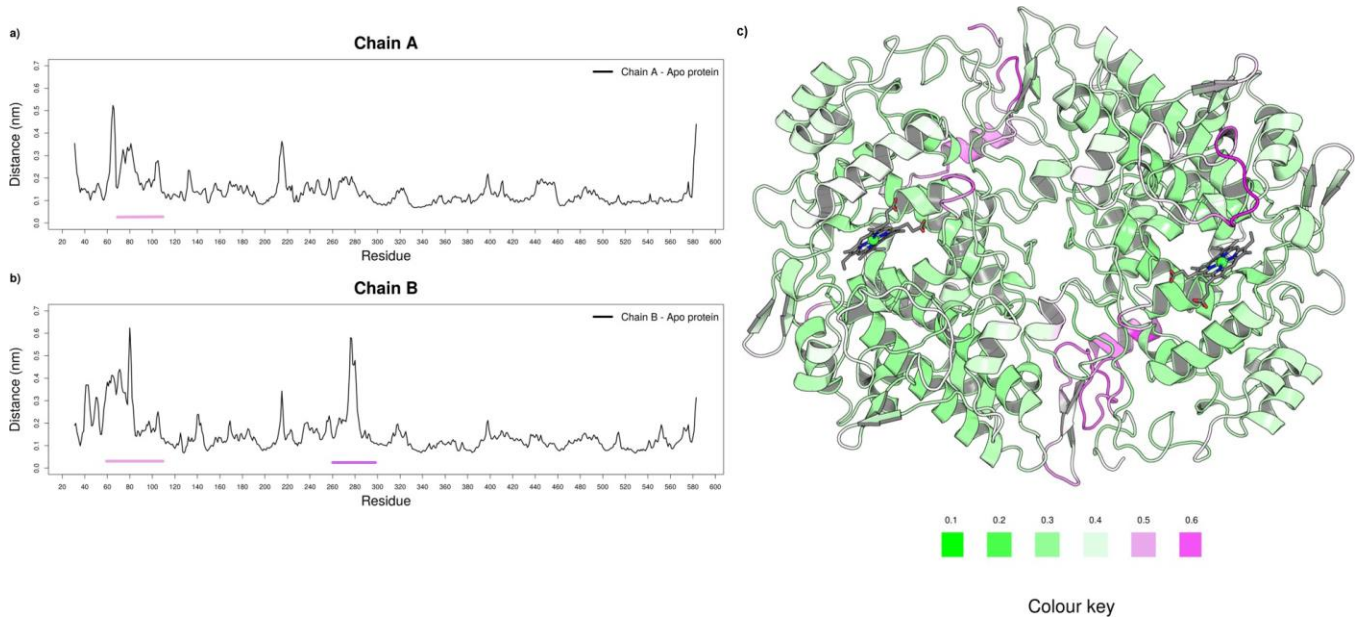


Figure 5.2: RMSF of COX-1 wild-type monomers, chain A (a) and chain B (b) during the MD simulation. Flexible regions are highlighted on the graphs and on the structure in (c) based on the colour key.

5.3.1.2 Principal Component Analysis

PCA of the wild-type showed conformational change over the length of the MD simulation. PC1 represented 52.6% of the variance, and as such, was the most significant, conveying global motion of the protein. As seen Figure 5.3 the distance between the point at frame 0 and the point at frame 150 is significantly large, along PC1.

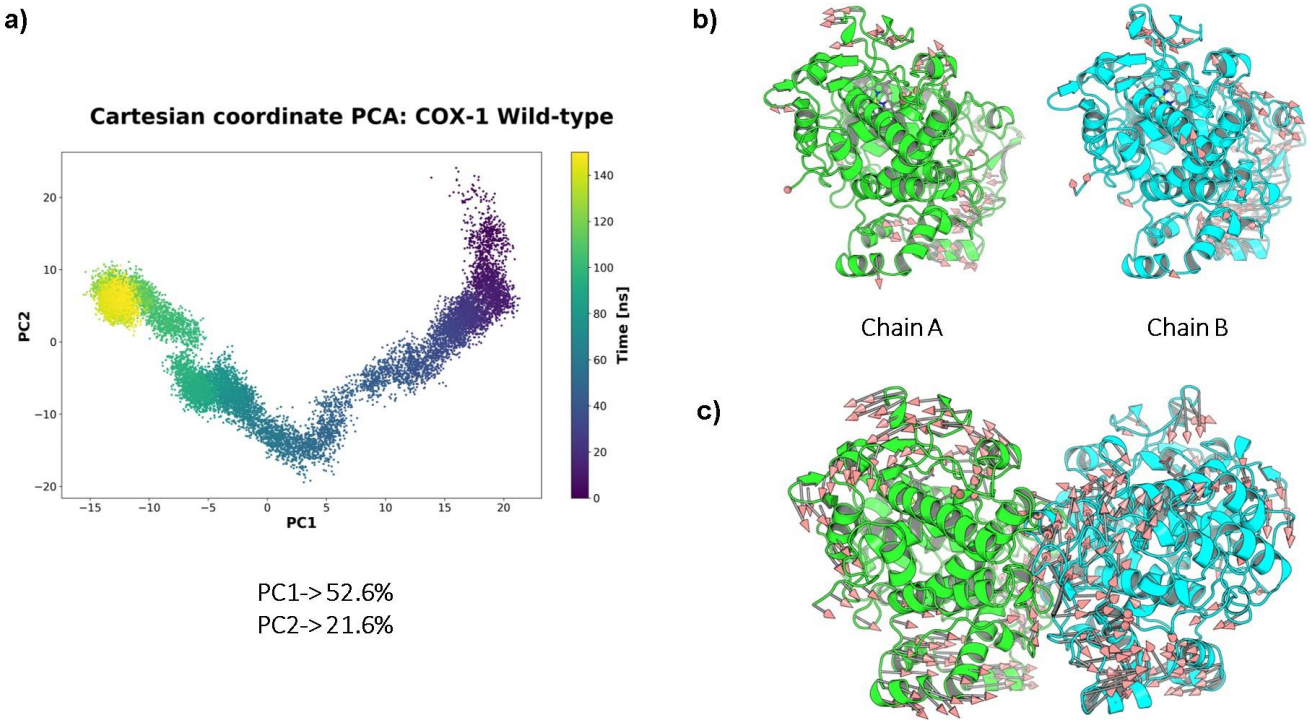


Figure 5.3: PCA plot of PC1 vs PC2 showing conformational change over the course of the simulation (a). Arrows in (b) and (c) show motions of the protein from the beginning of the simulation to the end.

Figures 5.3 (b) and (c) further supports the difference in chain movement, where loops in chain B (cyan) showed a more inward motion, contrasted by that of chain A. Movement around the heme cofactor also differed in both chains [Figure 5.4]. According to literature, the catalytic monomer has a higher binding affinity for heme than its allosteric partner [34] [38]. This difference in behaviour can be attributed to the different heme-monomer associations.

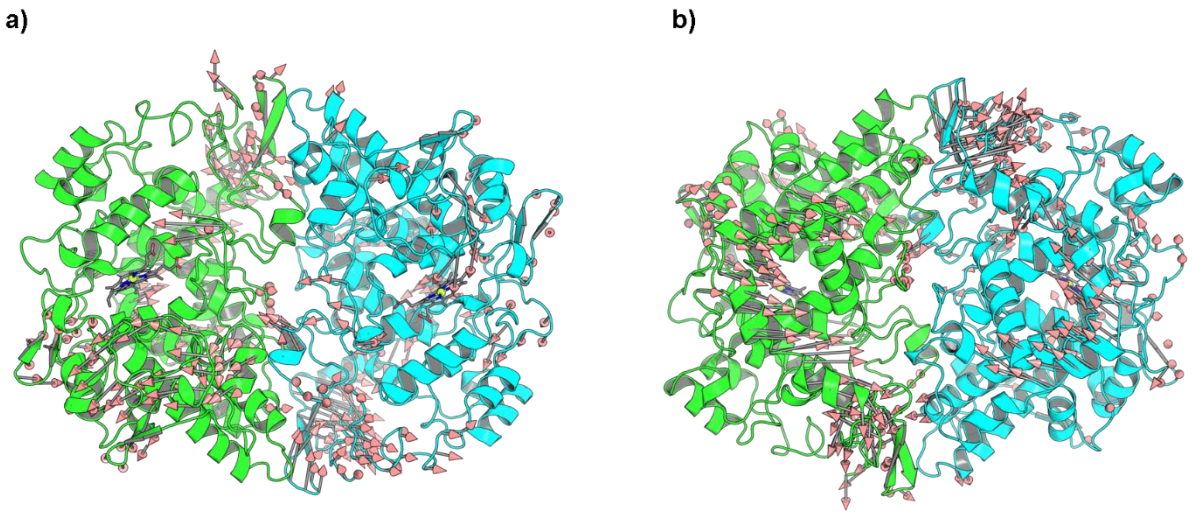


Figure 5.4:: Movement of COX-1 wild-type over the course of the MD simulation; (a) shows the top view of the protein, and (b) the bottom view.

5.3.1.3 Network Analysis

Average BC measures the frequency of contact with between residues, while *average L* measures the bond length of distance between bonds. BC tends to reflect the importance of a residue in a network, while L describes accessibility [84]. Typically, the two have an inversely proportional relationship, where residues with highest BC tend to have the lowest L.

Residues 372 and 371 respectively, had the highest average BC values, as seen in Fig.5.5. Although not catalytically significant, both residues were located in the core of the catalytic site, located in the middle of the protein. Active site residues such as Tyr-354 and Tyr-384 fall in this region. Other regions with higher *average BC*, 60-100, 130-150 and 210-230 fell in the dimer interface [29] [86] [240]. Notably residues from 123-129 and 505-515, have been reported to be important to dimer cross talk [41].

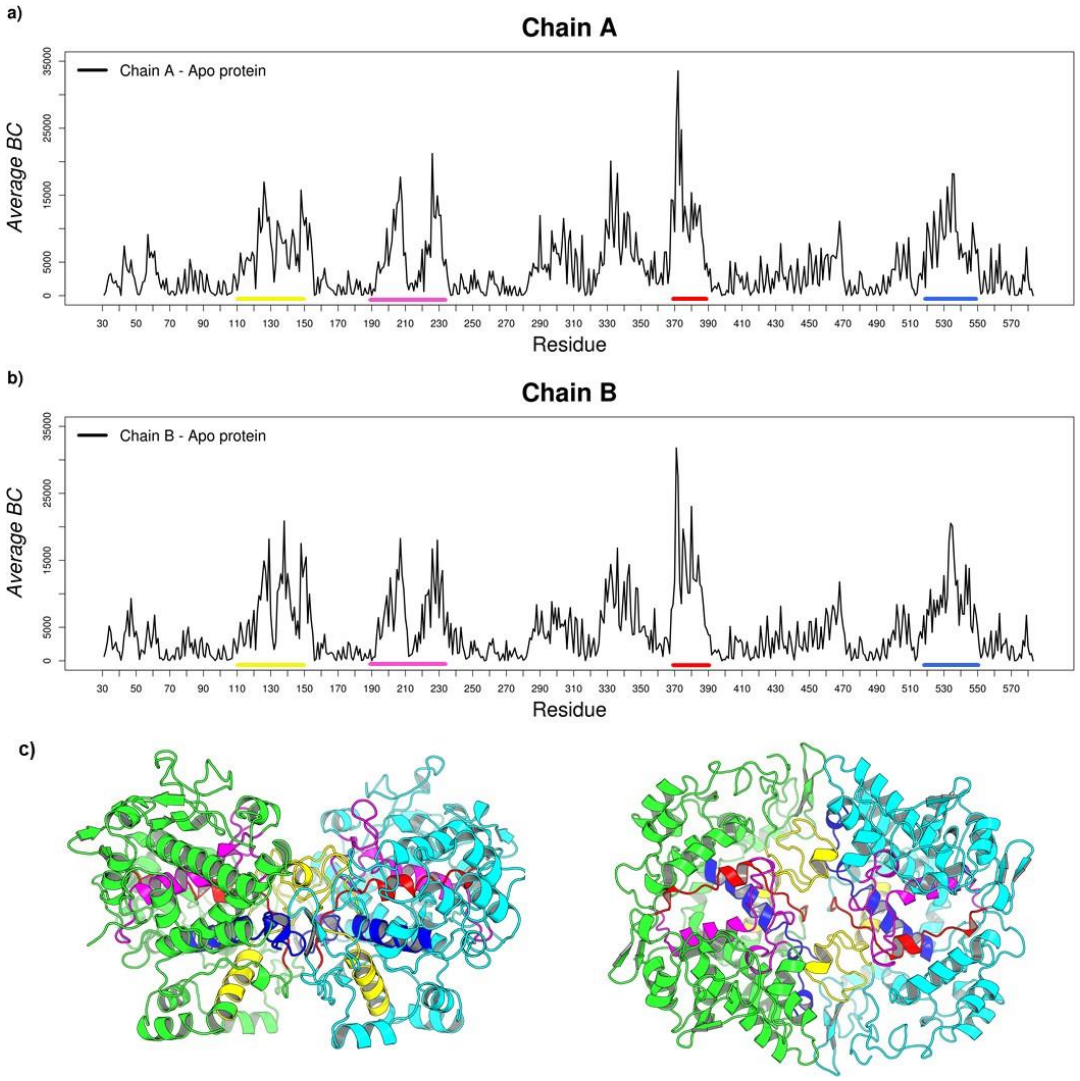


Figure 5.5: *Average BC* of wild-type monomer residues. Regions with peaks in BC are highlighted on the graphs and the protein structure with corresponding colours.

.

5.3.2 NsSNPs variants analysis

5.3.2.1 Global motions

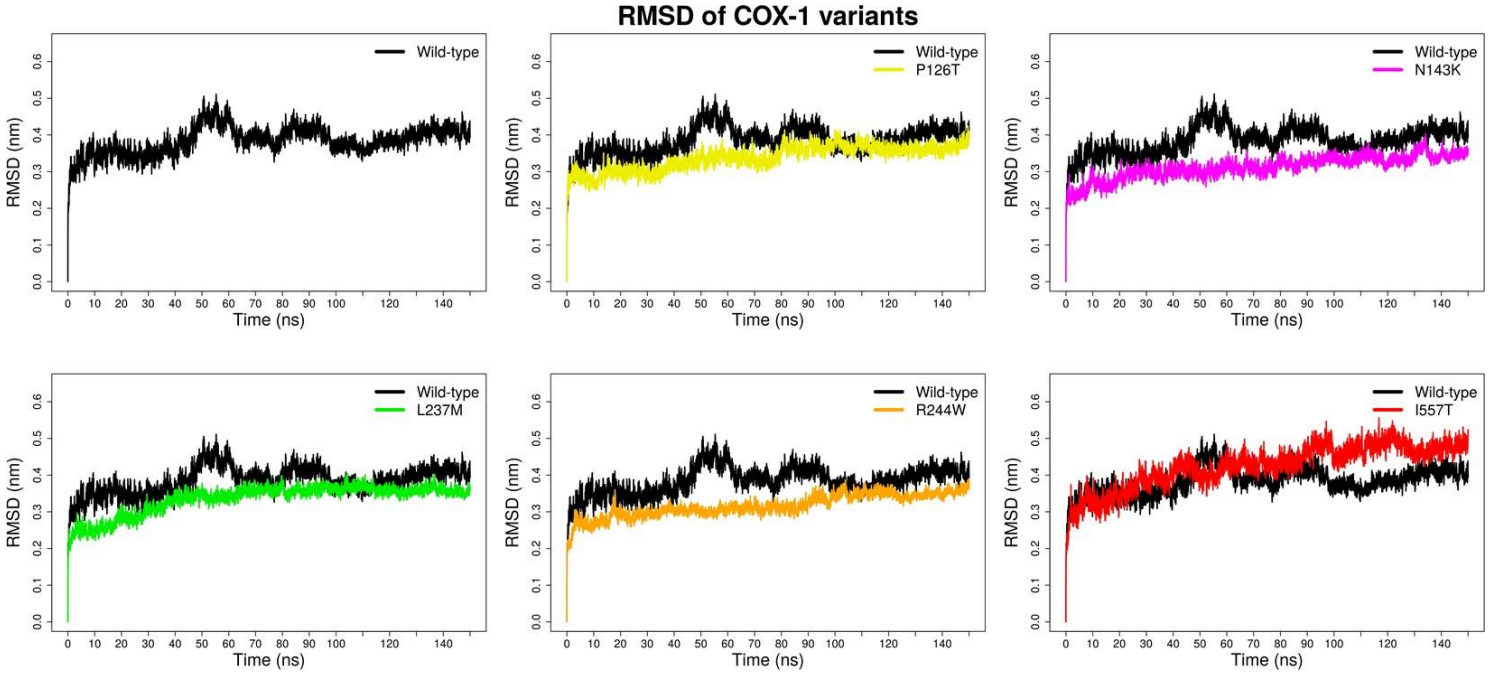


Figure 5.6: RMSD of COX-1 variants over the course of MD simulations in relation to the wild-type.

RMSD and RG RMSD and RG analyses of the COX-1 variants suggested a slight change in protein stability caused by presence of the SNPs. All deviations were within 0.1nm, except variation I557T, which exhibited a higher deviation, despite being predicted to contain the second least deleterious SNP [Figure 2.3} after L237M. The variant, however, does cause a very narrow change in radius of gyration.

Variants predicted to contain the most deleterious SNPs P126T, N143K and R244W, as well as L237M, confer a reduction in RMSD, suggesting the variants stabilised better than the wild-type over the course of the simulation.

True to the prediction of deleterious effects, variant R244W displays the largest decrease in radius of gyration at about 0.05nm, in Figure.5.7 while the rest of the variants converge with the wild-type. While the margin of difference is small, it implies a significant change in the internal amino acid residue network.

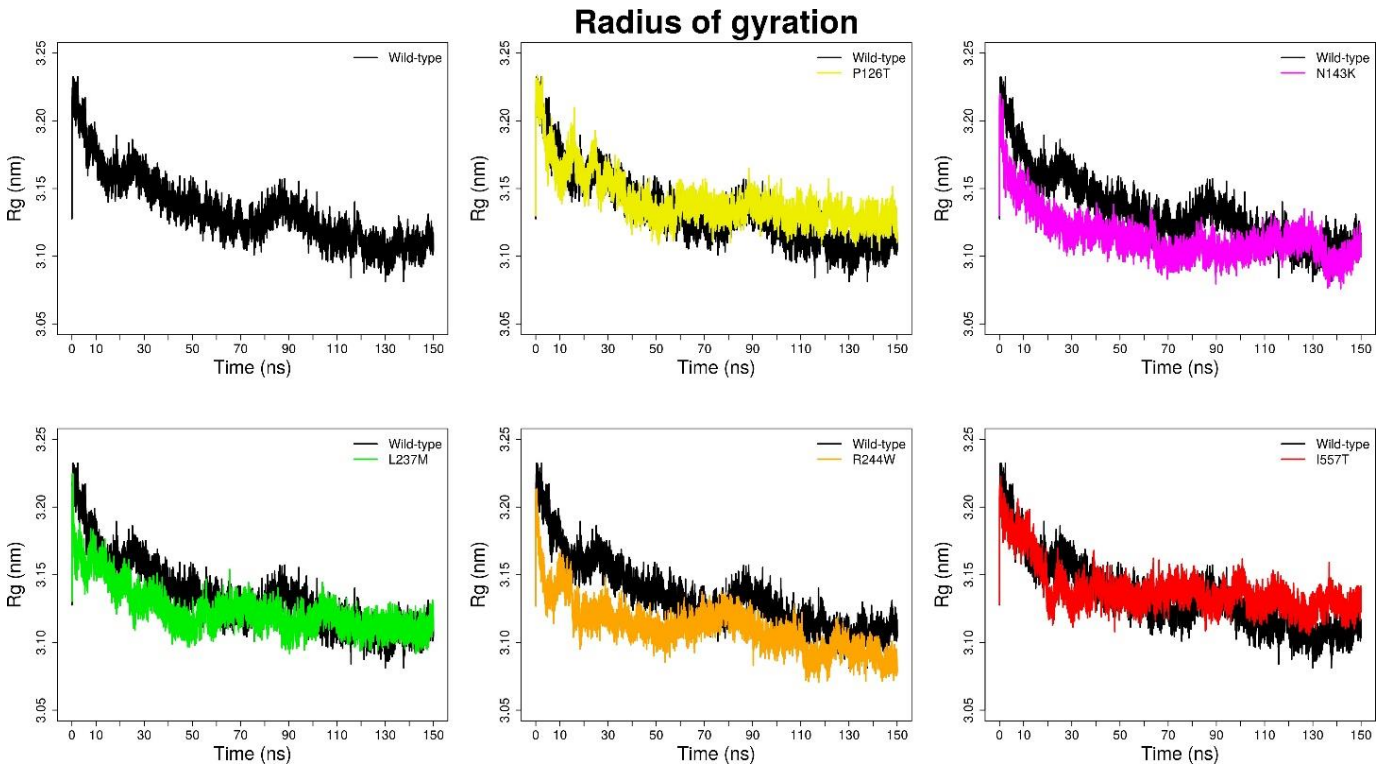


Figure 5.7: Radius of gyration of COX-1 variants in relation to the wild-type, over the length of the MD simulation.

.

5.3.2.2 COX-1 P126T

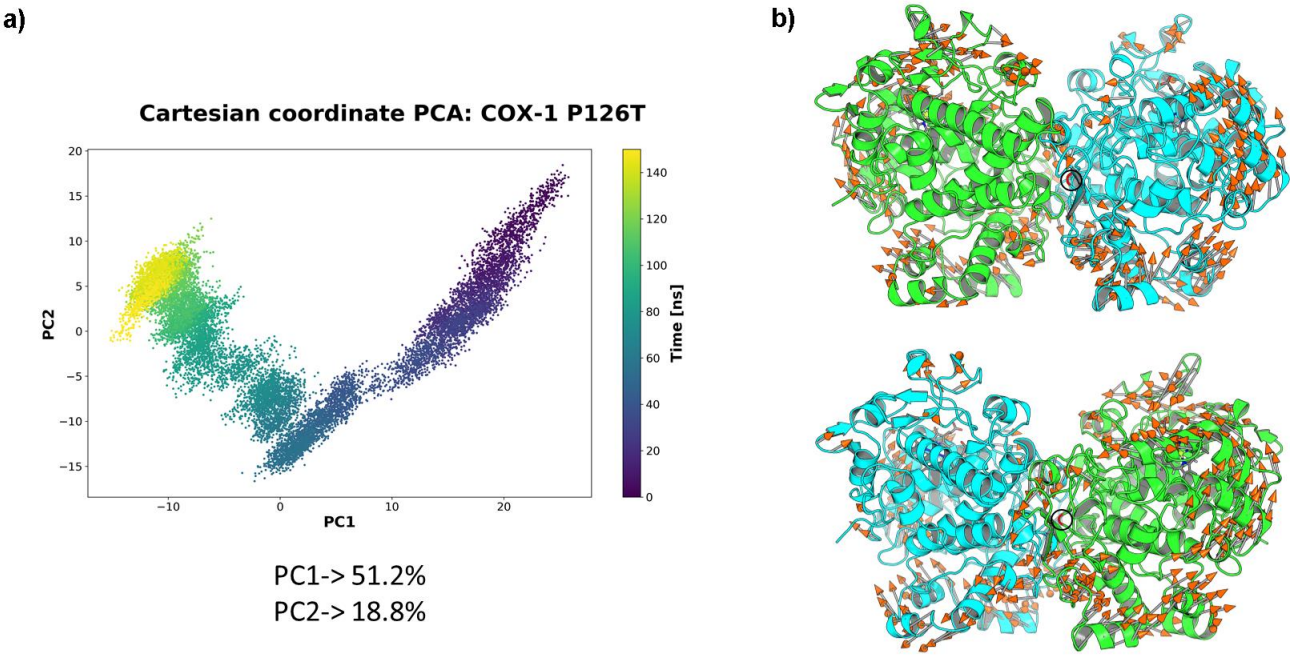


Figure 5.8: PCA plot of variant P126T over the course of the MD simulation, showing the variance represented by PC1 and PC2 (a). Motion of the protein is represented by arrows in (b), where the position of the mutation is encircled in black and highlighted in red.

PCA of variant P126T observed in Figure 5.8 showcased conformational change over the course of the trajectory, with PC1 representing 51.2% of the variance. Lack of motion in SNP positions and residues adjacent to them in both chains, suggested conformational change exhibited was more likely due to the asymmetrical behaviour of COX-1 monomers, than the presence of the SNPs. RMSD [Figure 5.6] and RG analyses [Figure 5.7] of variant P126T relative to the wild-type were in agreement with this, as the deviation seen in both is not significant.

RMSF and Network analysis Residue 126 is in what is referred to as the cross-talk region [] of the dimer interface, likely to be responsible for communication between the monomers, conveying allosteric behaviour from one to the other. This cross-talk can be seen in the contact maps [Figure 5.10], of both chain A and chain B, where the residue network was largely maintained, suggesting no adverse effects on cross-talk between the monomers.

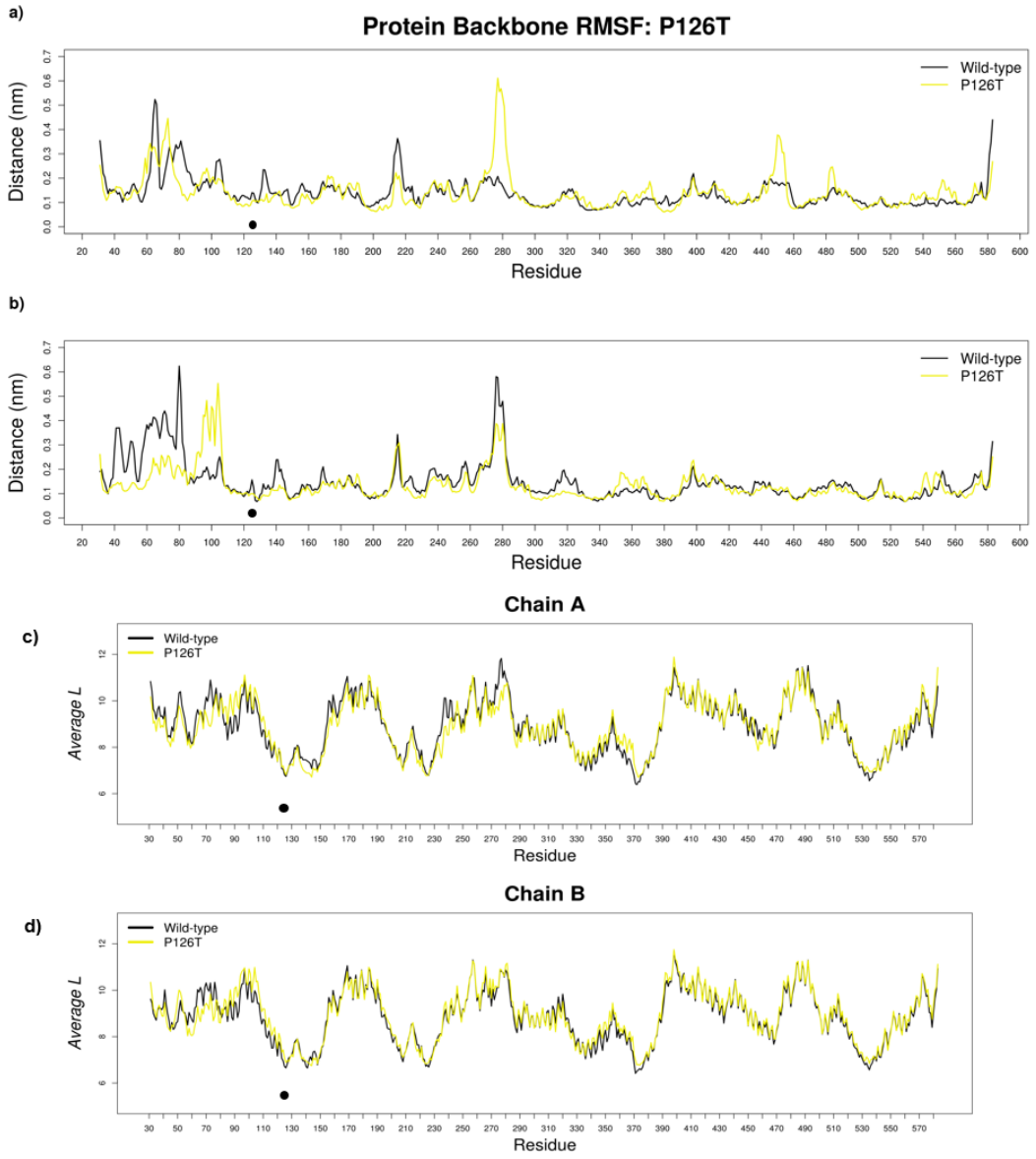


Figure 5.9: RMSF (a-b) and average L (c-d) of variant P126T, in relation to the wild-type, highlighting position of the SNP with a black dot.

The *average BC* [Figure 5.10] and *average L* showed no distinct change, further suggesting the variant did not affect residue interaction. The RMSF [Figure.5.9], however, demonstrated a general change in monomer flexibility in the variant, relative to the wild-type; but none directly in network with the position of the SNP. Chain A showed increased flexibility in the loop residues 260-280 and the helix from residue 440-460, not matched in chain B. The MDB region of chain B exhibited a decrease in flexibility. Despite change from a hydrophilic to a hydrophobic residue substitution P126T appeared to have no immense effect on the internal dynamics of the protein, and communication between the monomers.

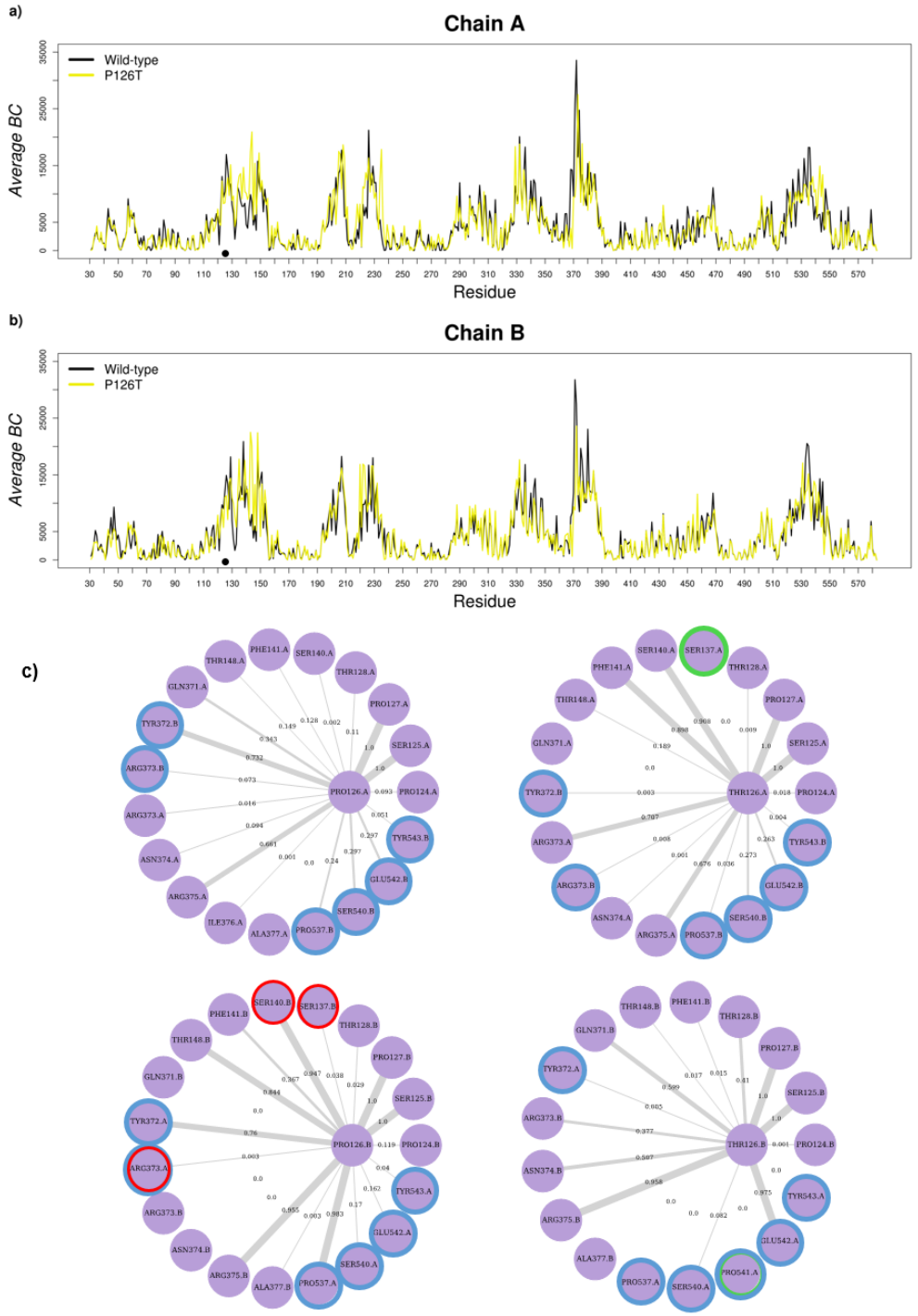


Figure 5.10: Average BC (a-b) and contact maps (c) of variant P126T, relative to the wild-type. SNP positions are represented by a black dot in the *Average BC* plot. The contact maps show residue interaction between the wild-type residues (right) and the variant (left). Interaction losses are circled in red, and gains in green. Residues interacting across the dimer interface are circled in blue.

.

5.3.2.3 COX-1 N143K

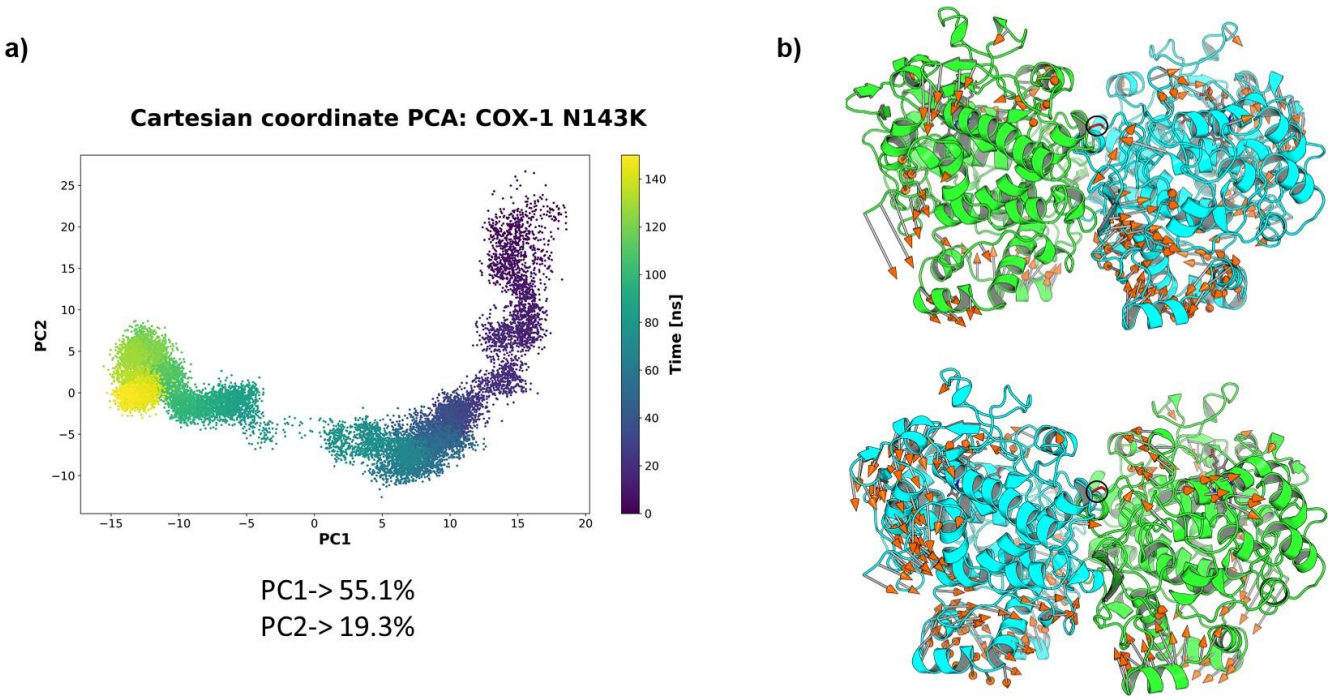


Figure 5.11: PCA plot of variant N143K over the course of the MD simulation, showing the variance represented by PC1 and PC2 (a). Motion of the protein is represented by arrows in (b), where the position of the mutation is encircled in black and highlighted in red.

PCA The PCA of variant N143K supported the notion of conformational of the protein during the simulation. PC1 covered the largest variance at 55.1% and PC2 19.3%. Even with the lower value, PC2 exhibited a relatively large Cartesian coordinate difference of 25. Chain A (in green) accounted for the largest motions, as shown in Figure 5.11(b), further adhering to the theory of monomer asymmetry.

RMSF and Network analysis Analysis of variant N143K, relative to the wild-type revealed a general change in the flexibility and *average L* in the EGF and MDB regions [Figure 5.12], which are significantly distant from position 143 in the dimer interface, and do not appear in the contact maps in Figure 5.13 .This change in RMSF and average was however not endorsed by *average BC*, where importance of residues in the network remained unaffected. This may be due to the preservation of hydrophilicity in the substitution.

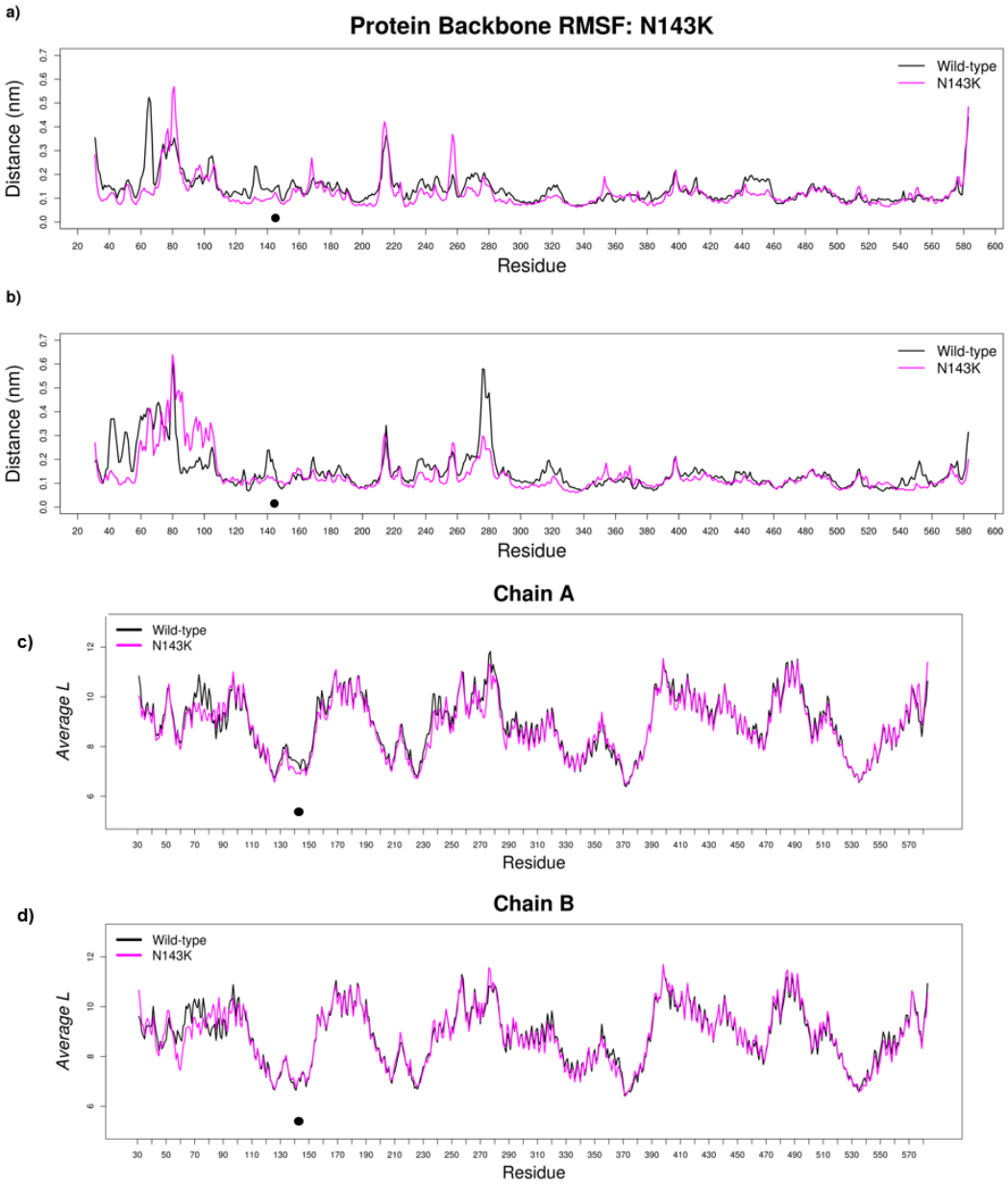


Figure 5.12: RMSF (a-b) and *average L* (c-d) of variant N143K, in relation to the wild-type, showing position of the SNP using a black dot.

Due its location in the dimer interface, the contact maps for each monomer showed numerous interactions with residues in the other. However, due to the obvious structural differences between asparagine and lysine, there was a decrease in cross-interface interactions in the variant, including one with SNP position Leu-237 in chain B. Interestingly, both monomers gained an interaction with Val-144 of the other (circled in blue in Figure 5.13), though the importance of said interaction was not evident. Chain A, however, showed an increase in residue-residue interactions

internally, which likely contributed to the decrease in RG seen in Figure 5.7).

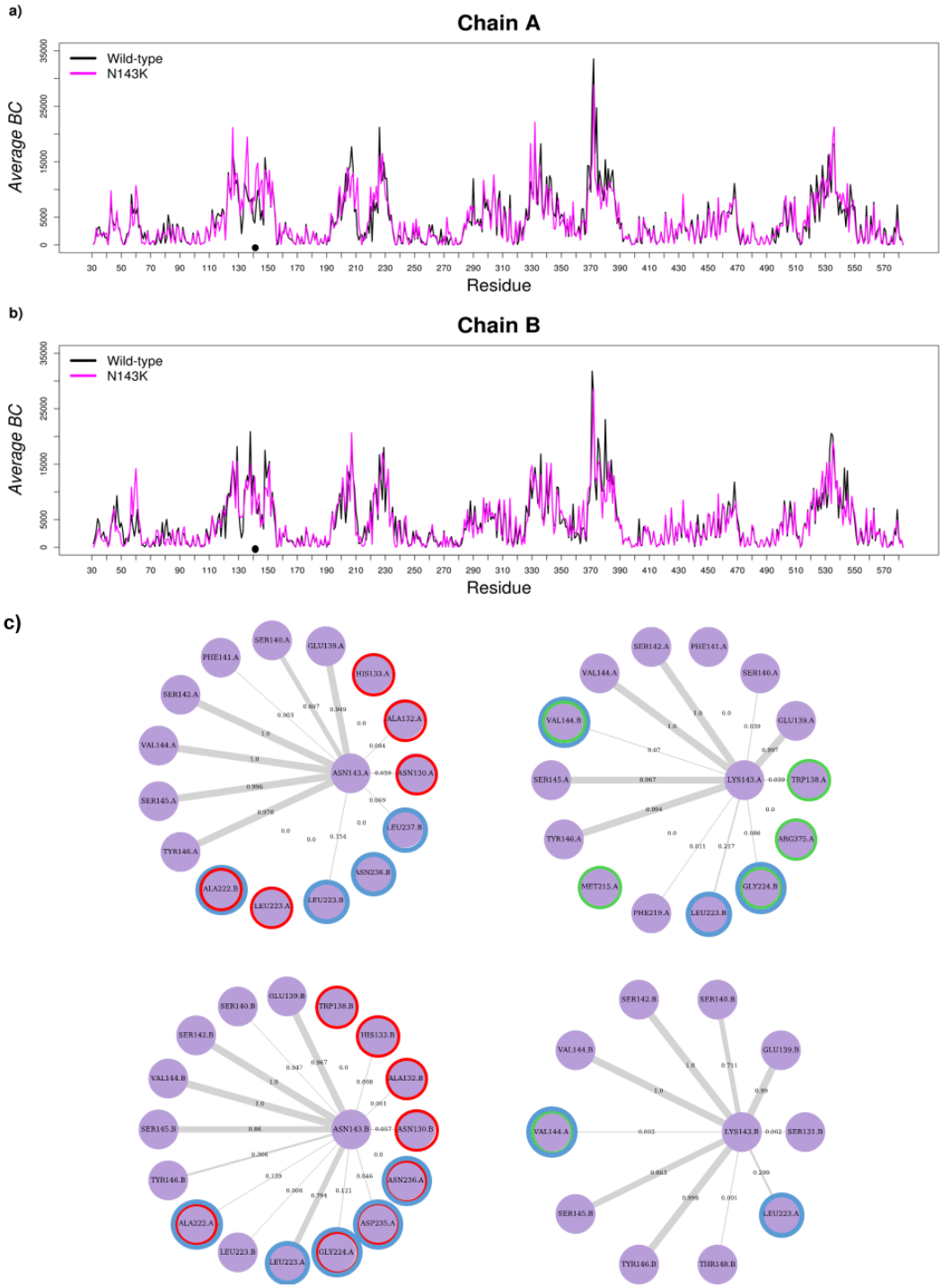


Figure 5.13: Average BC (a-b) and contact maps (c) of variant N143K, relative to the wild-type. SNP positions are represented by a black dot in the Average BC plot. The contact maps show residue interaction between the wild-type residues (right) and the variant (left). Interaction losses are circled in red, and gains in green. Residues interacting across the dimer interface are circled in blue.

As mentioned in chapter 2, Residue 143 is the site of N-linked glycosylation. COXs are glycosylated on asparagine in all organisms, but Asn-143 is the only one that is absolutely conserved [91] [Figure 2.1 & Figure 3.3)]. Though to date no studies exist on glycosylation in COX-1, several exist on its function and importance in COX-2. COX-2 exists as 72 and 74 kDa glycoforms, where glycosylation of the 72 kDa glycoform at residue Asn-580 affects COX-2 turnover and activity [241], homodimerisation [242] and efficacy of several NSAIDs, aspirin included [243]. Other researchers have found glycosylation of COX-2 at several asparagine residues (53,130 and 396) necessary for proper folding of COX-2 into an active enzyme [244]. These studies point to a probable importance of the Asn-143 glycosylation, lost in the substitution to lysine, explaining the SNP prediction scores in Chapter 2 [Figure 2.2]. In summation, while wet-lab studies imply importance of the position, *in silico* analysis of variant N143K provided insight into the effect of an amino acid substitution in said position. As such the two complemented each other in confirming probable impact of the SNP on protein function.

5.3.2.4 COX-1 L237M

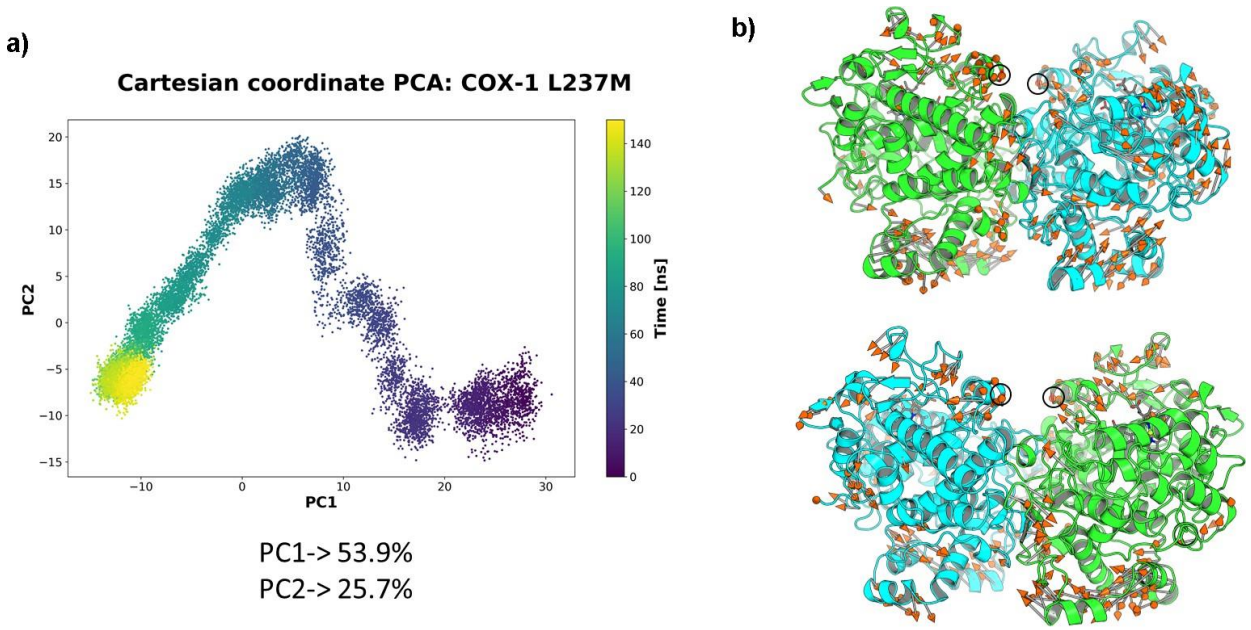


Figure 5.14: PCA plot of variant L237M over the course of the MD simulation, showing the variance represented by PC1 and PC2 (a). Motion of the protein is represented by in arrows in (b), where the position of the mutation is encircled in black and highlighted in red.

PCA The PCA plot of variant L37M in Figure 5.14a represented 87.7% of the variance observed in the system. PC1 covered the largest variance of the system, implying conformational change. The SNPs were located helices that contributed to movements [Figure 5.14b]. The motion observed in that particular helix in chain B (shown in cyan) was unique to the variant, implying an effect of the global protein motions.

RMSF and Network analysis The Leu-237 residue is conserved across the COX-1 and, interestingly, COX-2 sequences analysed in this study [Figure 3.3]. The residue is likely involved in the dimerisation interaction, with significant contacts to sugars linked to Asp-143 of the opposite monomer. This particular interaction was observed in the wild-type and variant contact maps of chain B in Figure 5.16. As such, the L237M variant was assumed to influence catalytic activity due to predicted impact on dimerization [93]. Interestingly, the contact maps showed a greater number of cross-monomer interactions in the wild-type chain B, which was inverted in chain A, exhibiting an obvious effect of the SNP on dimerisation. Though the methionine substitution maintains some interactions and hydrophobicity, the residue is larger and more flexible.

This structural difference is likely the cause of the changes in the residue network that were observed in the contact maps, and the *average BC* fluctuations, seen in both monomers from residue 120-150. Methionine can additionally be subject to oxidation that could further impact the dimer contacts.

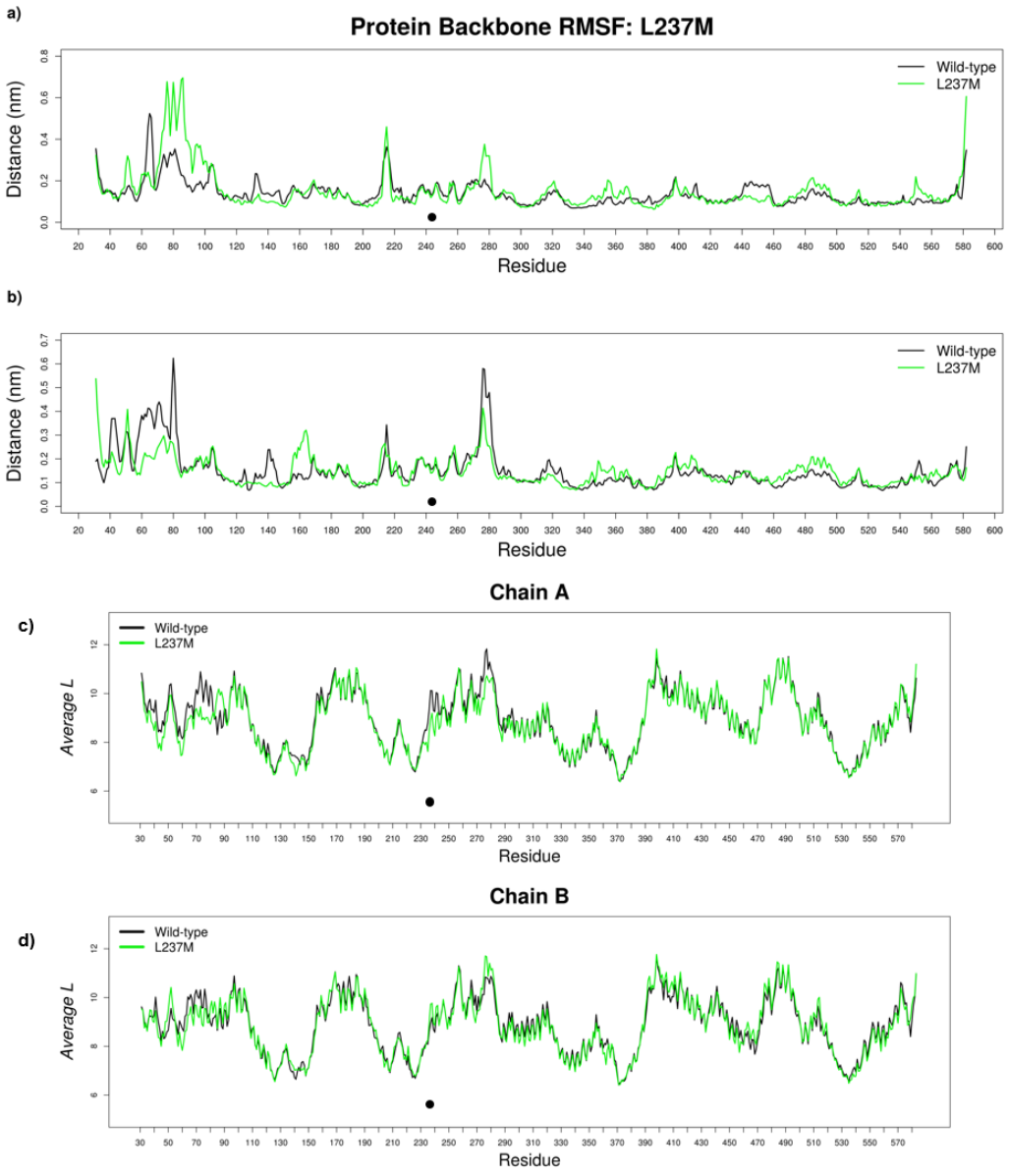


Figure 5.15: RMSF (a-b) and average L (c-d) of variant L237M, in relation to the wild-type, showing position of the SNP with a black dot.

The RMSF and average L analyses of the variant in Figure 5.15 showed changes in the EGF and MDB domains of the monomers, much like those observed in variant N143K [Figure 5.13] Whether this was due to the relation- ship between Leu-237 and Asp-143 is

subject to further analysis.

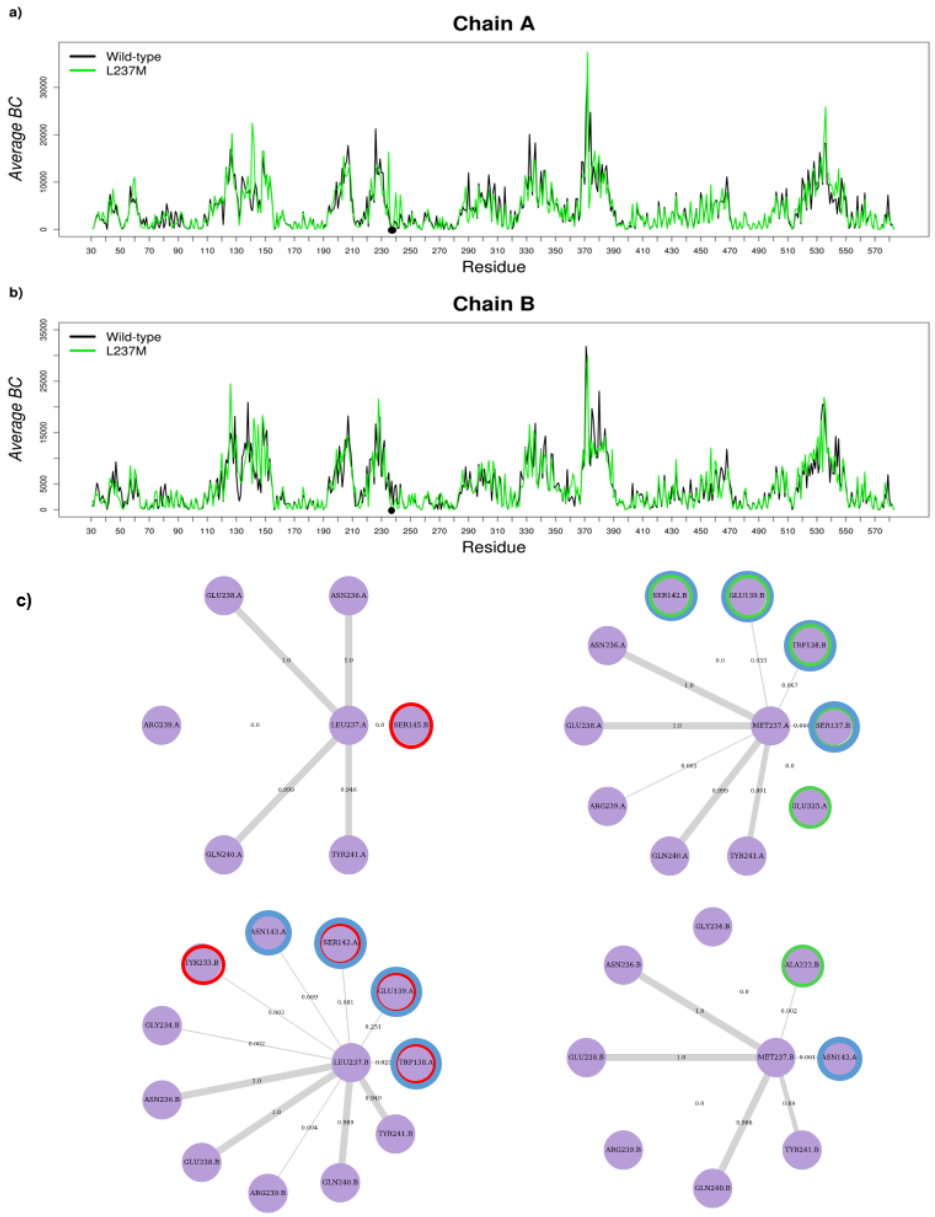


Figure 5.16: Average BC (a-b) and contact maps (c) of variant N143K, relative to the wild-type. SNP positions are represented by a black dot in the Average BC plot. The contact maps show residue interaction between the wild-type residues (right) and the variant (left). Interaction losses are circled in red, and gains in green. Residues interacting across the dimer interface are circled in blue.

Wet lab studies using the NSAID indomethacin [93], a non-selective COX inhibitor with greater potency for COX-1 than COX-2, showed no significant differences in L237M variant IC_{50} value compared to the wild-type. The L237M variants appeared more sensitive

to inhibition at higher concentration of the drug, though the IC_{50} value estimation was not statistically different. These findings suggested a reduction in enzyme activity in the L237M variant, as well as reduced COX-1 metabolic activity due to a combination of indomethacin treatment and other variants including L237M compared with indomethacin treatment of wild-type COX-1.

Analyses of the variant dynamics, seen in the RMSD, RG and PCA exhibited changes in global and local motion which may attribute to what is observed in the wet lab studies.

5.3.2.5 COX-1 R244W

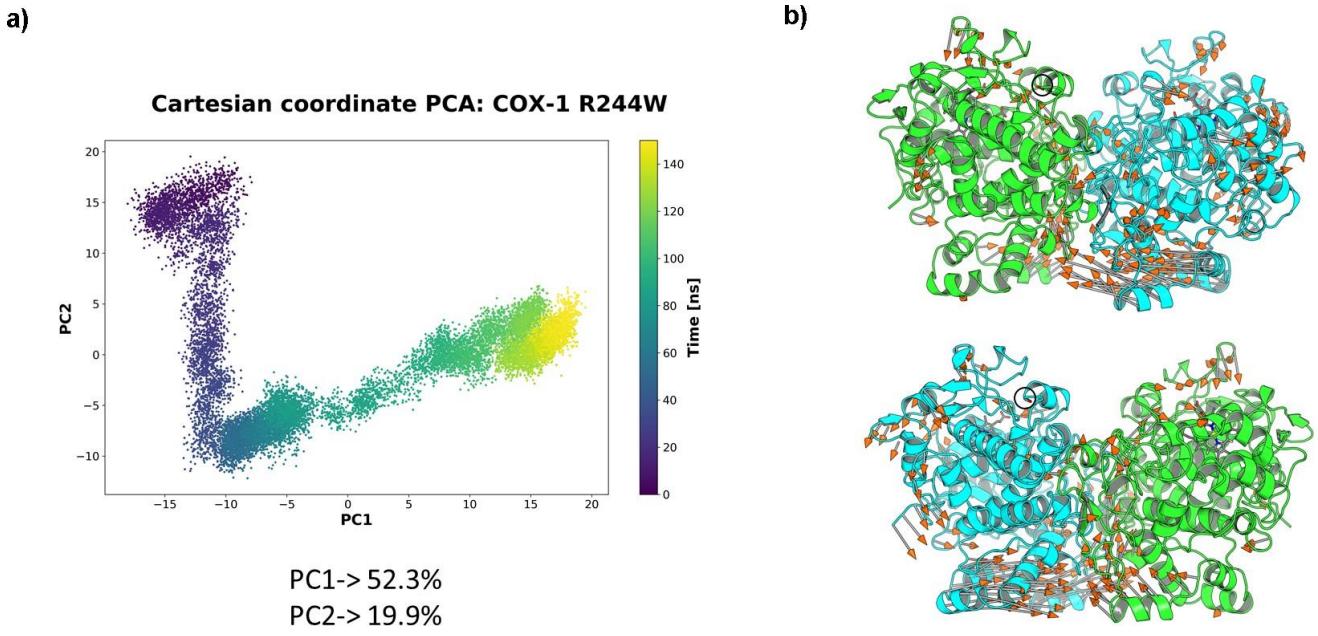


Figure 5.17: PCA plot of variant R244W over the course of the MD simulation, showing the variance represented by PC1 and PC2 (a). Motion of the protein is represented by arrows in (b), where the position of the SNP is encircled in black and highlighted in red.

PCA PC1 of the PCA analysis of variant R244W [Figure 5.17], explained 52.3% of the variance exhibited, and PC2, 19.9%. Chain B helices on the MBD were responsible for the greatest of the motions as seen in Figure 5.17b (in cyan); while no motion was observed at the site of, or regions adjacent to the SNP.

RMSF and Network analysis RMSF analysis of variant R244W in Figure 5.18 showed an increase in flexibility from residue 270-285 in chain A, that was complimented with a decrease in chain B. Whether this was due to possible monomer cross-talk cannot be confirmed, as it was not endorsed by the contact maps in Figure 5.19. These fluctuations were, however, in agreement with behaviour exhibited in the PCA [Figure 5.17b]. A decrease in *average L* between residues 220 and 250, in both monomers, was observed reflecting a decrease in distance between residues, likely due to the presence of hydrophobic tryptophan. This decrease in *average L* likely contributed to the lower RG seen in Figure 5.7, in relation to the wild-type.

Though predicted to be deleterious in Chapter 2, the substitution of arginine with tryptophan in a conserved helix of the catalytic domain appeared to minimally alter the residue network.

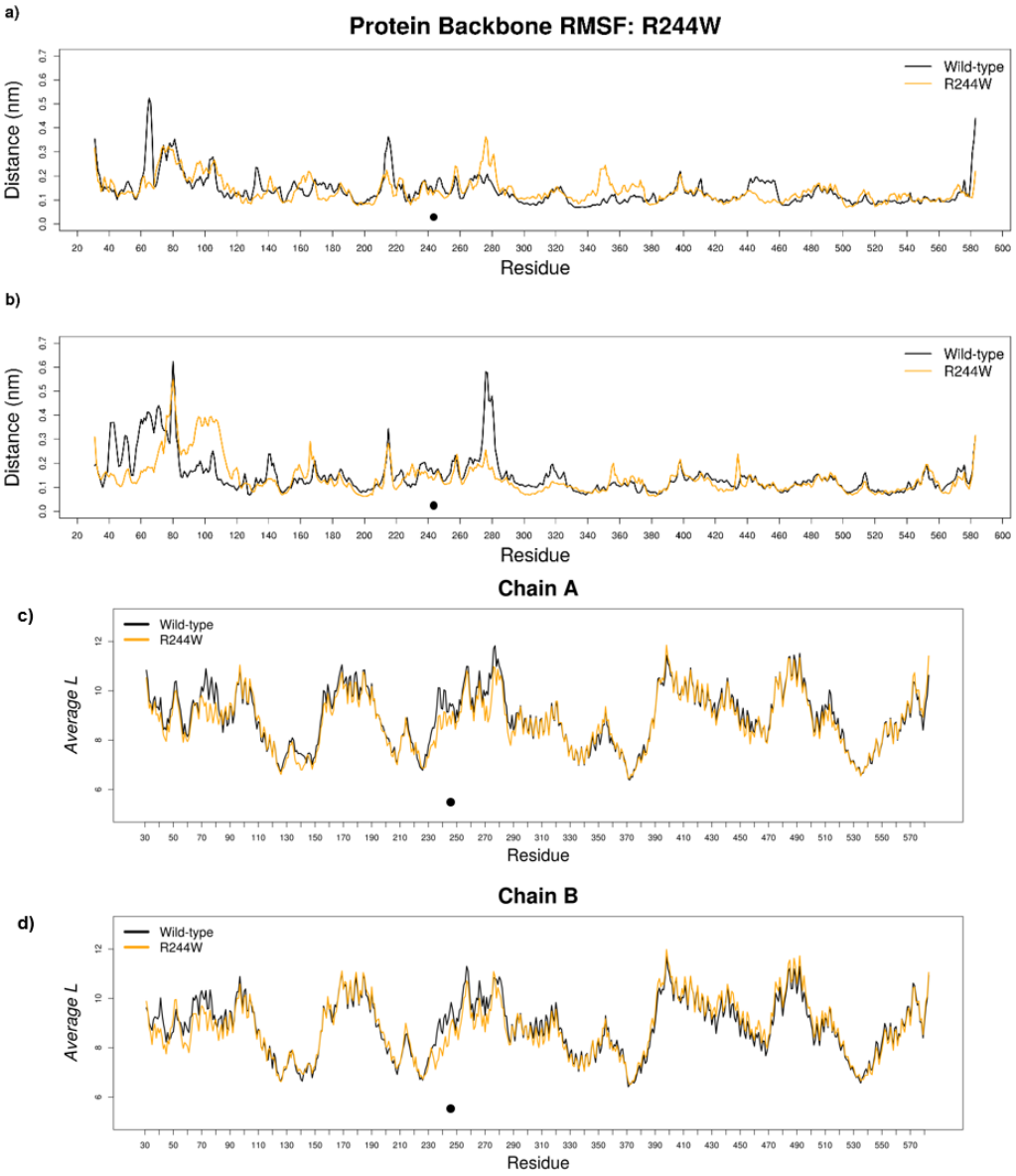


Figure 5.18: RMSF (a) and average L (b) of variant R244W, in relation to the wild-type, showing position of the SNP using a black dot.

The interactions in chain A remained ultimately unchanged [Figure 5.19], despite the differences in the substitution residue physiochemical properties. The similar sizes of the wild-type and variant residues may have been the cause of this.

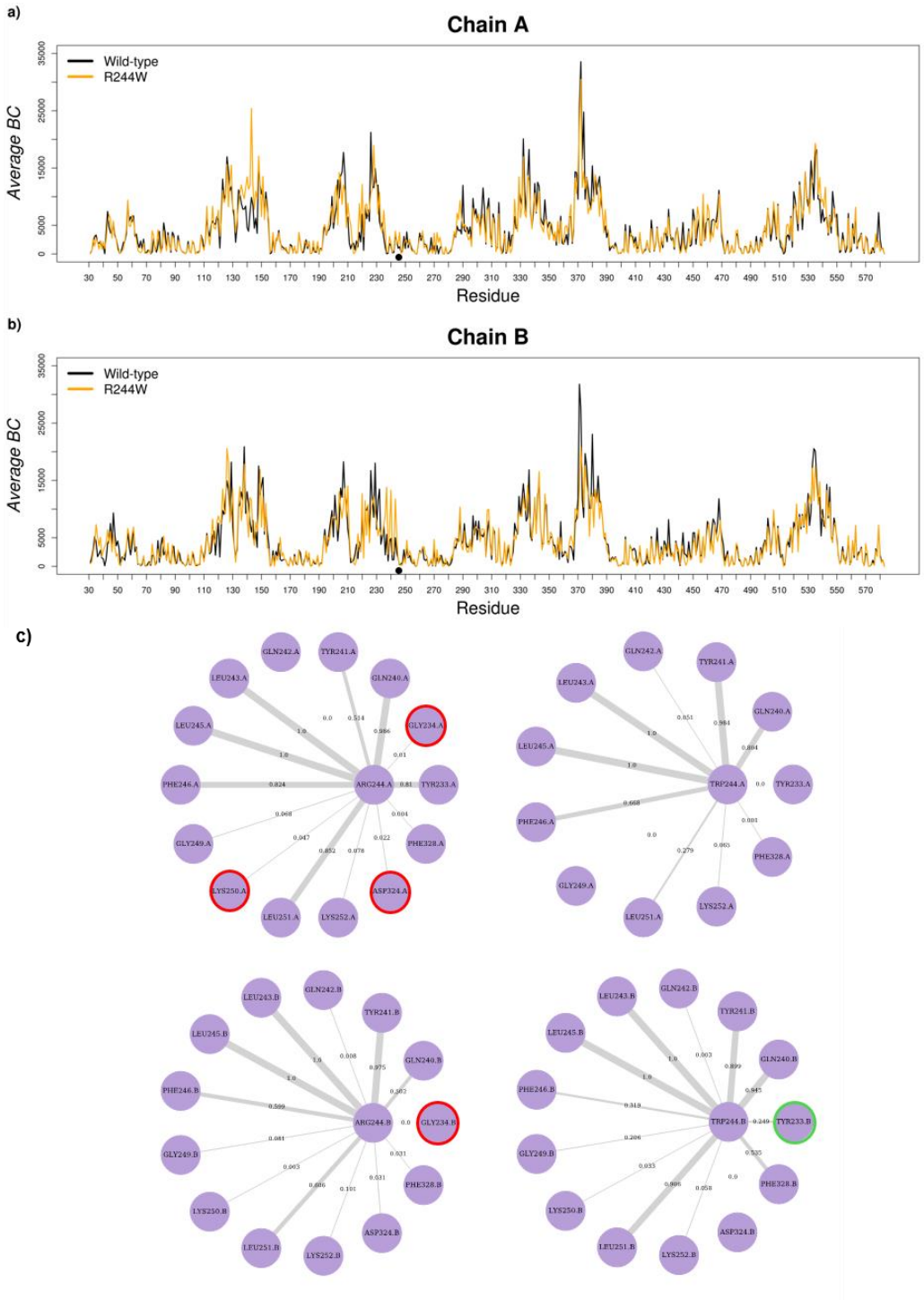


Figure 5.19: *Average BC* (a-b) and contact maps (c) of variant R244W, relative to the wild-type. SNP positions are represented by a black dot in the *Average BC* plot. The contact maps show residue interaction between the wild-type residues (right) and the variant (left) Interactions losses are shown in red and gains in green.

5.3.2.6 COX-1 I557T

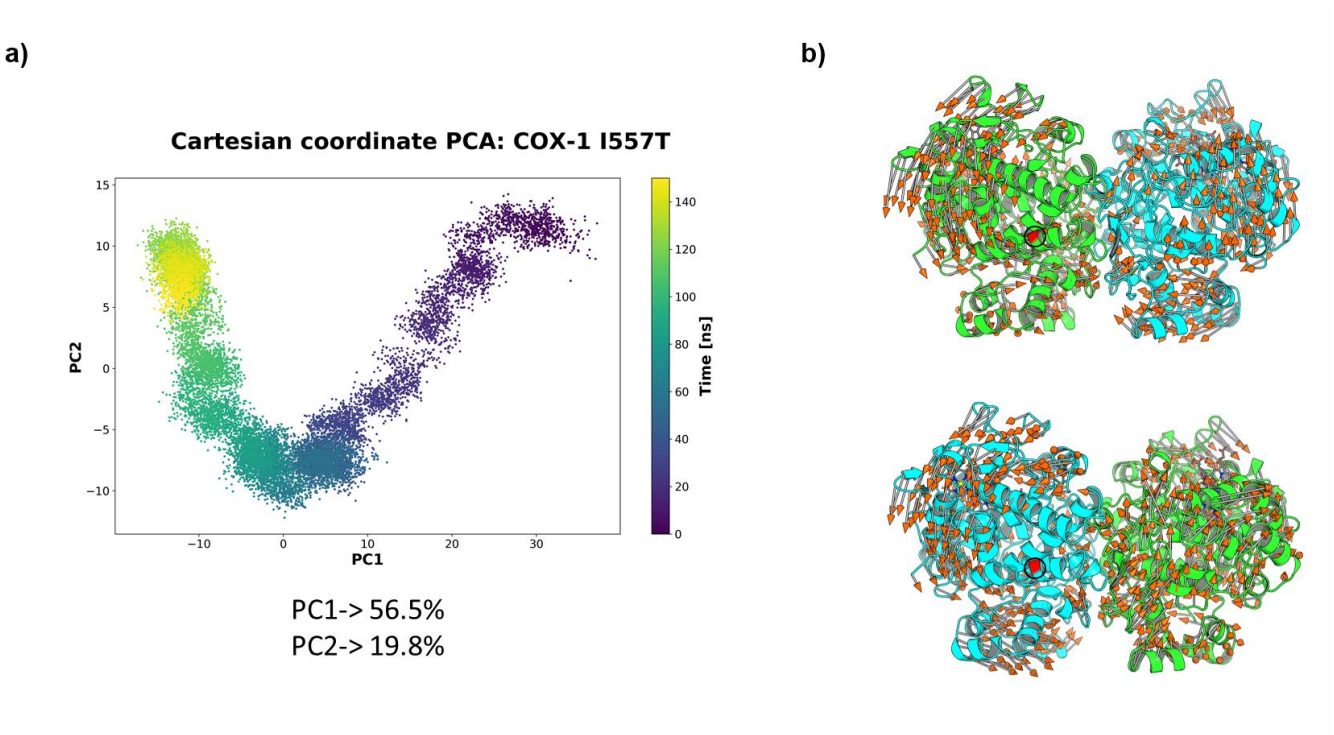


Figure 5.20: PCA plot of variant I557T over the course of the MD simulation, showing the variance represented by PC1 and PC2 (a). Motion of the protein is represented by arrows in (b), where the position of the SNP is encircled in black and highlighted in red.

PCA In the PCA of variant I557T shown in Figure 5.20, PC1 covered 56.5% of the variance, which was the largest of all the COX-1 variants. While the minimal movement was observed on the helices where the SNPs were located, the rest of the protein exhibited notable motion [Figure 5.20b], with the largest being observed in the loops. This behaviour tied in with deviations observed in variant I557T RMSD and RG [Figure 5.6 & Figure 5.7], in relation to the wild-type.

RMSF and Network analysis Residue fluctuation was seen in Figure 5.21 in the loops of EGF and MDB domains. Chain A exhibited the most deviation from the wild-type residue flexibility. The fluctuation was also observed in the *average L*.
Contrary to its lower deleterious scoring compared to the other SNPs, the *average BC* in Figure 5.22 of the variant was quite distinct from the wild-type, with highest *average BC* being attributed to residue 125, as opposed to 371, thus exhibiting immense change in the residue network. The change from hydrophobic, apolar residue to the exact opposite, likely disrupted the residue network, resulting in the changes in *average L* as well.

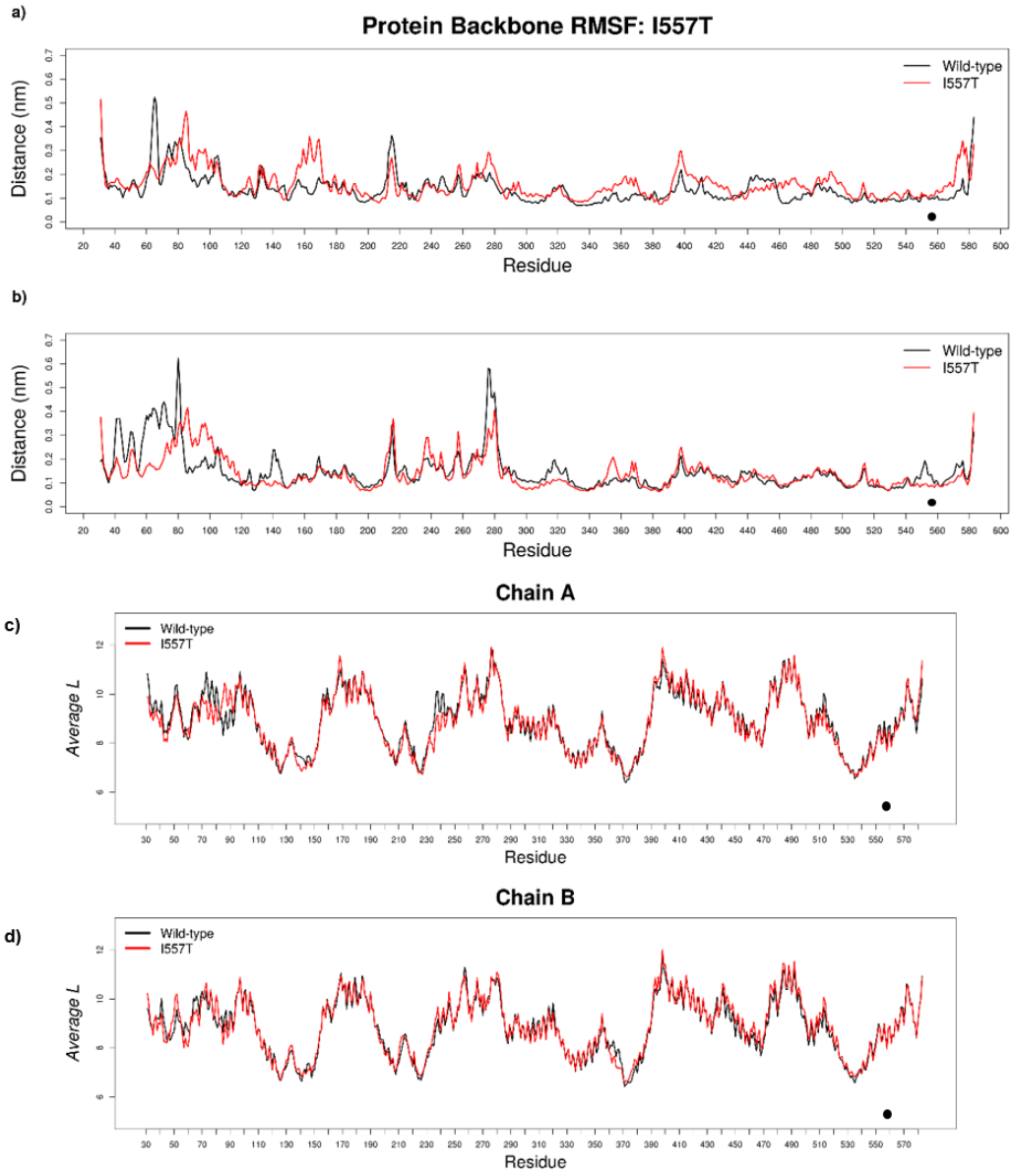


Figure 5.21: RMSF (a-b) and average L (c-d) of variant R244W, in relation to the wild-type, showing position of the SNP using a black dot.

Contact maps of the variant in Figure 5.22 show a general increase in interactions with the polar threonine. Interestingly, while global motion and behaviour of variant varied significantly from the wild-type, the monomers exhibited similar tends in RMSF, *average BC* and *L*; and contact maps.

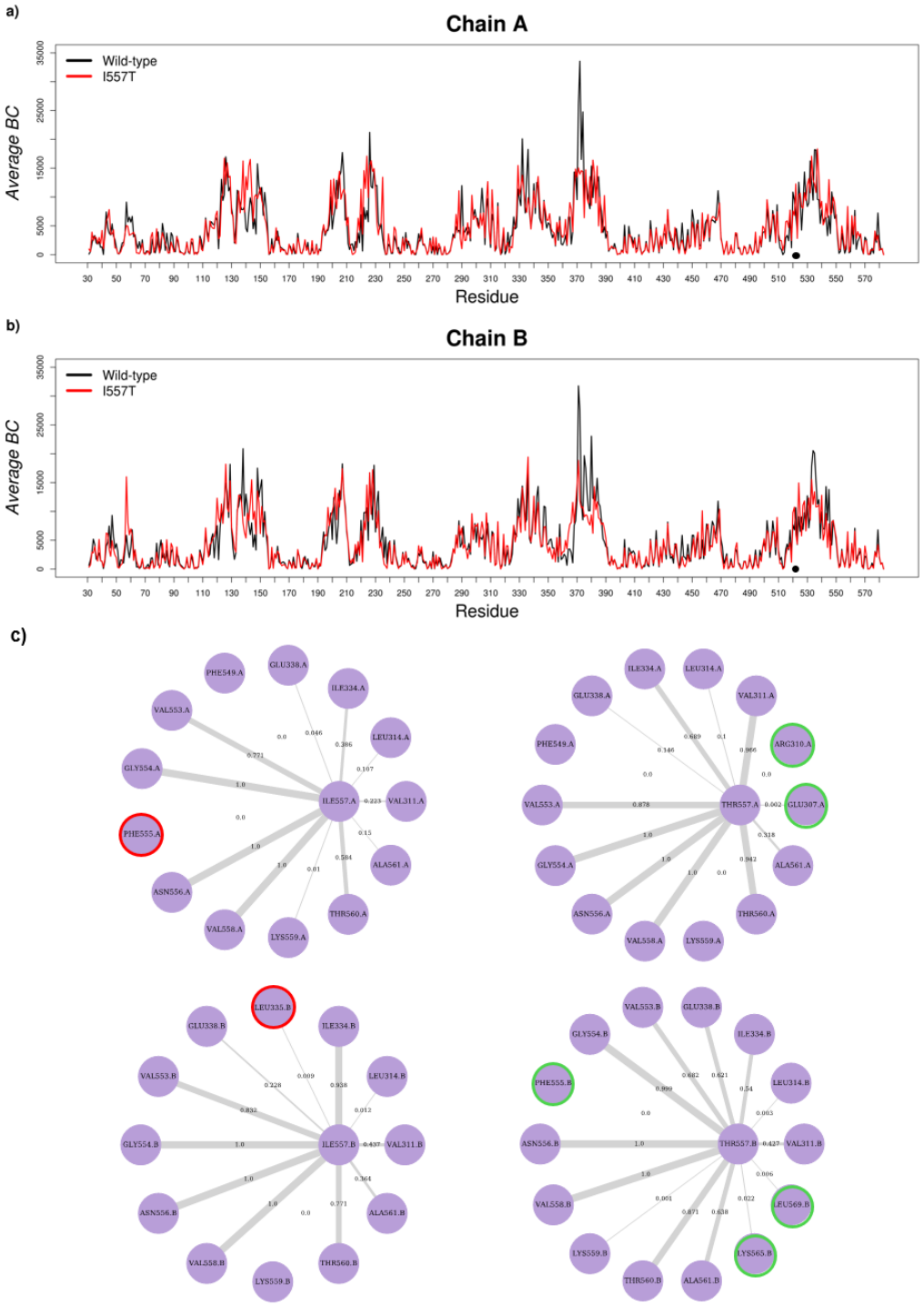


Figure 5.22: *Average BC* (a-b) and contact maps (c) of variant I557T, relative to the wild-type. SNP on *average BC* plot is represented by a black dot. The contact maps of the show residue interaction between the wild-type residue (right) and the variant (left). Interaction losses are circled in red, and gains in green.

5.3.3 Docking Analysis

In the drug discovery part of this study 11 ligands presented as viable drug compounds, based on molecular docking poses, protein-ligand interactions and binding energy. These ligands selected from the molecular docking aspect of the study were further screened by performing a 10ns MD simulation to assess ligand stability. The binding poses were re-scored after the short MD, and these scores are listed in Table 5.1. The 10ns binding energy scores and the ligand RMSDs over the MD run were used to further screen the hits. The MD simulations were extended to 100ns for further analysis.

Table 5.1: Ligand binding energies re-scored after a 10ns MD simulation

Active Site	Ligand	Binding energy t=0ns (kcal/mol)	Binding energy t=10ns (kcal/mol)	Binding energy t=0ns (kcal/mol)	Binding energy t=10ns (kcal/mol)
		Chain A		Chain B	
Cyclooxygenase	SANC239	-7.7	-4.9	-6.4	-6.8
	SANC521	-8.7	-6.3	-7.8	-7.2
	SANC627	-8.1	-8.2	-6.0	-6.8
	SANC721	-9.4	-8.7	-4.3	-8.9
	ZINC925	-10.5	-9.8	-6.4	-6.9
	ZINC3113	-10.3	-9.6	-7.5	-9.6
	ZINC4587	-9.0	-9.1	-8.2	-7.1
	ZINC4671	-10.7	-9.3	-6.6	-6.5
	ZINC4749	-10.6	-9.1	-8.8	-8.7
Peroxidase	ZINC4394	-9.4	-4.2	0.6	-4.8
	ZINC4591	-9.5	-5.9	-7.9	-4.7

Rescoring of top-ranking ligands binding energies after 10ns MD simulations, shown in Table.5.1, present irregular binding energy changes across the ligands and monomers. Binding energy increase in one monomer complimented by decrease in the other, and vice versa. Due to the inconsistency of the scores, the MD runs for all ligands were extended.

RMSD analysis of the ligands after a 90ns MD simulation extension, seen in Figure 5.23, shows changes in ligands stability for quite a number of ligands. As in the rescoring procedure, ligand behaviour is not particularly uniform across the monomers. Ligands SANC239, SANC721 and ZINC4671 are the exception to this behaviour, with the ligands exhibit a minimal amount of deviation for both chains over the duration of the simulation

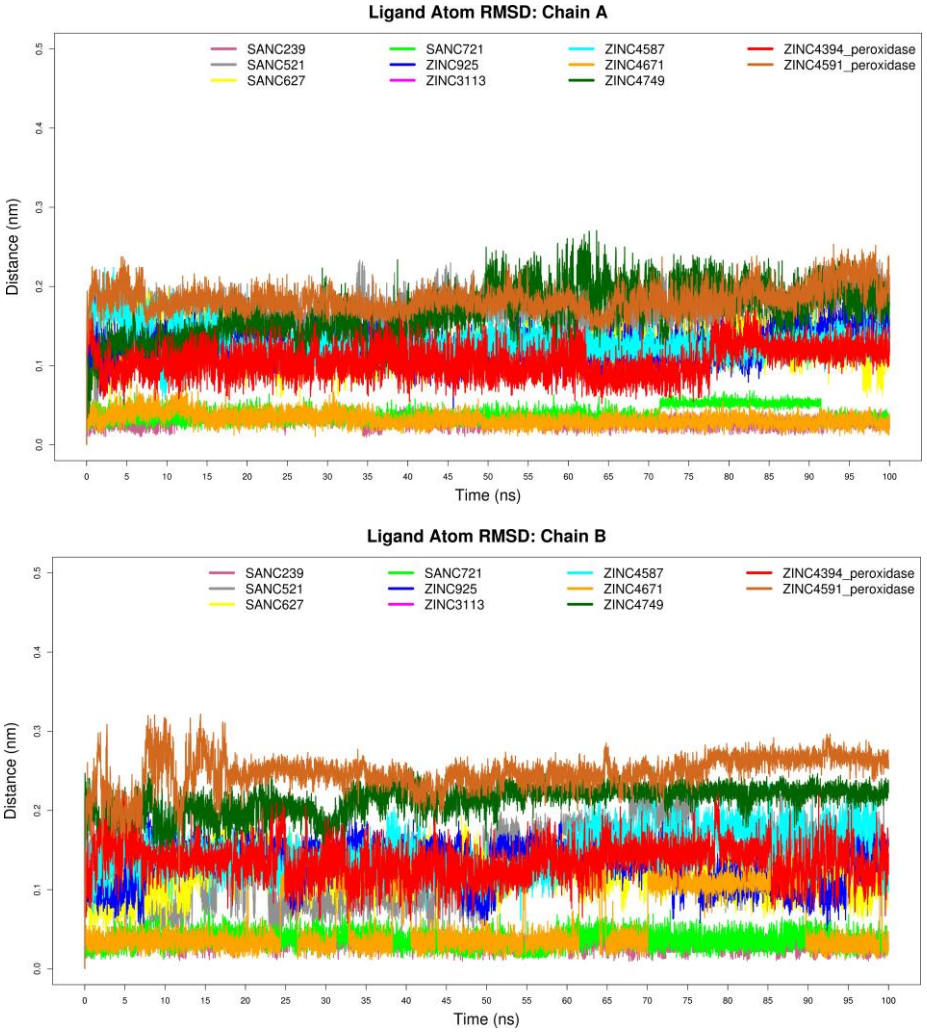


Figure 5.23: Ligand RMSDs in chain A (a) and chain B (b) during the 100ns MD simulation.

5.3.3.1 SANC239

Due to the asymmetric behaviour of ligands in each of the monomers, one monomer was assumed to be the catalytic monomer (E_{cat}) and discussed. In the case of SANC239, while both exhibit stability in RMSD [Figure 2.3 & S.Figure 8]; there was a drastic decrease in binding energy in chain A, which was contrasted by an increase in chain B [Table.5.1]. As such protein-ligand interaction in chain B will be discussed.

Visual analysis Snapshots of the protein-ligand complex at varying stages over the course of the MD simulation in Figure 5.24 showed a change from hydrogen bonding with Try-384 and Met-521, to Ser-529 and Tyr-386. The newly formed hydrogen bonds were maintained throughout the simulation, matching the interactions of salicylic acid.

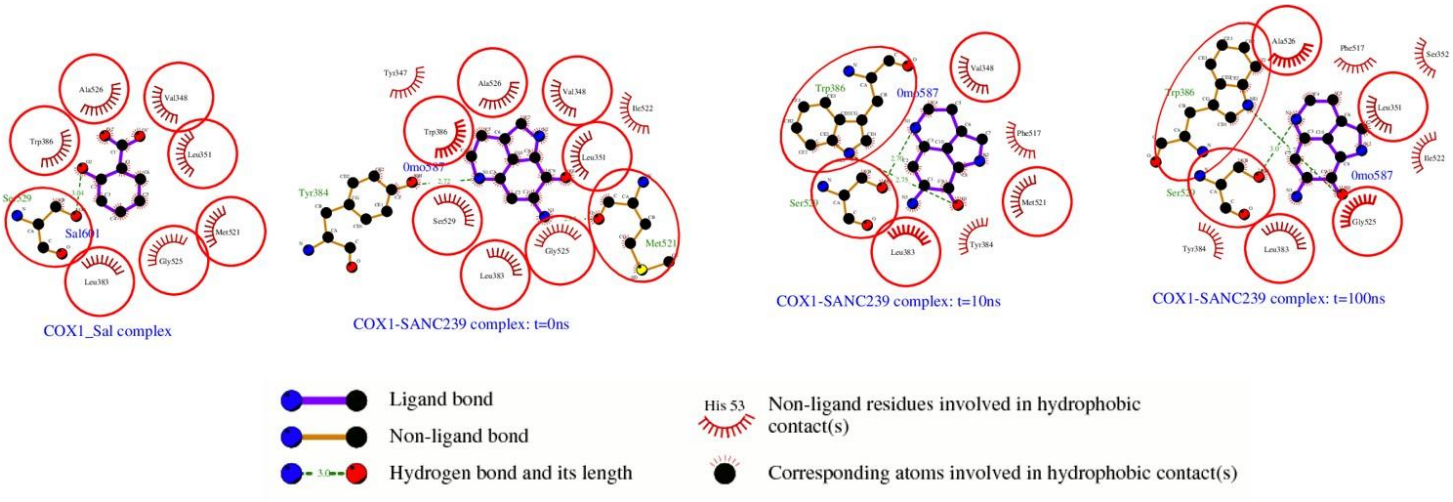


Figure 5.24: LigPlot+ protein-ligand interaction plot for SANC239 in chain B, showing the intermolecular interaction at different stages of the 100ns MD simulation. A plot of co-crystallised a COX-1-aspirin complex (PDB ID:5F1A) is used as the reference for the active site pocket and important residues.

Network analysis While presence of the SANC239 minimally changed residue interaction network for both chains as characterised by Figure 5.25, *average BC* was higher in chain B in the loop regions of residues 120-160, changes in *average BC* and *L* are not immensely different.

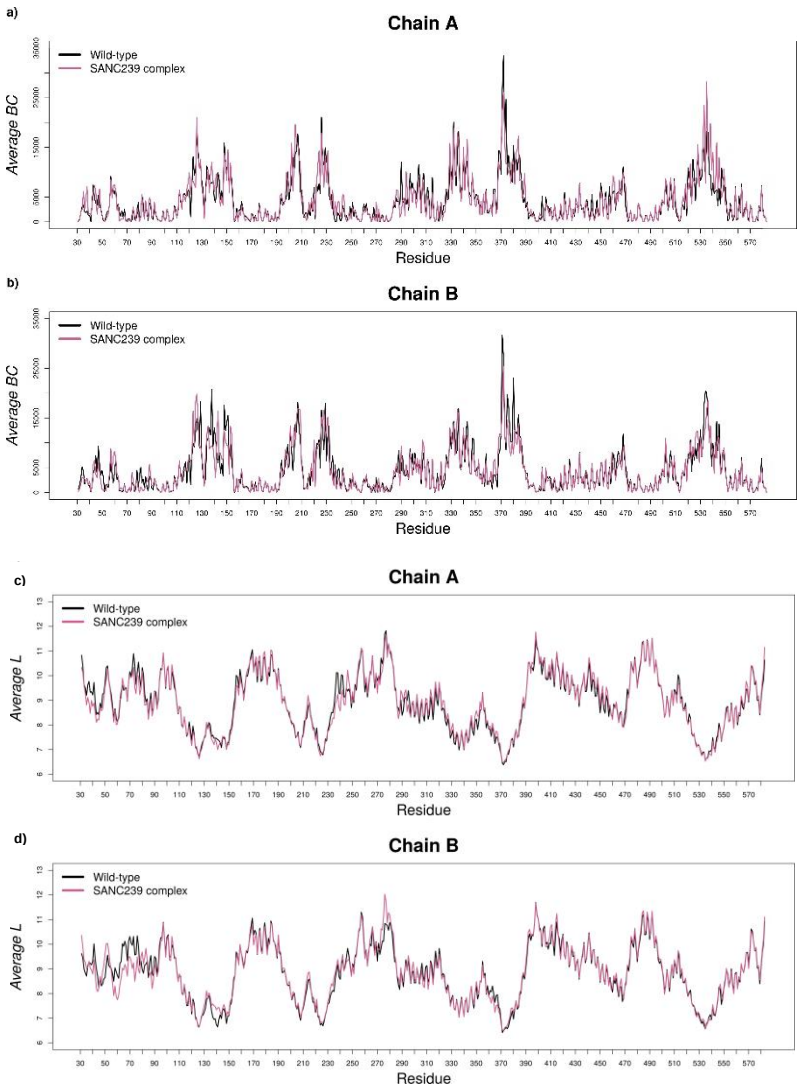


Figure 5.25: *Average BC* (a-b) and *L* (c-d) of SANC239 bound COX-1, relative to the apo-protein.

MM-PBSA calculation Residues Leu-351 and Phe-517, which shared VDW interactions with the ligand [Figure 5.24] in the active site pocket, contributed the highest to the free binding energy of the ligand. A significant number of residues in the active site, interacting with the ligand, contributed favourably to overall free binding energy. As shown in Figure 5.26, Trp-386, the electron donor in a hydrogen bond with the ligand; Gly-525 and Ala-526 participate in the interaction electrostatically, and Leu-383 which is a hydrogen donor from the main chain [S.Figure 11], all had considerable contributions.

Studies have shown other aspirin-like compounds are valuable in the inhibition of COX-enzymes [190]. SANC239 fits this bill, with is relatively small molecular weight

[shown in Table 4.3] comparable to that of aspirin and its interactions with notable residues of the aspirin active-site pocket.

Table 5.2: A decomposition of the binding energy components obtained from MM-PBSA conducted for SANC239.

Ligand	Energy (kJ/mol)				
	VDW	Electrostatic	Polar solvation	SASA	Binding
SANC239_A	-81.364 ± 11.341	-208.742 ± 19.426	178.669 ± 19.302	-11.073 ± 0.646	-122.509 ± 16.088
SANC239_B	-120.014 ± 9.685	-49.920 ± 7.296	104.718 ± 4.619	-11.031 ± 0.580	-76.247 ± 8.394

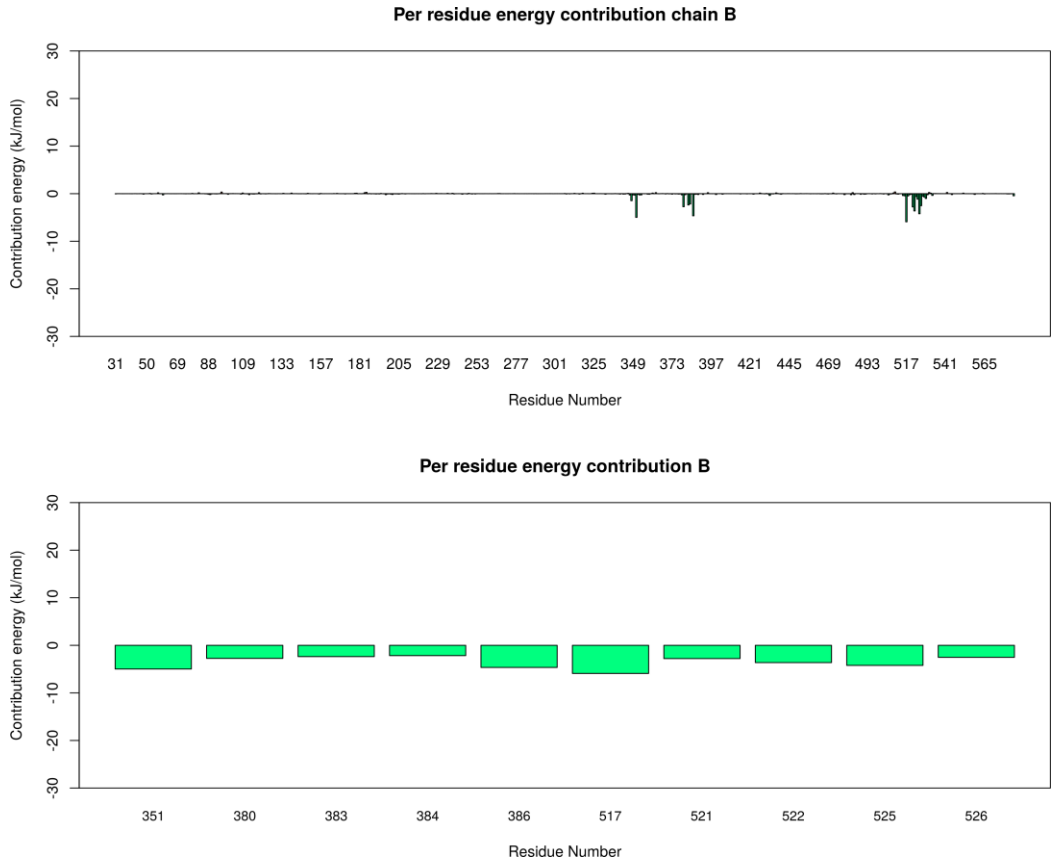


Figure 5.26: SANC239 MM-PBSA per residue energy contribution in chain B, highlighting residues with the highest contributions.

5.3.3.2 ZINC4671

The protein-ligand interactions of ligand ZINC 4671 in chain A were analysed, due to the better ligand stability exhibited in chain A as opposed to chain B [Figure 5.23 & S.Figure 10].

Visual analysis Visual analysis of protein-ligand interactions in Figure 5.27, showed constant hydrogen bonding with electron-donor Ser-529. The ligand seemed to dislodge from active site at 10ns, but re-established interactions by the 100ns mark. This behaviour was additionally corroborated by the slight deviation observed around the 10ns in the RMSD [(Figure 5.23 & S.Figure 10], after which the ligand remained stable.

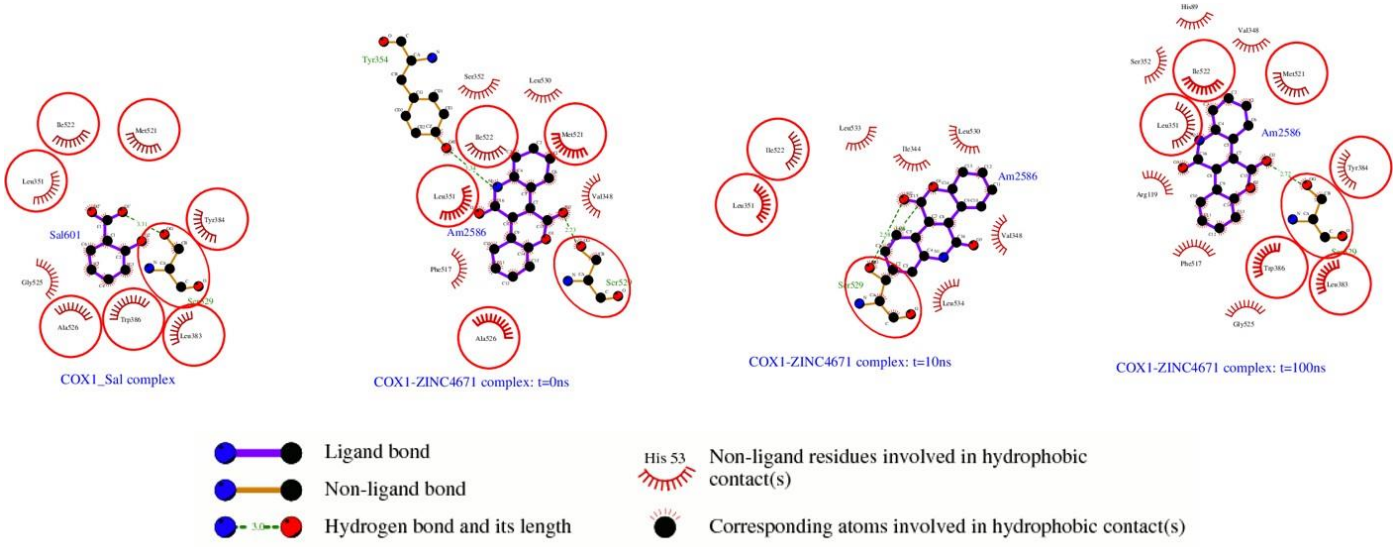


Figure 5.27: LigPlot protein-ligand interaction plot for ZINC461 in chain A, showing the intermolecular interaction at different stages of the 100ns MD simulation. A plot of co-crystallised a COX-1-aspirin complex (PDB ID: 5F1A) is used as the reference for the active site pocket and important residues.

Network analysis Network analysis of the protein in Fig.5.28, showed an increase in average BC between residue 135 and 150 in the catalytic domain. The increase was not mirrored in chain B which exhibits unfavourable interactions with the ligand. It was therefore highly likely the differences in BC were due to ligand presence.

MM-PBSA calculation The per residue contributions from free binding energy analyses, of ZINC 4671 conducted using MM-PBSA, showed that Arg-119 contributed unfavourably to the total ΔG.

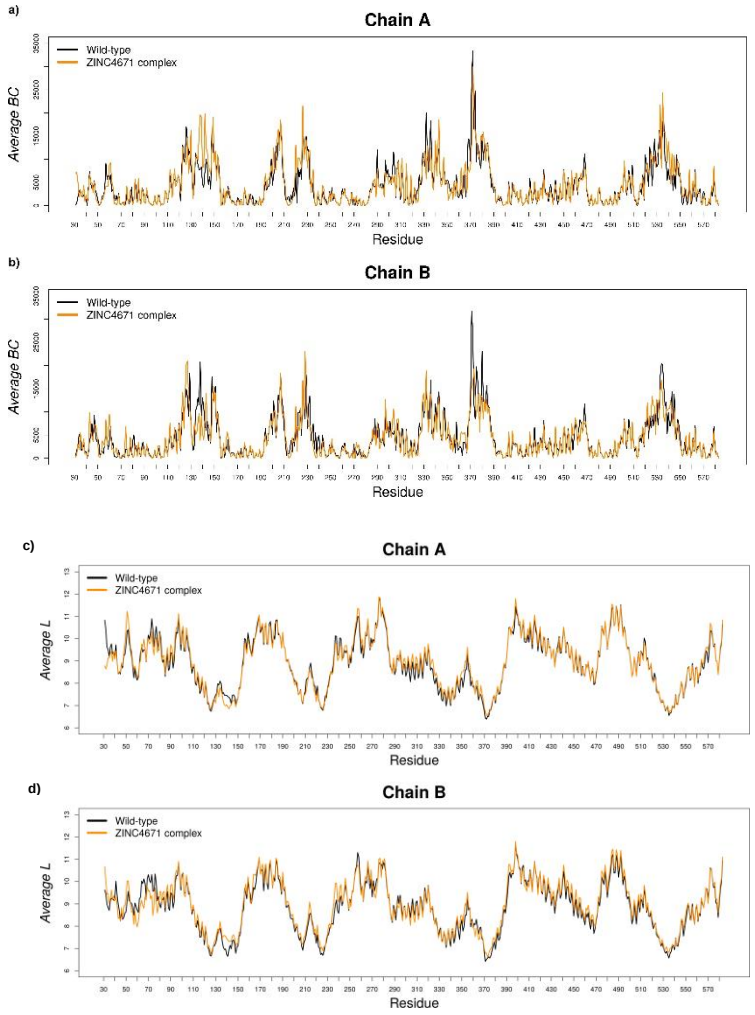


Figure 5.28: Average BC (a-b) and L (c-d) of ZINC4671 bound COX-1, relative to the apo-protein.

Arg-119 plays a key role in stabilising substrate, AA, in the active site. Though it does not appear in the LigPlot diagram, Arg-119 participates in a π-π stack with the ligand, as shown in S.Figure 11. Ile-522 and Leu-351, which are among the residues that contribute favourably to total energy, can be seen interacting electrostatically with the ligand. Leu-351, with the second largest contribution, participates in a residue ligand π-σ interaction [S.Figure 11].

Table 5.3: A decomposition of the binding energy components obtained from MM-PBSA conducted for ZINC4671.

Ligand	Energy (kJ/mol)				
	VDW	Electrostatic	Polar solvation	SASA	Binding
ZINC4671_A	-151.560 ± 0.265	-23.66 ± 0.205	99.210 ± 0.239	-14.871 ± 0.026	-90.849 ± 0.284
ZINC4671_B	-145.302 ± 0.299	-29.804 ± 0.394	93.035 ± 0.741	-11.031 ± 0.030	-95.732 ± 0.522

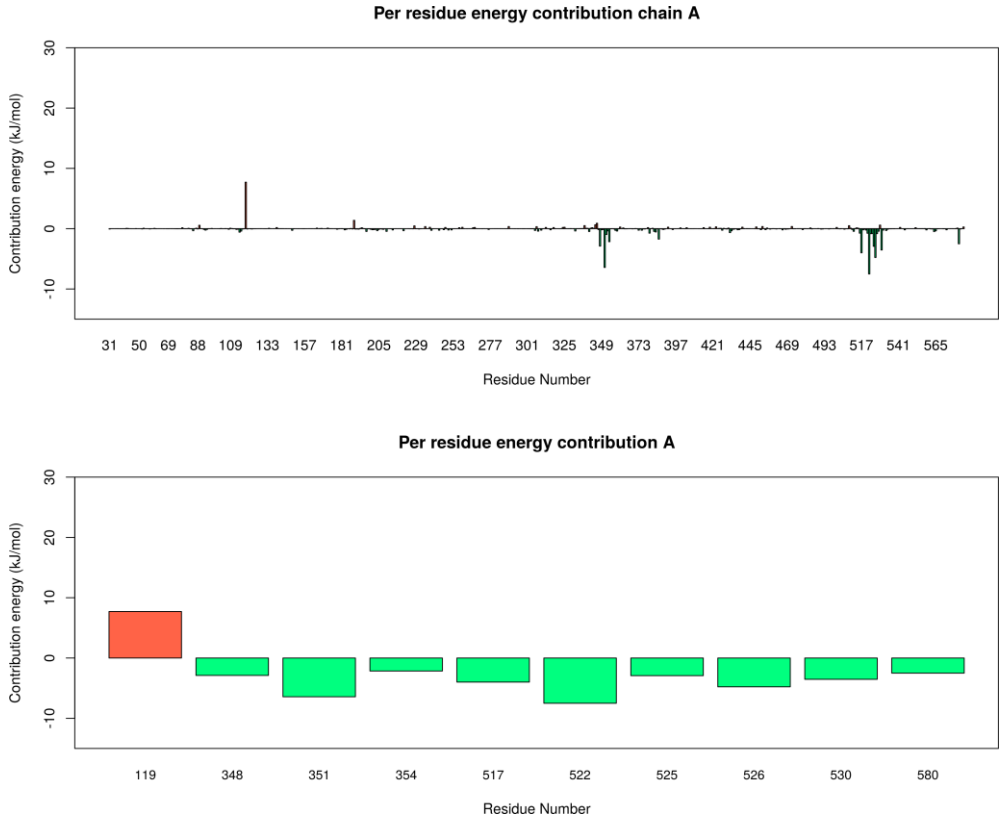


Figure 5.29: ZINC4671 MM-PBSA per residue energy contribution in chain A, highlighting residues with the highest contributions.

5.4 Conclusion

The objective of this chapter was to perform MD simulations on the COX-1 in three different states, a) as the wild-type, b) as structural variants due to SNPs and c) as a ligand-bound protein.

MD simulations conducted on the COX-1 wild-type were performed to provide a better understanding of protein's behaviour. The MD simulations revealed that the COX-1 exhibited dissimilar behaviour during the simulations, inferring that the enzyme is in fact a functional heterodimer. Dimer asymmetry has been cited to be important to catalytic activity of some enzymes [245].

MD simulations conducted on the COX-1 variants showed that presence of SNPs altered protein dynamics to varying degrees. I557T exhibited the largest change in global motion, despite its low evolutionary conservation and *in silico* prediction score. N143K and L237M showed the greatest change in DRN produced contact maps, affecting inter-monomer interactions. Based on the suggested importance of communication between monomers, this may affect protein catalytic function. P126T exhibited intermediate change, relative to the wild-type. Probable effect of the SNPs on enzyme interaction with ligands or potential drugs catalytic cannot be confirmed based on the trajectory analyses.

The simulations conducted on the protein-ligand complexes had the purpose of assessing ligand stability in screening for a potential drug compound. Due to being a functional heterodimer, the COX-1 monomers interacted with ligands in a dissimilar manner, as seen in the MD simulations on the protein ligand complexes. Ligands SANC239 and ZINC4671 showed reasonable stability over the course of the MD simulations and interacted with key residues in the cyclooxygenase active site. Both specifically bound in the aspirin pocket, though SANC 239 exhibited more hydrogen bonding. It is assumed SANC239 bound more effectively than ZINC4671 due to its smaller size relative to the other. The unique behaviour of ligands across the monomers further suggested that allostery can be switched between the two. Overall, the two ligands seem to be viable inhibitors that could benefit from further investigation in a wet-lab study.

6 Conclusion

6.1 Concluding Remarks

The research aims of this study were to identify potential drug compounds to inhibit COX-1 and analyse the effect of nsSNPs on structural variants of the protein associated with them. The experimental work outlined in this thesis investigates COX-1 in these two aspects.

6.1.1 SNP analysis

In the SNP analysis component of this study, the wild-type motions and conformations were studied as a reference for variant behaviour. The findings of the analysis corroborated previous observations, that the COX-1 is a sequence homodimer, but conformational heterodimer [] [] [].

Analysis of the structural variants of COX-1 revealed an overall effect on the global motions of the protein due to presence of the SNPs.I557T exhibited the largest changes in global motion, while N143K and L237M showed the biggest effect to the residue interaction network. Interaction across the dimer interface, was also affected by SNPs N143K and L237M. While the importance of communications is yet to be fully investigated [41], it is suggested to be vital to the catalytic activity of the protein, conferring allosteric and catalytic activity to the monomers [,].

As such, investigation into the SNPs revealed their probable effect on the structure and probably catalytic function of the protein. Functional effects of these SNPs can be further investigated and validated in wet-lab experimentation through series of functional assays.

6.1.2 Drug Discovery

In the drug discovery portion of the research, potential drug compounds that bound favourably to the COX-1 monomers were identified. However, due to COX-1 being conformational heterodimer, each monomer interacted with ligands uniquely, resulting in varying results.

Most promising ligands interacted with Ser-529, the site of aspirin acetylation in COX-1. Most promising ligands met at least three of the Lipinski rules for druggability. Aspirin- like natural product ligands, such as SANC 239, bound well and in a stable manner with COX-1, and is likely to be good inhibitor of COX-1. ZINC4671 also showed stability in the aspirin active site pocket. These observations could be further validated experimentally using NMR analysis and point-mutation studies and follow up assays.

6.2 Limitations of study

The primary limitation to this study was the selection of SNP/SNVs studied. As highlighted in Chapter 2, the criteria used in the study prioritised location of amino acid substitution in protein over variant frequency in population [60] and, as a result, some of the substitutions [Table 2.1] can be better categorised as private variants.

The consequence of this selection is the non-viability of the targets for large-scale pharmacogenomics and precision medicine. The method used however, did reveal probable effects in those small populations carrying these variants and is transferable to SNPs of higher frequency.

6.3 Recommendations and further work

Further *in silico* analysis into effect of SNPs on ligand binding and interaction could shed light on the effect of the polymorphisms on protein catalytic activity. Of great interest is L237M, which has been cited and investigated for playing a role in aspirin resistance, especially in Asian populations.

Additionally, the use of tailor-made ligands through scaffold-hopping [246] and fragment linking [247] could be employed to improve *in silico* drug design efforts.

References

[1] W. H. Organization, World health statistics 2016: monitoring health for the SDGs sustainable development goals, World Health Organization, 2016.

[2] D. Mozaffarian, E. J. Benjamin, A. S. Go, D. K. Arnett, M. J. Blaha, M. Cushman, S. R. Das, S. Ferranti, J.-P. Després, H. J. Fullerton, V. J. Howard, M. D. Huffman, C. R. Isasi, M. C. Jiménez, S. E. Judd, B. M. Kissela, J. H. Lichtman, L. D. Lisabeth, S. Liu, R. H. Mackey, D. J. Magid, D. K. McGuire, E. R. Mohler, C. S. Moy, P. Muntner, M. E. Mussolino, K. Nasir, R. W. Neumar, G. Nichol, L. Palaniappan, D. K. Pandey, M. J. Reeves, C. J. Rodriguez, W. Rosamond, P. D. Sorlie, J. Stein, A. Towfighi, T. N. Turan, S. S. Virani, D. Woo, R. W. Yeh and M. B. Turner, "Heart Disease and Stroke Statistics—2016 Update," *Circulation,* 2015.

[3] S. C. Johnston, S. Mendis and C. D. Mathers, "Global variation in stroke burden and mortality: estimates from monitoring, surveillance, and modelling," *The Lancet Neurology,* vol. 8, pp. 345-354, 2009.

[4] K. Strong, C. Mathers and R. Bonita, "Preventing stroke: saving lives around the world," *The Lancet Neurology,* vol. 6, pp. 182-187, 2007.

[5] V. L. Feigin, B. Norrving and G. A. Mensah, "Global burden of stroke," *Circulation research,* vol. 120, pp. 439-448, 2017.

[6] M. J. O'Donnell, D. Xavier, L. Liu, H. Zhang, S. L. Chin, P. Rao-Melacini, S. Rangarajan, S. Islam, P. Pais, M. J. McQueen and others, "Risk factors for ischaemic and intracerebral haemorrhagic stroke in 22 countries (the INTERSTROKE study): a case-control study," *The Lancet,* vol. 376, pp. 112-123, 2010.

[7] R. O. Akinyemi, B. Ovbiagele, M. Gebreziabher, S. Warth, D. Lackland, A. Akpalu, K. Sagoe, L. Owolabi, F. Sarfo, R. Obiako and others, "Stroke genomics in people of African ancestry: charting new paths," *Cardiovascular journal of Africa,* vol. 26, p. S39, 2015.

[8] R. O. Akinyemi, I. M. H. Izzeldin, C. Dotchin, W. K. Gray, O. Adeniji, O. A. Seidi, J. J. Mwakisambwe, C. J. Mhina, F. Mutesi, H. Z. Msechu and others, "Contribution of Noncommunicable Diseases to Medical Admissions of Elderly Adults in Africa: A Prospective, Cross-Sectional Study in Nigeria, Sudan, and Tanzania," *Journal of the American Geriatrics Society,* vol. 62, pp. 1460-1466, 2014.

[9] M. O. Owolabi, O. Arulogun, S. Melikam, A. M. Adeoye, S. Akarolo-Anthony, R. Akinyemi, D. Arnett, H. Tiwari, M. Gebregziabher, C. Jenkins and others, "The burden of stroke in Africa: a glance at the present and a glimpse into the future," *Cardiovascular journal of Africa,* vol. 26, p. S27, 2015.

[10] M. D. Connor, R. Walker, G. Modi and C. P. Warlow, "Burden of stroke in black populations in sub-Saharan Africa," *The Lancet Neurology,* vol. 6, pp. 269-278, 2007.

[11] W. R. I. T. I. N. G. G. R. O. U. P. MEMBERS, E. J. Benjamin, M. J. Blaha, S. E. Chiuve, M. Cushman, S. R. Das, R. Deo, S. D. Ferranti, J. Floyd, M. Fornage and others, "Heart disease and stroke statistics—2017 update: a report from the American Heart Association," *Circulation,* vol. 135, p. e146, 2017.

[12] M. J. O'Donnell, S. L. Chin, S. Rangarajan, D. Xavier, L. Liu, H. Zhang, P. Rao-Melacini, X. Zhang, P. Pais, S. Agapay and others, "Global and regional effects of potentially modifiable risk factors associated with acute stroke in 32 countries (INTERSTROKE): a case-control study,"

The Lancet, vol. 388, pp. 761-775, 2016.

[13] C. L. Allen and U. Bayraktutan, "Risk Factors for Ischaemic Stroke," *International Journal of Stroke,* vol. 3, pp. 105-116, 2008.

[14] S. Vidale and E. Agostoni, "Endovascular treatment of ischemic stroke: an updated meta-analysis of efficacy and safety," *Vascular and endovascular surgery,* vol. 51, pp. 215-219, 2017.

[15] R. A. G. Patel and P. W. McMullen, "Neuroprotection in the treatment of acute ischemic stroke," *Progress in cardiovascular diseases,* vol. 59, pp. 542-548, 2017.

[16] E. C. Jauch, J. L. Saver, H. P. Adams, A. Bruno, B. M. Demaerschalk, P. Khatri, P. W. McMullan, A. I. Qureshi, K. Rosenfield, P. A. Scott and others, "Guidelines for the early management of patients with acute ischemic stroke: a guideline for healthcare professionals from the American Heart Association/American Stroke Association," *Stroke,* vol. 44, pp. 870-947, 2013.

[17] M. D. Ginsberg, "Neuroprotection for ischemic stroke: past, present and future," *Neuropharmacology,* vol. 55, pp. 363-389, 2008.

[18] R. L. Sacco, J. Y. Chong, S. Prabhakaran and M. S. V. Elkind, "Experimental treatments for acute ischaemic stroke," *The Lancet,* vol. 369, pp. 331-341, 2007.

[19] E. H. Lo, T. Dalkara and M. A. Moskowitz, "Neurological diseases: Mechanisms, challenges and opportunities in stroke," *Nature reviews neuroscience,* vol. 4, p. 399, 2003.

[20] N. Patel, M. B. Lanktree and R. A. Hegele, "Genetic risk factors for stroke in the genome-wide association era," *Expert opinion on medical diagnostics,* vol. 5, pp. 75-84, 2011.

[21] R. A. Hegele and M. Dichgans, "Advances in stroke 2009: update on the genetics of stroke and cerebrovascular disease 2009," *Stroke,* vol. 41, pp. e63--e66, 2010.

[22] M. Rask-Andersen, M. S. Almén and H. B. Schiöth, "Trends in the exploitation of novel drug targets," *Nature reviews Drug discovery,* vol. 10, p. 579, 2011.

[23] M. Huang, G. Cheng, H. Tan, R. Qin, Y. Zou, Y. Wang and Y. Zhang, "Capsaicin protects cortical neurons against ischemia/reperfusion injury via down-regulating NMDA receptors," *Experimental neurology,* vol. 295, pp. 66-76, 2017.

[24] J.-Q. Liu, S.-X. Dai, J.-J. Zheng, Y.-C. Guo, W.-X. Li, G.-H. Li and J.-F. Huang, "The identification and molecular mechanism of anti-stroke traditional Chinese medicinal compounds," *Scientific reports,* vol. 7, p. 41406, 2017.

[25] J. A. Saugstad, "Non-coding RNAs in stroke and neuroprotection," *Frontiers in neurology,* vol. 6, p. 50, 2015.

[26] W. L. Smith and D. L. Dewitt, "Prostaglandin endoperoxide H synthases-1 and-2," in *Advances in immunology*, vol. 62, Elsevier, 1996, pp. 167-215.

[27] D. Picot, P. J. Loll and R. M. Garavito, "The X-ray crystal structure of the membrane protein prostaglandin H2 synthase-1," *Nature,* vol. 367, p. 243, 1994.

[28] K. Gupta, B. S. Selinsky, C. J. Kaub, A. K. Katz and P. J. Loll, "The 2.0 Å resolution crystal structure of prostaglandin H2 synthase-1: structural insights into an unusual peroxidase," *Journal of molecular biology,* vol. 335, pp. 503-518, 2004.

[29] R. M. Garavito and A. M. Mulichak, "The structure of mammalian cyclooxygenases," *Annual review of biophysics and biomolecular structure,* vol. 32, pp. 183-206, 2003.

[30] P. W. Fowler and P. V. Coveney, "A computational protocol for the integration of the monotopic protein prostaglandin H2 synthase into a phospholipid bilayer," *Biophysical journal,* vol. 91, pp. 401-410, 2006.

[31] L. Toth, L. Muszbek and I. Komaromi, "Mechanism of the irreversible inhibition of human cyclooxygenase-1 by aspirin as predicted by QM/MM calculations," *Journal of Molecular Graphics and Modelling,* vol. 40, pp. 99-109, 2013.

[32] R. G. Kurumbail, A. M. Stevens, J. K. Gierse, J. J. McDonald, R. A. Stegeman, J. Y. Pak, D. Gildehaus, T. D. Penning, K. Seibert, P. C. Isakson and others, "Structural basis for selective inhibition of cyclo-oxygenase-2 by anti-inflammatory agents," *Nature,* vol. 384, p. 644, 1996.

[33] M. L. Plount Price and W. L. Jorgensen, "Analysis of binding affinities for celecoxib analogues with COX-1 and COX-2 from combined docking and Monte Carlo simulations and insight into the COX-2/COX-1 selectivity," *Journal of the American Chemical Society,* vol. 122, pp. 9455-9466, 2000.

[34] L. Dong, N. P. Sharma, B. J. Jurban and W. L. Smith, "Preexistent Asymmetry in the Human Cyclooxygenase-2 Sequence Homodimer," *Jou-rnal of Biological Chemistry,* pp. jbc--M113, 2013.

[35] N. P. Sharma, L. Dong, C. Yuan, K. R. Noon and W. L. Smith, "Asymmetric Acetylation of the Cyclooxygenase-2 Homodimer by Aspirin and Its Effects on the Oxygenation of Arachidonic, Eicosapentaenoic and Docosahexaenoic Acids," *Molecular pharmacology,* pp. mol--109, 2010.

[36] R. S. Sidhu, J. Y. Lee, C. Yuan and W. L. Smith, "Comparison of cyclooxygenase-1 crystal structures: cross-talk between monomers comprising cyclooxygenase-1 homodimers," *Biochemistry,* vol. 49, pp. 7069-7079, 2010.

[37] R. J. Kulmacz and W. E. Lands, "Prostaglandin H synthase. Stoichiometry of heme cofactor.," *Journal of Biological Chemistry,* vol. 259, pp. 6358-6363, 1984.

[38] R. J. Kulmacz and W. E. Lands, "Stoichiometry and kinetics of the interaction of prostaglandin H synthase with anti-inflammatory agents.," *Journal of Biological Chemistry,* vol. 260, pp. 12572-12578, 1985.

[39] C. Yuan, C. J. Rieke, G. Rimon, B. A. Wingerd and W. L. Smith, "Partnering between monomers of cyclooxygenase-2 homodimers," *Proceedings of the National Academy of Sciences,* vol. 103, pp. 6142-6147, 2006.

[40] H. Zou, C. Yuan, L. Dong, R. S. Sidhu, Y. H. Hong, D. V. Kuklev and W. L. Smith, "Human cyclooxygenase-1 activity and its responses to COX inhibitors are allosterically regulated by non-substrate fatty acids," *Journal of lipid research,* pp. jlr--M026856, 2012.

[41] C. Yuan, R. S. Sidhu, D. V. Kuklev, Y. Kado, M. Wada, I. Song and W. L. Smith, "Cyclooxygenase allosterism: fatty acid mediated cross-talk between monomers of cyclooxygenase homodimers," *Journal of Biological Chemistry,* 2009.

[42] G. Rimon, R. S. Sidhu, D. A. Lauver, J. Y. Lee, N. P. Sharma, C. Yuan, R. A. Frieler, R. C. Trievel, B. R. Lucchesi and W. L. Smith, "Coxibs interfere with the action of aspirin by binding tightly to one monomer of cyclooxygenase-1," *Proceedings of the National Academy of Sciences,* vol. 107, pp. 28-33, 2010.

[43] S. D. Skaper, "The brain as a target for inflammatory processes and neuroprotective strategies," *Annals of the New York Academy of Sciences,* vol. 1122, pp. 23-34, 2007.

[44] H. Lin, T.-N. Lin, W.-M. Cheung, G.-M. Nian, P.-H. Tseng, S.-F. Chen, J.-J. Chen, S.-K. Shyue, J.-Y. Liou, C.-W. Wu and others, "Cyclooxygenase-1 and bicistronic cyclooxygenase-1/ prostacyclin synthase gene transfer protect against ischemic cerebral infarction," *Circulation,* vol. 105, pp. 1962-1969, 2002.

[45] J. W. Phillis, L. A. Horrocks and A. A. Farooqui, "Cyclooxygenases, lipoxygenases, and epoxygenases in CNS: their role and involvement in neurological disorders," *Brain research reviews,* vol. 52, pp. 201-243, 2006.

[46] W. J. Streit, R. E. Mrak and W. S. T. Griffin, "Microglia and neuroinflammation: a pathological perspective," *Journal of neuroinflammation,* vol. 1, p. 14, 2004.

[47] E. Candelario-Jalil, A. C. P. Oliveira, S. Gräf, H. S. Bhatia, M. Hüll, E. Muñoz and B. L. Fiebich, "Resveratrol potently reduces prostaglandin E 2 production and free radical formation in lipopolysaccharide-activated primary rat microglia," *Journal of neuroinflammation,* vol. 4, p. 25,

2007.

[48] S. Aïd and F. Bosetti, "Gene expression of cyclooxygenase-1 and Ca2+-independent phospholipase A2 is altered in rat hippocampus during normal aging," *Brain research bulletin,* vol. 73, pp. 108-113, 2007.

[49] J. M. Schwab, T. D. Nguyen, E. Postler, R. Meyermann and H. J. Schluesener, "Selective accumulation of cyclooxygenase-1-expressing microglial cells/ macrophages in lesions of human focal cerebral ischemia," *Acta neuropathologica,* vol. 99, pp. 609-614, 2000.

[50] S.-H. Choi, S. Aid and F. Bosetti, "The distinct roles of cyclooxygenase-1 and-2 in neuroinflammation: implications for translational research," *Trends in pharmacological sciences,* vol. 30, pp. 174-181, 2009.

[51] F. Yi, L. Sun, L.-j. Xu, Y. Peng, H.-b. Liu, C.-n. He and P.-g. Xiao, "In silico approach for anti-thrombosis drug discovery: P2Y1R structure-based TCMs screening," *Frontiers in pharmacology,* vol. 7, p. 531, 2017.

[52] C. Reis, O. Akyol, W. M. Ho, C. Araujo, L. Huang, I. I. Applegate, J. H. Zhang and others, "Phase I and phase II therapies for acute ischemic stroke: an update on currently studied drugs in clinical research," *BioMed research international,* vol. 2017, 2017.

[53] Z. Zhou, J. Lu, W.-W. Liu, A. Manaenko, X. Hou, Q. Mei, J.-L. Huang, J. Tang, J. H. Zhang, H. Yao and others, "Advances in stroke pharmacology," *Pharmacology & therapeutics,* 2018.

[54] A. Pannunzio and M. Coluccia, "Cyclooxygenase-1 (COX-1) and COX-1 inhibitors in cancer: a review of oncology and medicinal chemistry literature," *Pharmaceuticals,* vol. 11, p. 101, 2018.

[55] A. D. Malvezi, C. Panis, R. V. Silva, R. C. Freitas, M. I. L. Martins, V. L. H. Tatakihara, N. G. Zanluqui, E. C. Neto, S. Goldenberg, J. Bordignon and others, "Inhibition of cyclooxygenase-1 and cyclooxygenase-2 impairs Trypanosoma cruzi entry in cardiac cell and promotes differential modulation of inflammatory response," *Antimicrobial agents and chemotherapy,* pp. AAC--02752, 2014.

[56] K. A. Babaheydari, "In Silico Drug Design on Aspirin for Cyclooxygenase I and II, Target for Reduce the Effects of Inflammatory," *Biosciences Biotechnology Research Asia,* vol. 12, no. 1, pp. 433-444, 6 2015.

[57] L. Dewi, "In Silico Analysis of the Potential of the Active Compounds Fucoidan and Alginate Derived from Sargassum Sp. as Inhibitors of COX-1 and COX-2," *Medical Archives,* vol. 70, p. 172, 2016.

[58] M. J. H. L. Mulder, I. G. H. Jansen, R.-J. B. Goldhoorn, E. Venema, V. Chalos, K. C. J. Compagne, B. Roozenbeek, H. F. Lingsma, W. J. Schonewille, I. R. Wijngaard and others, "Time to endovascular treatment and outcome in acute ischemic stroke: MR CLEAN registry results," *Circulation,* pp. CIRCULATIONAHA--117, 2018.

[59] J. T. Mendell and H. C. Dietz, "When the message goes awry: disease-producing mutations that influence mRNA content and performance," *Cell,* vol. 107, pp. 411-414, 2001.

[60] A. Riva and I. S. Kohane, "SNPper: retrieval and analysis of human SNPs," *Bioinformatics,* vol. 18, pp. 1681-1685, 2002.

[61] F. S. Collins, L. D. Brooks and A. Chakravarti, "A DNA polymorphism discovery resource for research on human genetic variation," *Genome research,* vol. 8, pp. 1229-1231, 1998.

[62] A. Uzun, C. M. Leslin, A. Abyzov and V. Ilyin, "Structure SNP (StSNP): a web server for mapping and modeling nsSNPs on protein structures with linkage to metabolic pathways," *Nucleic acids research,* vol. 35, pp. W384--W392, 2007.

[63] D. Chasman and R. M. Adams, "Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation," *Journal of molecular biology,* vol. 307, pp. 683-706, 2001.

[64] J. L. Lahti, G. W. Tang, E. Capriotti, T. Liu and R. B. Altman, "Bioinformatics and variability in drug response: a protein structural perspective," *Journal of The Royal Society Interface,* vol. 9, pp. 1409-1437, 2012.

[65] N. H. Lee, "Pharmacogenetics of drug metabolizing enzymes and transporters: effects on pharmacokinetics and pharmacodynamics of anticancer agents," *Anti-Cancer Agents in Medicinal Chemistry (Formerly Current Medicinal Chemistry-Anti-Cancer Agents),* vol. 10, pp. 583-592, 2010.

[66] N. Nagasundaram, H. Zhu, J. Liu, V. Karthick, C. Chakraborty, L. Chen and others, "Analysing the effect of mutation on protein function and discovering potential inhibitors of CDK4: molecular modelling and dynamics studies," *PLoS One,* vol. 10, p. e0133969, 2015.

[67] A. Hillisch and R. Hilgenfeld, "The role of protein 3D-structures in the drug discovery process," in *Modern methods of drug discovery*, Springer, 2003, pp. 157-181.

[68] E. Sitbon and S. Pietrokovski, "Occurrence of protein structure elements in conserved sequence regions," *BMC structural biology,* vol. 7, p. 3, 2007.

[69] A. K. Mitra, S. V. Singh, V. K. Garg, M. Sharma, R. Chaturvedi and S. K. Rath, "Protective association exhibited by the single nucleotide polymorphism (SNP) rs1052133 in the gene human 8- oxoguanine DNA glycosylase (hOGG1) with the risk of squamous cell carcinomas of the head & neck (SCCHN) among north Indians," *The Indian journal of medical research,* vol. 133, p. 605, 2011.

[70] P. M. Visscher, N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy, M. A. Brown and J. Yang, "10 years of GWAS discovery: biology, function, and translation," *The American Journal of Human Genetics,* vol. 101, pp. 5-22, 2017.

[71] M. Cargill, D. Altshuler, J. Ireland, P. Sklar, K. Ardlie, N. Patil, C. R. Lane, E. P. Lim, N. Kalyanaraman, J. Nemesh and others, "Characterization of single-nucleotide polymorphisms in coding regions of human genes," *Nature genetics,* vol. 22, p. 231, 1999.

[72] L. Bao and Y. Cui, "Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information," *Bioinformatics,* vol. 21, pp. 2185-2190, 2005.

[73] P. Yue and J. Moult, "Identification and analysis of deleterious human SNPs," *Journal of molecular biology,* vol. 356, pp. 1263-1274, 2006.

[74] N.-L. Sim, P. Kumar, J. Hu, S. Henikoff, G. Schneider and P. C. Ng, "SIFT web server: predicting effects of amino acid substitutions on proteins," *Nucleic acids research,* vol. 40, pp. W452--W457, 2012.

[75] Y. Bromberg and B. Rost, "SNAP: predict effect of non-synonymous polymorphisms on function," *Nucleic acids research,* vol. 35, pp. 3823-3835, 2007.

[76] B. Li, V. G. Krishnan, M. E. Mort, F. Xin, K. K. Kamati, D. N. Cooper, S. D. Mooney and P. Radivojac, "Automated inference of molecular mechanisms of disease from amino acid substitutions," *Bioinformatics,* vol. 25, pp. 2744-2750, 2009.

[77] H. Mi, X. Huang, A. Muruganujan, H. Tang, C. Mills, D. Kang and P. D. Thomas, "PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements," *Nucleic acids research,* vol. 45, pp. D183--D189, 2016.

[78] S. Yin, F. Ding and N. V. Dokholyan, "Eris: an automated estimator of protein stability," *Nature methods,* vol. 4, p. 466, 2007.

[79] V. Parthiban, M. M. Gromiha and D. Schomburg, "CUPSAT: prediction of protein stability upon point mutations," *Nucleic acids research,* vol. 34, pp. W239--W242, 2006.

[80] Y. Choi and A. P. Chan, "PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels," *Bioinformatics,* vol. 31, pp. 2745-2747, 2015.

[81] I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov and S. R. Sunyaev, "A method and server for predicting damaging missense mutations," *Nature methods,* vol. 7, p. 248, 2010.

[82] J. Cheng, A. Randall and P. Baldi, "Prediction of protein stability changes for single-site mutations using support vector machines," *Proteins: Structure, Function, and Bioinformatics,* vol. 62, pp. 1125-1132, 2006.

[83] M. A. Care, C. J. Needham, A. J. Bulpitt and D. R. Westhead, "Deleterious SNP prediction: be mindful of your training data!," *Bioinformatics,* vol. 23, pp. 664-672, 2007.

[84] D. K. Brown and Ö. T. Bishop, "Role of structural bioinformatics in drug discovery by computational SNP analysis: analyzing variation at the protein level," *Global heart,* vol. 12, pp. 151-161, 2017.

[85] J. Bendl, J. Stourac, O. Salanda, A. Pavelka, E. D. Wieben, J. Zendulka, J. Brezovsky and J. Damborsky, "PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations," *PLoS computational biology,* vol. 10, p. e1003440, 2014.

[86] C. M. Ulrich, J. Bigler, J. Sibert, E. A. Greene, R. Sparks, C. S. Carlson and J. D. Potter, "Cyclooxygenase 1 (COX1) polymorphisms in African-American and Caucasian populations," *Human mutation,* vol. 20, pp. 409-410, 2002.

[87] J. A. G. Agúndez, M. Blanca, J. A. Cornejo-García and E. García-Martín, "Pharmacogenomics of cyclooxygenases," *Pharmacogenomics,* vol. 16, pp. 501-522, 2015.

[88] M. K. Halushka, L. P. Walker and P. V. Halushka, "Genetic variation in cyclooxygenase 1: effects on response to aspirin," *Clinical Pharmacology & Therapeutics,* vol. 73, pp. 122-130, 2003.

[89] J. Shi, N. L. A. Misso, D. L. Duffy, B. Bradley, R. Beard, P. J. Thompson and M. A. Kedda, "Cyclooxygenase-1 gene polymorphisms in patients with different asthma phenotypes and atopy," *European Respiratory Journal,* vol. 26, pp. 249-256, 2005.

[90] T. Arisawa, T. Tahara, T. Shibata, M. Nagasaka, M. Nakamura, Y. Kamiya, H. Fujita, D. Yoshioka, Y. Arima, M. Okubo and others, "Genetic polymorphisms of cyclooxygenase-1 (COX-1) are associated with functional dyspepsia in Japanese women," *Journal of women's health,* vol. 17, pp. 1039-1043, 2008.

[91] N. V. Chandrasekharan and D. L. Simmons, "The cyclooxygenases," *Genome biology,* vol. 5, p. 241, 2004.

[92] J. A. G. Agundez, C. Martínez, D. Pérez-Sala, M. Carballo, M. J. Torres and E. García-Martín, "Pharmacogenomics in aspirin intolerance," *Current drug metabolism,* vol. 10, pp. 998-1008, 2009.

[93] C. R. Lee, F. G. Bottone Jr, J. M. Krahn, L. Li, H. W. Mohrenweiser, M. E. Cook, R. M. Petrovich, D. A. Bell, T. E. Eling and D. C. Zeldin, "Identification and functional characterization of polymorphisms in human cyclooxygenase-1 (PTGS1)," *Pharmacogenetics and genomics,* vol. 17, p. 145, 2007.

[94] C. R. Lee, K. E. North, M. S. Bray, D. J. Couper, G. Heiss and D. C. Zeldin, "Cyclooxygenase polymorphisms and risk of cardiovascular events: the Atherosclerosis Risk in Communities (ARIC) study," *Clinical Pharmacology & Therapeutics,* vol. 83, pp. 52-60, 2008.

[95] X. L. Li, J. Cao, L. Fan, L. Ye, Q. Wang, C. P. Cui, L. Liu and F. C. Zhou, "Correlation analysis of aspirin resistance and cycloxygenase-1 haplotype in old Chinese patients with cardio-cerebrovascular diseases," *Zhongguo ying yong sheng li xue za zhi= Zhongguo yingyong shenglixue zazhi= Chinese journal of applied physiology,* vol. 28, pp. 225-229, 2012.

[96] Z. Wang, Y. Chen, S. Hu, R. Liu and W. Yang, "A Meta-analysis of the Association of COX-1 Gene rs3842788 and rs1330344 Polymorphism with Aspirin Resistance in Chinese," *Journal of Medical Diagnostic Methods,* vol. 06, 1 2017.

[97] D. R. Zerbino, P. Achuthan, W. Akanni, M. R. Amode, D. Barrell, J. Bhai, K. Billis, C. Cummins, A. Gall, C. G. Girón and others, "Ensembl 2018," *Nucleic acids research,* vol. 46, pp. D754--D761, 2017.

[98] E. Capriotti, R. Calabrese and R. Casadio, "Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information," *Bioinformatics,* vol. 22, pp. 2729-2734, 2006.

[99] Y. Choi, G. E. Sims, S. Murphy, J. R. Miller and A. P. Chan, "Predicting the functional effect of amino acid substitutions and indels," *PloS one,* vol. 7, p. e46688, 2012.

[100] V. Ramensky, P. Bork and S. Sunyaev, "Human non-synonymous SNPs: server and survey," *Nucleic acids research,* vol. 30, pp. 3894-3900, 2002.

[101] P. C. Ng and S. Henikoff, "Accounting for human polymorphisms predicted to affect protein function," *Genome research,* vol. 12, pp. 436-446, 2002.

[102] P. D. Thomas, A. Kejariwal, M. J. Campbell, H. Mi, K. Diemer, N. Guo, I. Ladunga, B. Ulitsky-Lazareva, A. Muruganujan, S. Rabkin and others, "PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification," *Nucleic acids research,* vol. 31, pp. 334-341, 2003.

[103] E. A. Stone and A. Sidow, "Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity," *Genome research,* vol. 15, pp. 978-986, 2005.

[104] L. Bao, M. Zhou and Y. Cui, "nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms," *Nucleic acids research,* vol. 33, pp. W480--W482, 2005.

[105] D. K. Brown and Ö. Tastan Bishop, "HUMA: A platform for the analysis of genetic variation in humans," *Human mutation,* vol. 39, pp. 40-51, 2018.

[106] E. Capriotti, P. Fariselli and R. Casadio, "I-Mutant2. 0: predicting stability changes upon mutation from the protein sequence or structure," *Nucleic acids research,* vol. 33, pp. W306--W310, 2005.

[107] H. Ashkenazy, S. Abadi, E. Martz, O. Chay, I. Mayrose, T. Pupko and N. Ben-Tal, "ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules," *Nucleic acids research,* vol. 44, pp. W344--W350, 2016.

[108] P. C. Ng and S. Henikoff, "Predicting the effects of amino acid substitutions on protein function," *Annu. Rev. Genomics Hum. Genet.,* vol. 7, pp. 61-80, 2006.

[109] G. Thiltgen and R. A. Goldstein, "Assessing predictors of changes in protein stability upon mutation using self-consistency," *PloS one,* vol. 7, p. e46084, 2012.

[110] E. Capriotti, P. Fariselli and R. Casadio, "A neural-network-based method for predicting protein stability changes upon single point mutations," *Bioinformatics,* vol. 20, pp. i63--i68, 2004.

[111] Y. Zhu, M. R. Spitz, C. I. Amos, J. Lin, M. B. Schabath and X. Wu, "An evolutionary perspective on single-nucleotide polymorphism screening in molecular cancer epidemiology," *Cancer research,* vol. 64, pp. 2251-2257, 2004.

[112] S. Malleshappa Gowder, J. Chatterjee, T. Chaudhuri and K. Paul, "Prediction and analysis of surface hydrophobic residues in tertiary structure of proteins," *The Scientific World Journal,* vol. 2014, 2014.

[113] J. Ipe, M. Swart, K. S. Burgess and T. C. Skaar, "High-throughput assays to assess the functional impact of genetic variants: A road towards genomic-driven medicine," *Clinical and translational science,* vol. 10, pp. 67-77, 2017.

[114] D. K. Brown, O. S. Amamuddy and Ö. T. Bishop, "Structure-based analysis of single nucleotide variants in the renin-angiotensinogen complex," *Global heart,* vol. 12, pp. 121-132, 2017.

[115] C. N. Cavasotto and S. S. Phatak, "Homology modeling in drug discovery: current trends and

applications," *Drug discovery today,* vol. 14, pp. 676-683, 2009.

[116] A. Bishop, T. A. P. De Beer and F. Joubert, "Protein homology modelling and its use in South Africa," *South African Journal of Science,* vol. 104, pp. 2-6, 2008.

[117] C. Chothia and A. M. Lesk, "The relation between the divergence of sequence and structure in proteins.," *The EMBO journal,* vol. 5, pp. 823-826, 1986.

[118] C. Sander and R. Schneider, "Database of homology-derived protein structures and the structural meaning of sequence alignment," *Proteins: Structure, Function, and Bioinformatics,* vol. 9, pp. 56-68, 1991.

[119] K. Illergård, D. H. Ardell and A. Elofsson, "Structure is three to ten times more conserved than sequence_a study of structural response in protein cores," *Proteins: Structure, Function, and Bioinformatics,* vol. 77, pp. 499-508, 2009.

[120] W. R. Pearson, "An introduction to sequence similarity (_homology_) searching," *Current protocols in bioinformatics,* vol. 42, pp. 3-1, 2013.

[121] E. Luccio and P. Koehl, "A quality metric for homology modeling: the H-factor," *BMC bioinformatics,* vol. 12, p. 48, 2011.

[122] C. G. Roessler, B. M. Hall, W. J. Anderson, W. M. Ingram, S. A. Roberts, W. R. Montfort and M. H. J. Cordes, "Transitive homology-guided structural studies lead to discovery of Cro proteins with 40% sequence identity but different folds," *Proceedings of the National Academy of Sciences,* vol. 105, pp. 2343-2348, 2008.

[123] P. A. Alexander, Y. He, Y. Chen, J. Orban and P. N. Bryan, "The design and characterization of two proteins with 88% sequence identity but different structure and function," *Proceedings of the National Academy of Sciences,* vol. 104, pp. 11963-11968, 2007.

[124] B. Rost, "Twilight zone of protein sequence alignments," *Protein engineering,* vol. 12, pp. 85-94, 1999.

[125] D. M. Nikolaev, A. A. Shtyrov, M. S. Panov, A. Jamal, O. B. Chakchir, V. A. Kochemirovsky, M. Olivucci and M. N. Ryazantsev, "A Comparative Study of Modern Homology Modeling Algorithms for Rhodopsin Structure Prediction," *ACS omega,* vol. 3, pp. 7555-7566, 2018.

[126] E. Krieger, S. B. Nabuurs and G. Vriend, "Homology modeling," *Methods of biochemical analysis,* vol. 44, pp. 509-524, 2003.

[127] Z. Xiang, "Advances in homology protein structure modeling," *Current Protein and Peptide Science,* vol. 7, pp. 217-227, 2006.

[128] J. Park, K. Karplus, C. Barrett, R. Hughey, D. Haussler, T. Hubbard and C. Chothia, "Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods1," *Journal of molecular biology,* vol. 284, pp. 1201-1210, 1998.

[129] K. Katoh, J. Rozewicki and K. D. Yamada, "MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization," *Briefings in bioinformatics,* 2017.

[130] F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding and others, "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega," *Molecular systems biology,* vol. 7, p. 539, 2011.

[131] R. C. Edgar, "MUSCLE: multiple sequence alignment with high accuracy and high throughput," *Nucleic acids research,* vol. 32, pp. 1792-1797, 2004.

[132] E. Papaleo, G. Saladino, M. Lambrughi, K. Lindorff-Larsen, F. L. Gervasio and R. Nussinov, "The role of protein loops and linkers in conformational dynamics and allostery," *Chemical reviews,* vol. 116, pp. 6391-6423, 2016.

[133] J. Mendes, A. M. Baptista, M. A. Carrondo and C. M. Soares, "Improved modeling of side-chains in proteins with rotamer-based methods: A flexible rotamer model," *Proteins: Structure, Function, and Bioinformatics,* vol. 37, pp. 530-543, 1999.

[134] S. Liang and N. V. Grishin, "Side-chain modeling with an optimized scoring function," *Protein Science,* vol. 11, pp. 322-331, 2002.

[135] H. Park, S. Ovchinnikov, D. E. Kim, F. DiMaio and D. Baker, "Protein homology model refinement by large-scale energy optimization," *Proceedings of the National Academy of Sciences,* p. 201719115, 2018.

[136] N. Eswar, B. Webb, M. A. Marti-Renom, M. S. Madhusudhan, D. Eramian, M.-y. Shen, U. Pieper and A. Sali, "Comparative protein structure modeling using Modeller," *Current protocols in bioinformatics,* vol. 15, pp. 5-6, 2006.

[137] Y. Song, F. DiMaio, R. Y.-R. Wang, D. Kim, C. Miles, T. J. Brunette, J. Thompson and D. Baker, "High-resolution comparative modeling with RosettaCM," *Structure,* vol. 21, pp. 1735-1742, 2013.

[138] A. Waterhouse, M. Bertoni, S. Bienert, G. Studer, G. Tauriello, R. Gumienny, F. T. Heer, T. A. P. Beer, C. Rempfer, L. Bordoli and others, "SWISS-MODEL: homology modelling of protein structures and complexes," *Nucleic acids research,* 2018.

[139] R. Hatherley, D. K. Brown, M. Glenister and Ö. T. Bishop, "PRIMO: An interactive homology modeling pipeline," *PloS one,* vol. 11, p. e0166698, 2016.

[140] A. Fiser and A. Sali, "ModLoop: automated modeling of loops in protein structures," *Bioinformatics,* vol. 19, pp. 2500-2501, 2003.

[141] G. G. Krivov, M. V. Shapovalov and R. L. Dunbrack, "Improved prediction of protein side-chain conformations with SCWRL4," *Proteins: Structure, Function, and Bioinformatics,* vol. 77, pp. 778-795, 2009.

[142] A. Šali and T. L. Blundell, "Comparative protein modelling by satisfaction of spatial restraints," *Journal of molecular biology,* vol. 234, pp. 779-815, 1993.

[143] A. D. MacKerell Jr, D. Bashford, M. L. D. R. Bellott, R. L. Dunbrack Jr, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha and others, "All-atom empirical potential for molecular modeling and dynamics studies of proteins," *The journal of physical chemistry B,* vol. 102, pp. 3586-3616, 1998.

[144] M.-y. Shen and A. Sali, "Statistical potential for assessment and prediction of protein structures," *Protein science,* vol. 15, pp. 2507-2524, 2006.

[145] F. Melo, R. Sánchez and A. Sali, "Statistical potentials for fold assessment," *Protein science,* vol. 11, pp. 430-448, 2002.

[146] M. Wiederstein and M. J. Sippl, "ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins," *Nucleic acids research,* vol. 35, pp. W407--W410, 2007.

[147] D. Eisenberg, R. Lüthy and J. U. Bowie, "VERIFY3D: assessment of protein models with three-dimensional profiles," in *Methods in enzymology*, vol. 277, Elsevier, 1997, pp. 396-404.

[148] R. A. Laskowski, M. W. MacArthur, D. S. Moss and J. M. Thornton, "PROCHECK: a program to check the stereochemical quality of protein structures," *Journal of applied crystallography,* vol. 26, pp. 283-291, 1993.

[149] P. Benkert, M. Künzli and T. Schwede, "QMEAN server for protein model quality estimation," *Nucleic acids research,* vol. 37, pp. W510--W514, 2009.

[150] R. W. W. Hooft, "Errors in protein structures," *Nature,* vol. 381, p. 272, 1996.

[151] M. J. Sippl, "Knowledge-based potentials for proteins," *Current opinion in structural biology,* vol. 5, pp. 229-235, 1995.

[152] W. G. Touw, C. Baakman, J. Black, T. A. H. Beek, E. Krieger, R. P. Joosten and G. Vriend, "A series of PDB-related databanks for everyday needs," *Nucleic acids research,* vol. 43, pp. D364--D368, 2014.

[153] U. Consortium, "UniProt: the universal protein knowledgebase," *Nucleic acids research,* vol. 45, pp. D158--D169, 2016.

[154] S. F. Altschul, W. Gish, W. Miller, E. W. Myers and D. J. Lipman, "Basic local alignment search tool," *Journal of molecular biology,* vol. 215, pp. 403-410, 1990.

[155] L. Zimmermann, A. Stephens, S.-Z. Nam, D. Rau, J. Kübler, M. Lozajic, F. Gabler, J. Söding, A. N. Lupas and V. Alva, "A completely Reimplemented MPI bioinformatics toolkit with a new HHpred server at its Core," *Journal of molecular biology,* vol. 430, pp. 2237-2243, 2018.

[156] A. G. Murzin, S. E. Brenner, T. Hubbard and C. Chothia, "SCOP: a structural classification of proteins database for the investigation of sequences and structures," *Journal of molecular biology,* vol. 247, pp. 536-540, 1995.

[157] A. Bateman, L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. L. Sonnhammer and others, "The Pfam protein families database," *Nucleic acids research,* vol. 32, pp. D138--D141, 2004.

[158] A. Marchler-Bauer, Y. Bo, L. Han, J. He, C. J. Lanczycki, S. Lu, F. Chitsaz, M. K. Derbyshire, R. C. Geer, N. R. Gonzales and others, "CDD/SPARCLE: functional classification of proteins via subfamily domain architectures," *Nucleic acids research,* vol. 45, pp. D200--D203, 2016.

[159] F. S.-M. Pais, P. Cássia Ruy, G. Oliveira and R. S. Coimbra, "Assessing the efficiency of multiple sequence alignment programs," *Algorithms for Molecular Biology,* vol. 9, p. 4, 2014.

[160] J. Daugelaite, A. O'Driscoll and R. D. Sleator, "An overview of multiple sequence alignments and cloud computing in bioinformatics," *ISRN Biomathematics,* vol. 2013, 2013.

[161] H. Fan and A. E. Mark, "Refinement of homology-based protein structures by molecular dynamics simulation techniques," *Protein Science,* vol. 13, pp. 211-220, 2004.

[162] N. Moitessier, P. Englebienne, D. Lee, J. Lawandi, Corbeil and CR, "Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go," *British journal of pharmacology,* vol. 153, pp. S7--S26, 2008.

[163] B. J. McConkey, V. Sobolev and M. Edelman, "The performance of current methods in ligand--protein docking," *Current Science,* pp. 845-856, 2002.

[164] C. Hetényi and D. Spoel, "Blind docking of drug-sized compounds to proteins with up to a thousand residues," *FEBS letters,* vol. 580, pp. 1447-1450, 2006.

[165] D. E. Koshland Jr, "The key--lock theory and the induced fit theory," *Angewandte Chemie International Edition in English,* vol. 33, pp. 2375-2378, 1995.

[166] G. G. Hammes, "Multiple conformational changes in enzyme catalysis," *Biochemistry,* vol. 41, pp. 8221-8228, 2002.

[167] M. Berry, B. Fielding and J. Gamieldien, "Practical considerations in virtual screening and molecular docking," *Emerging Trends in Computational Biology, Bioinformatics, and Systems Biology; Tran, QN, Hamid, AR, Eds,* pp. 487-502, 2015.

[168] G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew and A. J. Olson, "Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function," *Journal of computational chemistry,* vol. 19, pp. 1639-1662, 1998.

[169] M. Rarey, B. Kramer, T. Lengauer and G. Klebe, "A fast flexible docking method using an incremental construction algorithm," *Journal of molecular biology,* vol. 261, pp. 470-489, 1996.

[170] X.-Y. Meng, H.-X. Zhang, M. Mezei and M. Cui, "Molecular docking: a powerful approach for structure-based drug discovery," *Current computer-aided drug design,* vol. 7, pp. 146-157, 2011.

[171] J. Liu and R. Wang, "Classification of current scoring functions," *Journal of chemical information and modeling,* vol. 55, pp. 475-482, 2015.

[172] J. Baell and M. A. Walters, "Chemistry: Chemical con artists foil drug discovery," *Nature News,*

vol. 513, p. 481, 2014.

[173] W. M. Pardridge, "The blood-brain barrier: bottleneck in brain drug development," *NeuroRx,* vol. 2, pp. 3-14, 2005.

[174] P. Ballabh, A. Braun and M. Nedergaard, "The blood--brain barrier: an overview: structure, regulation, and clinical implications," *Neurobiology of disease,* vol. 16, pp. 1-13, 2004.

[175] A. C. Wallace, R. A. Laskowski and J. M. Thornton, "LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions," *Protein engineering, design and selection,* vol. 8, pp. 127-134, 1995.

[176] D. S. BIOVIA, "Dassault Systèmes," *Discovery Studio Modeling Environment: San Diego, CA, USA,* 2015.

[177] J. J. Irwin and B. K. Shoichet, "ZINC- A free database of commercially available compounds for virtual screening," *Journal of chemical information and modeling,* vol. 45, pp. 177-182, 2005.

[178] Q. Li, T. Cheng, Y. Wang and S. H. Bryant, "PubChem as a public resource for drug discovery," *Drug discovery today,* vol. 15, pp. 1052-1057, 2010.

[179] M. M. Mysinger, M. Carchia, J. J. Irwin and B. K. Shoichet, "Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking," *Journal of medicinal chemistry,* vol. 55, pp. 6582-6594, 2012.

[180] A. Gaulton, A. Hersey, M. Nowotka, A. P. Bento, J. Chambers, D. Mendez, P. Mutowo, F. Atkinson, L. J. Bellis, E. Cibrián-Uhalte and others, "The ChEMBL database in 2017," *Nucleic acids research,* vol. 45, pp. D945--D954, 2016.

[181] D. J. Newman and G. M. Cragg, "Natural products as sources of new drugs over the 30 years from 1981 to 2010," *Journal of natural products,* vol. 75, pp. 311-335, 2012.

[182] C. Y.-C. Chen, "TCM Database@ Taiwan: the world's largest traditional Chinese medicine database for drug screening in silico," *PloS one,* vol. 6, p. e15939, 2011.

[183] F. Ntie-Kang, P. A. Onguéné, M. Scharfe, L. C. O. Owono, E. Megnassan, L. M. Mbaze, W. Sippl and S. M. N. Efange, "ConMedNP: a natural product library from Central African medicinal plants for drug discovery," *Rsc Advances,* vol. 4, pp. 409-419, 2014.

[184] R. Hatherley, D. K. Brown, T. M. Musyoka, D. L. Penkler, N. Faya, K. A. Lobb and Ö. T. Bishop, "SANCDB: a South African natural compound database," *Journal of cheminformatics,* vol. 7, p. 29, 2015.

[185] G. P. Hochgesang, S. W. Rowlinson and L. J. Marnett, "Tyrosine-385 is critical for acetylation of cyclooxygenase-2 by aspirin," *Journal of the American Chemical Society,* vol. 122, pp. 6514-6515, 2000.

[186] J. Henriques, P. J. Costa, M. J. Calhorda and M. Machuqueiro, "Charge Parametrization of the D v H-c 3 Heme Group: Validation Using Constant-(pH, E) Molecular Dynamics Simulations," *The Journal of Physical Chemistry B,* vol. 117, pp. 70-82, 2012.

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

[188] W. L. DeLano, "The PyMOL molecular graphics system," *http://www. pymol. org,* 2002.

[189] C. F. Stratton, D. J. Newman and D. S. Tan, "Cheminformatic comparison of approved drugs from natural product versus synthetic origins," *Bioorganic & medicinal chemistry letters,* vol. 25, pp. 4802-4807, 2015.

[190] A. S. Kalgutkar, B. C. Crews, S. W. Rowlinson, C. Garner, K. Seibert and L. J. Marnett, "Aspirin-like molecules that covalently inactivate cyclooxygenase-2," *Science,* vol. 280, pp. 1268-1270, 1998.

[191] R. E. Hubbard, "Hydrogen bonds in proteins: role and strength," *eLS,* 2001.

[192] C. A. Lipinski, "Lead-and drug-like compounds: the rule-of-five revolution," *Drug Discovery Today: Technologies,* vol. 1, pp. 337-341, 2004.

[193] Z. Xiao, N. C. Waters, C. L. Woodard, Z. Li and P.-K. Li, "Design and synthesis of Pfmrk inhibitors as potential antimalarial agents," *Bioorganic & medicinal chemistry letters,* vol. 11, pp. 2875-2878, 2001.

[194] C. E. Whibley, R. A. Keyzers, A. G. Soper, M. I. C. H. A. E. L. T. DAVIES-COLEMAN, T. Samaai and D. T. Hendricks, "Antiesophageal cancer activity from Southern African marine organisms," *Annals of the New York Academy of Sciences,* vol. 1056, pp. 405-412, 2005.

[195] S. Pather, "Marine biotechnology: evaluation and development of methods for the discovery of natural products from fungi," 2004.

[196] S. P. N. Mativandlela, T. Muthivhi, H. Kikuchi, Y. Oshima, C. Hamilton, A. A. Hussein, M. L. Walt, P. J. Houghton and N. Lall, "Antimycobacterial flavonoids from the leaf extract of Galenia africana," *Journal of natural products,* vol. 72, pp. 2169-2171, 2009.

[197] H. Alonso, A. A. Bliznyuk and J. E. Gready, "Combining docking and molecular dynamic simulations in drug design," *Medicinal research reviews,* vol. 26, pp. 531-568, 2006.

[198] T. Sakano, M. I. Mahamood, T. Yamashita and H. Fujitani, "Molecular dynamics analysis to evaluate docking pose prediction," *Biophysics and physicobiology,* vol. 13, pp. 181-194, 2016.

[199] M. Karplus and J. A. McCammon, "Molecular dynamics simulations of biomolecules," *Nature Structural and Molecular Biology,* vol. 9, p. 646, 2002.

[200] T. M. Musyoka, A. M. Kanzi, K. A. Lobb and Ö. T. Bishop, "Structure based docking and molecular dynamic studies of plasmodial cysteine proteases against a South African natural compound and its analogs," *Scientific reports,* vol. 6, p. 23690, 2016.

[201] A. Raval, S. Piana, M. P. Eastwood, R. O. Dror and D. E. Shaw, "Refinement of protein structure homology models via long, all-atom molecular dynamics simulations," *Proteins: Structure, Function, and Bioinformatics,* vol. 80, pp. 2071-2079, 2012.

[202] D. L. Penkler, C. Atilgan and Ö. Tastan Bishop, "Allosteric Modulation of Human Hsp90α Conformational Dynamics," *Journal of chemical information and modeling,* vol. 58, pp. 383-404, 2018.

[203] H. A. Scheraga, M. Khalili and A. Liwo, "Protein-folding dynamics: overview of molecular simulation techniques," *Annu. Rev. Phys. Chem.,* vol. 58, pp. 57-83, 2007.

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

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

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

[207] M. T. Nelson, W. Humphrey, A. Gursoy, A. Dalke, L. V. Kalé, R. D. Skeel and K. Schulten, "NAMD: a parallel, object-oriented molecular dynamics program," *The International Journal of Supercomputer Applications and High Performance Computing,* vol. 10, pp. 251-268, 1996.

[208] T. Hansson, C. Oostenbrink and W. Gunsteren, "Molecular dynamics simulations," *Current opinion in structural biology,* vol. 12, pp. 190-196, 2002.

[209] M. Rueda, C. Ferrer-Costa, T. Meyer, A. Pérez, J. Camps, J. L. Gelpí, M. Orozco and others, "A consensus view of protein dynamics," *Proceedings of the National Academy of Sciences,* vol. 104, pp. 796-801, 2007.

[210] J. W. Ponder and D. A. Case, "Force fields for protein simulations," in *Advances in protein chemistry*, vol. 66, Elsevier, 2003, pp. 27-85.

[211] K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov and others, "CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields," *Journal of computational chemistry,* vol. 31, pp. 671-690, 2010.

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

[213] W. L. Jorgensen, D. S. Maxwell and J. Tirado-Rives, "Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids," *Journal of the American Chemical Society,* vol. 118, pp. 11225-11236, 1996.

[214] D. Bashford and D. A. Case, "Generalized born models of macromolecular solvation effects," *Annual review of physical chemistry,* vol. 51, pp. 129-152, 2000.

[215] I. Kufareva and R. Abagyan, "Methods of protein structure comparison," in *Homology Modeling*, Springer, 2011, pp. 231-257.

[216] M. Orozco and F. J. Luque, "Theoretical methods for the description of the solvent effect in biomolecular systems," *Chemical Reviews,* vol. 100, pp. 4187-4226, 2000.

[217] R. Anandakrishnan, A. Drozdetski, R. C. Walker and A. V. Onufriev, "Speed of conformational change: comparing explicit and implicit solvent molecular dynamics simulations," *Biophysical journal,* vol. 108, pp. 1153-1164, 2015.

[218] A. W. S. Silva and W. F. Vranken, "ACPYPE-Antechamber python parser interface," *BMC research notes,* vol. 5, p. 367, 2012.

[219] A. W. Schüttelkopf and D. M. F. Van Aalten, "PRODRG: a tool for high-throughput crystallography of protein--ligand complexes," *Acta Crystallographica Section D: Biological Crystallography,* vol. 60, pp. 1355-1363, 2004.

[220] A. A. S. T. Ribeiro, B. A. C. Horta and R. B. d. Alencastro, "MKTOP: a program for automatic construction of molecular topologies," *Journal of the Brazilian Chemical Society,* vol. 19, pp. 1433-1435, 2008.

[221] A. K. Malde, L. Zuo, M. Breeze, M. Stroet, D. Poger, P. C. Nair, C. Oostenbrink and A. E. Mark, "An automated force field topology builder (ATB) and repository: version 1.0," *Journal of chemical theory and computation,* vol. 7, pp. 4026-4037, 2011.

[222] H. J. C. Berendsen, "Dynamic simulation as an essential tool in molecular modeling," *Journal of computer-aided molecular design,* vol. 2, pp. 217-221, 1988.

[223] H. C. Andersen, "Molecular dynamics simulations at constant pressure and/or temperature," *The Journal of chemical physics,* vol. 72, pp. 2384-2393, 1980.

[224] P. H. Hünenberger, "Thermostat algorithms for molecular dynamics simulations," in *Advanced computer simulation*, Springer, 2005, pp. 105-149.

[225] H. J. C. Berendsen, J. P. M. v. Postma, W. F. Gunsteren, A. R. H. J. DiNola and J. R. Haak, "Molecular dynamics with coupling to an external bath," *The Journal of chemical physics,* vol. 81, pp. 3684-3690, 1984.

[226] G. J. Martyna, D. J. Tobias and M. L. Klein, "Constant pressure molecular dynamics algorithms," *The Journal of Chemical Physics,* vol. 101, pp. 4177-4189, 1994.

[227] K. Sargsyan, C. Grauffel and C. Lim, "How molecular size impacts RMSD applications in molecular dynamics simulations," *Journal of chemical theory and computation,* vol. 13, pp. 1518-1524, 2017.

[228] F. Sittel, A. Jain and G. Stock, "Principal component analysis of molecular dynamics: On the use of Cartesian vs. internal coordinates," *The Journal of Chemical Physics,* vol. 141, p. 07B6051, 2014.

[229] P. A. Kollman, I. Massova, C. Reyes, B. Kuhn, S. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang and others, "Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models," *Accounts of chemical research,* vol. 33, pp. 889-897, 2000.

[230] W. Humphrey, A. Dalke and K. Schulten, "VMD: visual molecular dynamics," *Journal of molecular graphics,* vol. 14, pp. 33-38, 1996.

[231] A. Hospital, J. R. Goñi, M. Orozco and J. L. Gelpí, "Molecular dynamics simulations: advances and applications," *Advances and applications in bioinformatics and chemistry: AABC,* vol. 8, p. 37, 2015.

[232] P. Larsson, B. Hess and E. Lindahl, "Algorithm improvements for molecular dynamics simulations," *Wiley interdisciplinary reviews: computational molecular science,* vol. 1, pp. 93-108, 2011.

[233] K. A. Beauchamp, Y.-S. Lin, R. Das and V. S. Pande, "Are protein force fields getting better? A systematic benchmark on 524 diverse NMR measurements," *Journal of chemical theory and computation,* vol. 8, pp. 1409-1414, 2012.

[234] P. Robustelli, S. Piana and D. E. Shaw, "Developing a molecular dynamics force field for both folded and disordered protein states," *Proceedings of the National Academy of Sciences,* p. 201800690, 2018.

[235] N. Schmid, A. P. Eichenberger, A. Choutko, S. Riniker, M. Winger, A. E. Mark and W. F. Gunsteren, "Definition and testing of the GROMOS force-field versions 54A7 and 54B7," *European biophysics journal,* vol. 40, p. 843, 2011.

[236] D. K. Brown, D. L. Penkler, O. Sheik Amamuddy, C. Ross, A. R. Atilgan, C. Atilgan and Ö. Tastan Bishop, "MD-TASK: a software suite for analyzing molecular dynamics trajectories," *Bioinformatics,* vol. 33, pp. 2768-2771, 2017.

[237] R. Kumari, R. Kumar, O. S. D. D. Consortium and A. Lynn, "gmmpbsa A GROMACS tool for high-throughput MM-PBSA calculations," *Journal of chemical information and modeling,* vol. 54, pp. 1951-1962, 2014.

[238] P. Wang, H.-W. Bai and B. T. Zhu, "Structural basis for certain naturally occurring bioflavonoids to function as reducing co-substrates of cyclooxygenase I and II," *PLoS One,* vol. 5, p. e12316, 2010.

[239] W. L. Smith and I. Song, "The enzymology of prostaglandin endoperoxide H synthases-1 and-2," *Prostaglandins & other lipid mediators,* vol. 68, pp. 115-128, 2002.

[240] K. E. Furse, D. A. Pratt, N. A. Porter and T. P. Lybrand, "Molecular dynamics simulations of arachidonic acid complexes with COX-1 and COX-2: insights into equilibrium behavior," *Biochemistry,* vol. 45, pp. 3189-3205, 2006.

[241] M. B. Sevigny, C.-F. Li, M. Alas and M. Hughes-Fulford, "Glycosylation regulates turnover of cyclooxygenase-2," *FEBS letters,* vol. 580, pp. 6533-6536, 2006.

[242] D. D. C. Vann-Victorino, J. Cunanan, M. Chen, R. Chan, R. W. Hall and M. B. Sevigny, "Effect of glycosylation of cyclooxygenase-2 (COX-2) on homodimerization," *The FASEB Journal,* vol. 31, pp. lb79--lb79, 2017.

[243] M. B. Sevigny, K. Graham, E. Ponce, M. C. Louie and K. Mitchell, "Glycosylation of human cyclooxygenase-2 (COX-2) decreases the efficacy of certain COX-2 inhibitors," *Pharmacological research,* vol. 65, pp. 445-450, 2012.

[244] J. C. Otto, D. L. DeWitt and W. L. Smith, "N-glycosylation of prostaglandin endoperoxide

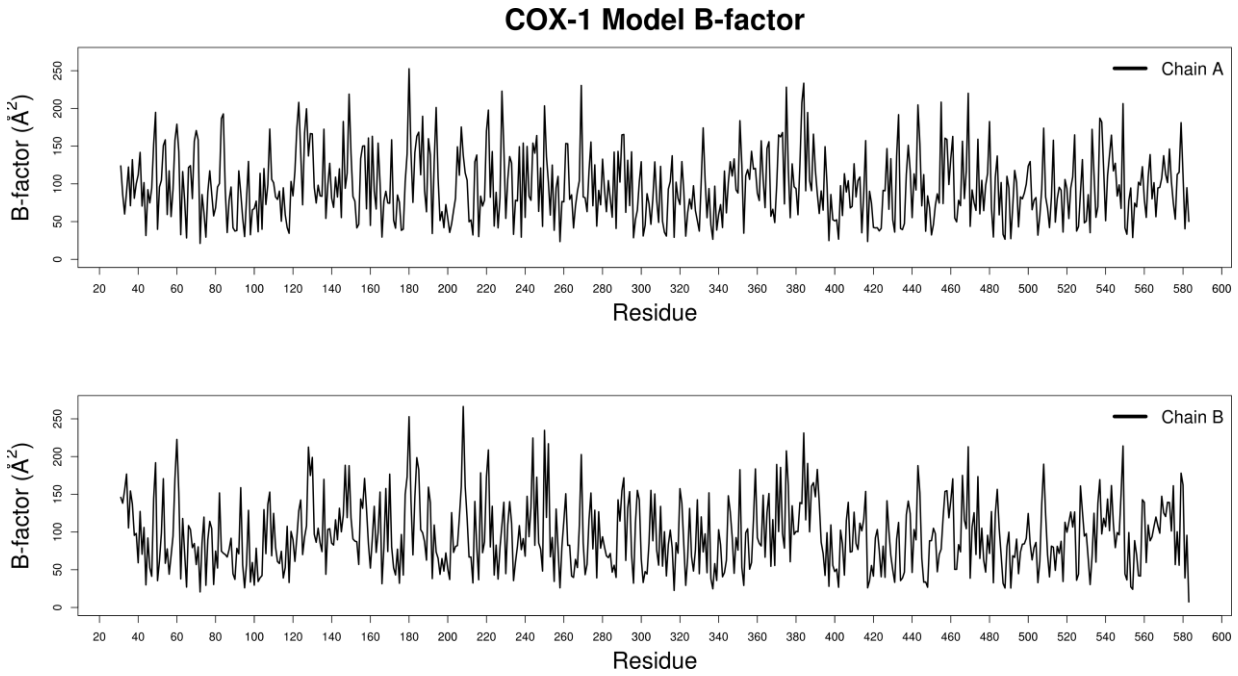
synthases-1 and-2 and their orientations in the endoplasmic reticulum.," *Journal of Biological Chemistry,* vol. 268, pp. 18234-18242, 1993.

[245] T. H. Kim, P. Mehrabi, Z. Ren, A. Sljoka, C. Ing, A. Bezginov, L. Ye, R. Pomès, R. S. Prosser and E. F. Pai, "The role of dimer asymmetry and protomer dynamics in enzyme catalysis," *Science,* vol. 355, p. eaag2355, 2017.

[246] H.-J. Böhm, A. Flohr and M. Stahl, "Scaffold hopping," *Drug discovery today: Technologies,* vol. 1, pp. 217-224, 2004.

[247] C. De Fusco, P. Brear, J. Iegre, K. H. Georgiou, H. F. Sore, M. Hyvönen and D. R. Spring, "A fragment-based approach leading to the discovery of a novel binding site and the selective CK2 inhibitor CAM4066," *Bioorganic & medicinal chemistry,* vol. 25, pp. 3471-3482, 2017.

Supplementary Material

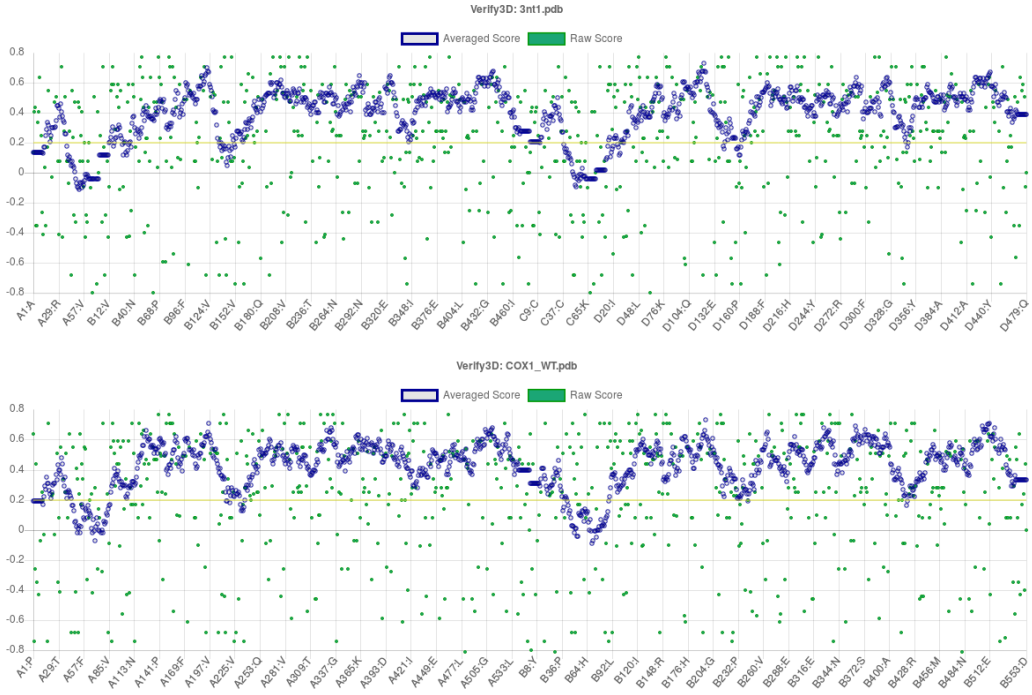


Supplemental Figure 1: B-factors of COX-1 model.

I

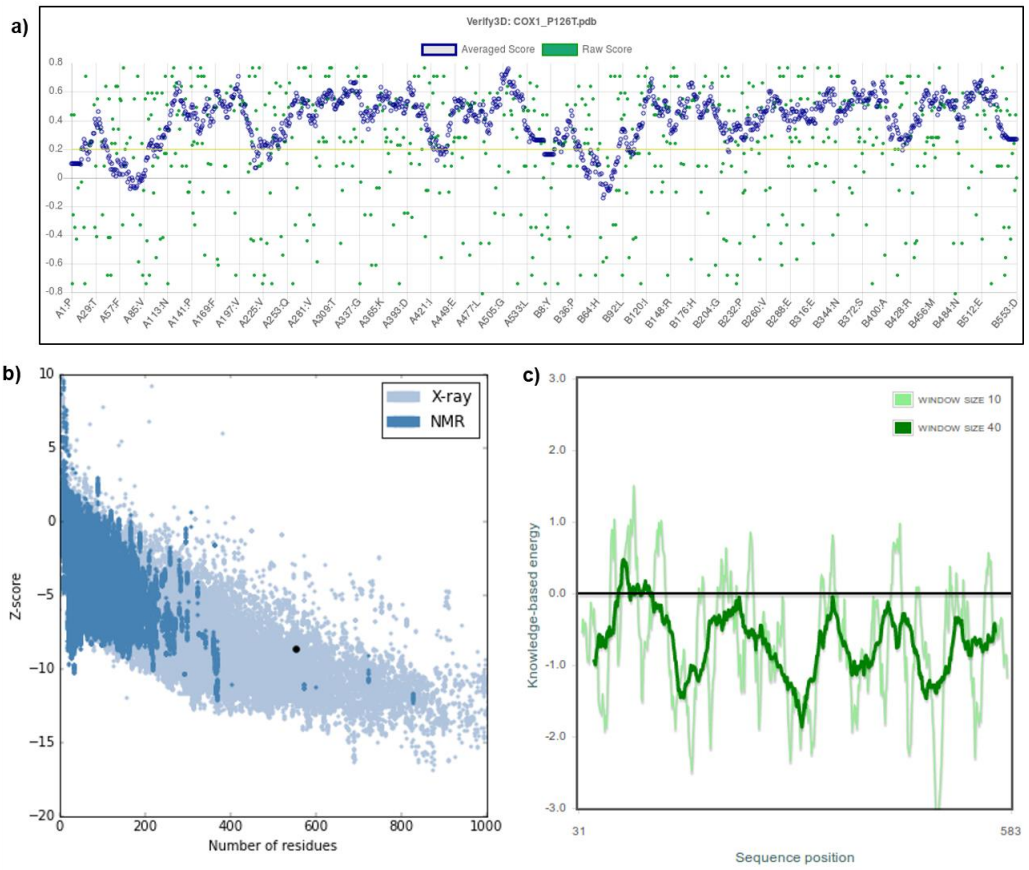
Wild residue	Position	Target residue	PredictSNP prediction	PredictSNP expected accuracy	MAPP prediction	MAPP expected accuracy	PhD-SNP prediction	PhD-SNP expected accuracy	PolyPhen-1 prediction	PolyPhen-1 expected accuracy	PolyPhen-2 prediction	PolyPhen-2 expected accuracy	SIFT prediction	SIFT expected accuracy	SNAP prediction	SNAP expected accuracy	PANTHER prediction	PANTHER expected accuracy
P	126	T	DELETERIOUS	0.86908365	DELETERIOUS	0.85832084	DELETERIOUS	0.73260309	DELETERIOUS	0.74491225	DELETERIOUS	0.63431877	DELETERIOUS	0.79280784	DELETERIOUS	0.72038776	NEUTRAL	0.48032407
N	143	K	DELETERIOUS	0.86908365	DELETERIOUS	0.76611694	DELETERIOUS	0.88474971	DELETERIOUS	0.74491225	DELETERIOUS	0.67524116	DELETERIOUS	0.79280784	DELETERIOUS	0.72038776	UNKNOWN	0
L	237	M	NEUTRAL	0.68365861	NEUTRAL	0.79559471	NEUTRAL	0.68183996	NEUTRAL	0.66884082	NEUTRAL	0.67635271	DELETERIOUS	0.42969871	DELETERIOUS	0.55551884	NEUTRAL	0.47222222
R	244	W	DELETERIOUS	0.86908365	DELETERIOUS	0.7711928	DELETERIOUS	0.87523992	DELETERIOUS	0.74491225	DELETERIOUS	0.81142888	DELETERIOUS	0.79280784	DELETERIOUS	0.88519637	UNKNOWN	0
I	557	T	DELETERIOUS	0.86908365	DELETERIOUS	0.7711928	DELETERIOUS	0.7733853	DELETERIOUS	0.74491225	DELETERIOUS	0.55077121	DELETERIOUS	0.79280784	DELETERIOUS	0.62208185	NEUTRAL	0.47222222

Supplemental Table 1: : PredictSNP summary table, showing SNP effect predictions from integrated tools, including associated accuracy scores.

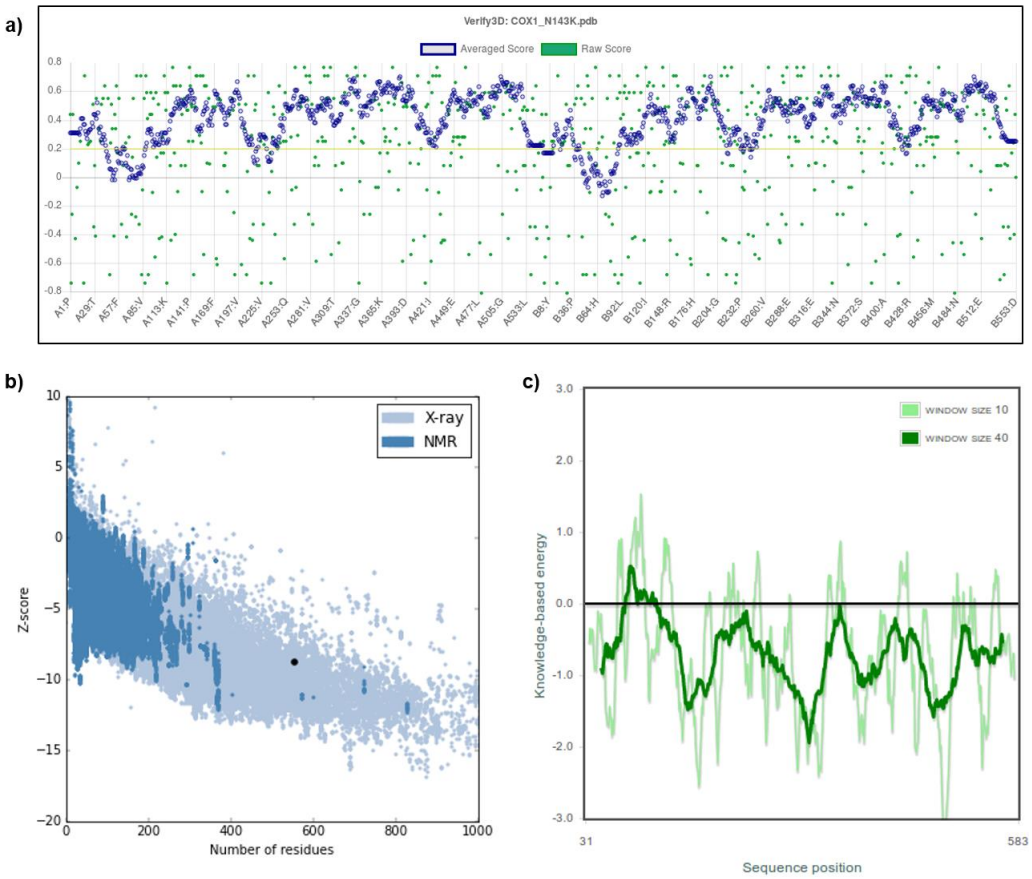


Supplemental Figure 2: Verify3D results for the template 3nt1 (top) and the top ranking wild- type model generated using MODELLER (bottom), showing regions scoring below 0.2.

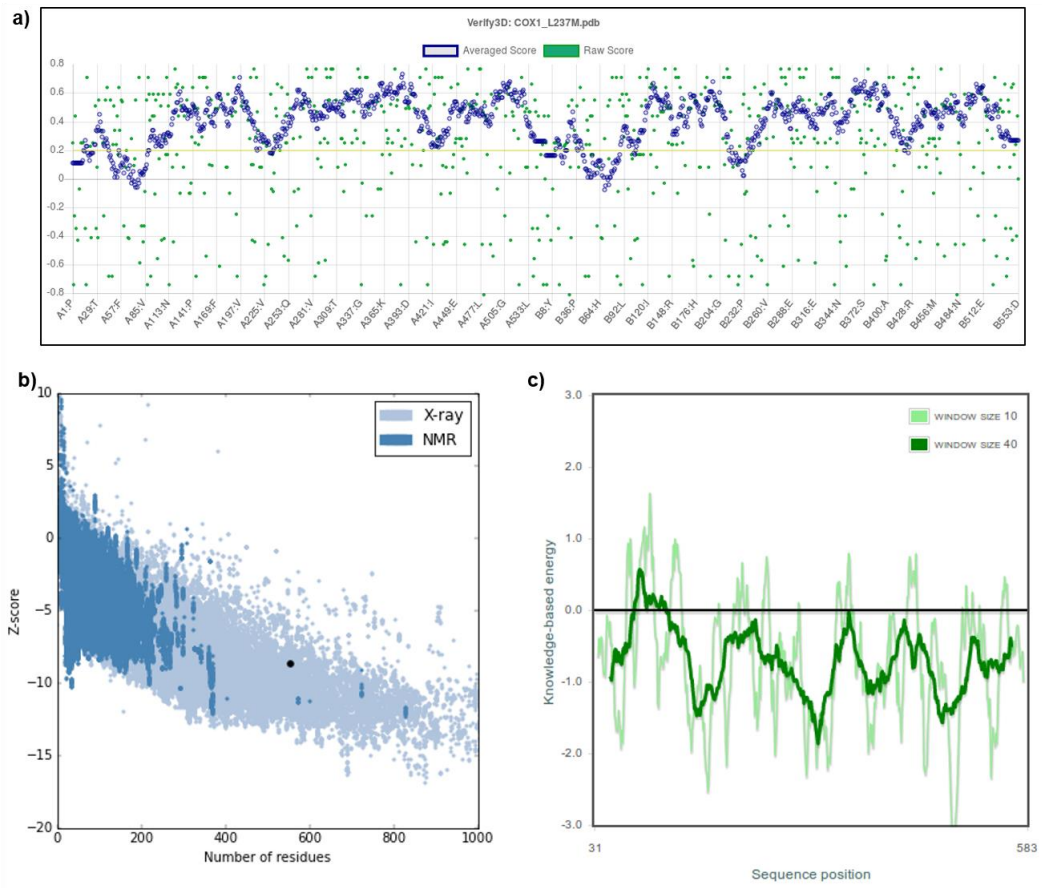
.



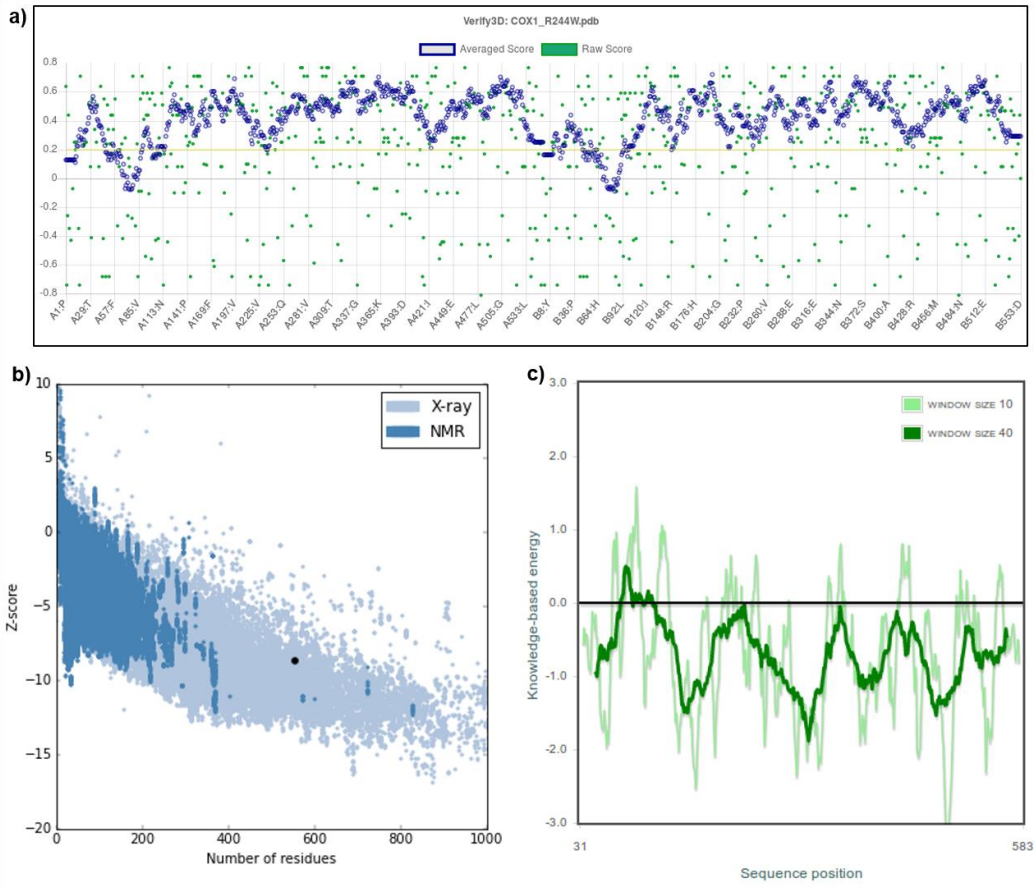
Supplemental Figure 3: P126T validation results, from Verify3D (a), and ProSA-web overall (b) and local (c) model quality



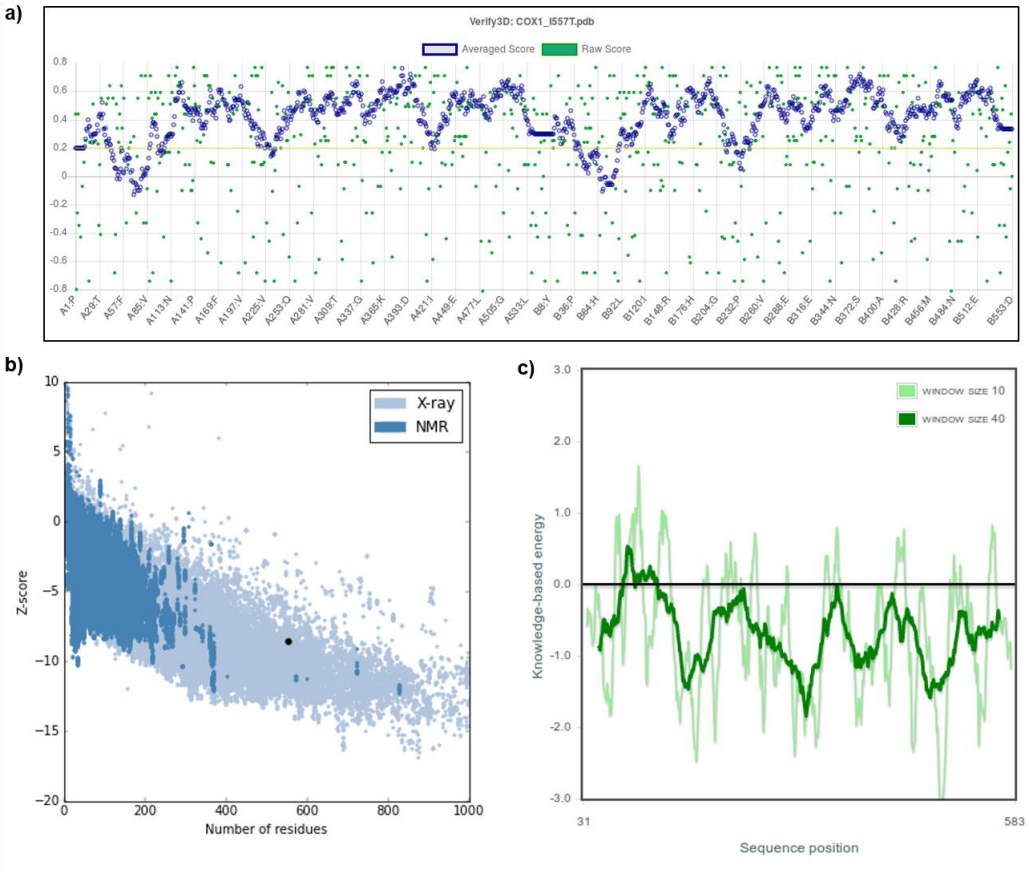
Supplemental Figure 4: : N143k validation results, from Verify3D (a), and ProSA-web overall (b) and local (c) model quality.



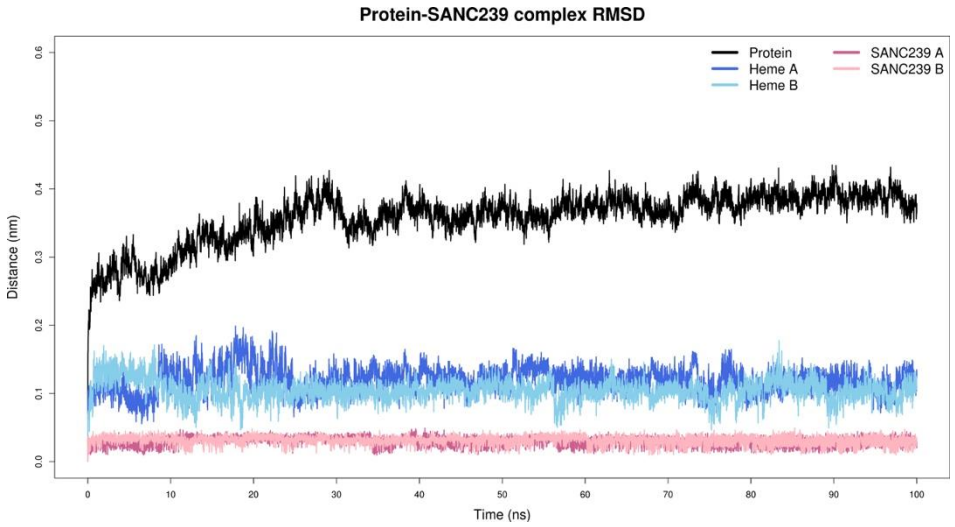
Supplemental Figure 5: L237M validation results, from Verify3D (a), and ProSA-web overall (b) and local (c) model quality



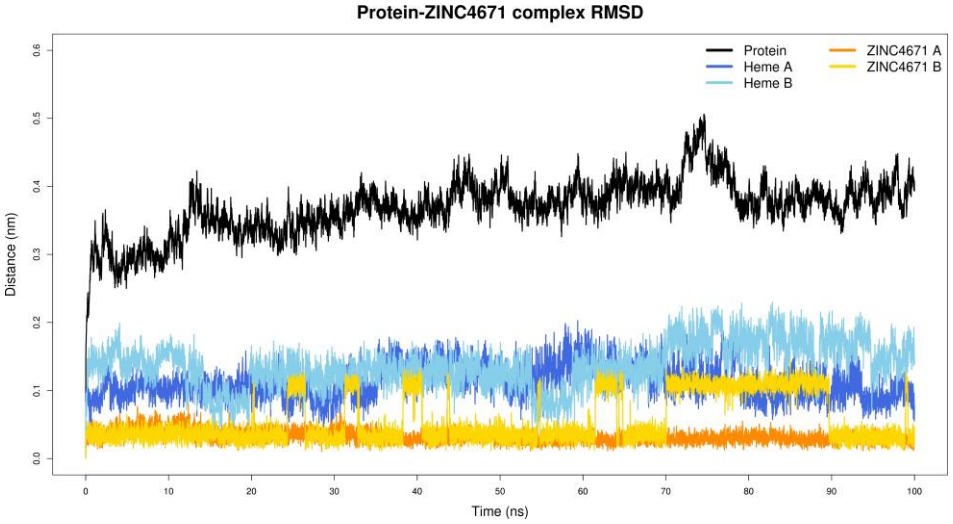
Supplemental Figure 6: R244W validation results, from Verify3D (a), and ProSA-web overall (b) and local (c) model quality.



Supplemental Figure 7: I557T validation results, from Verify3D (a), and ProSA-web overall (b) and (c) local model quality

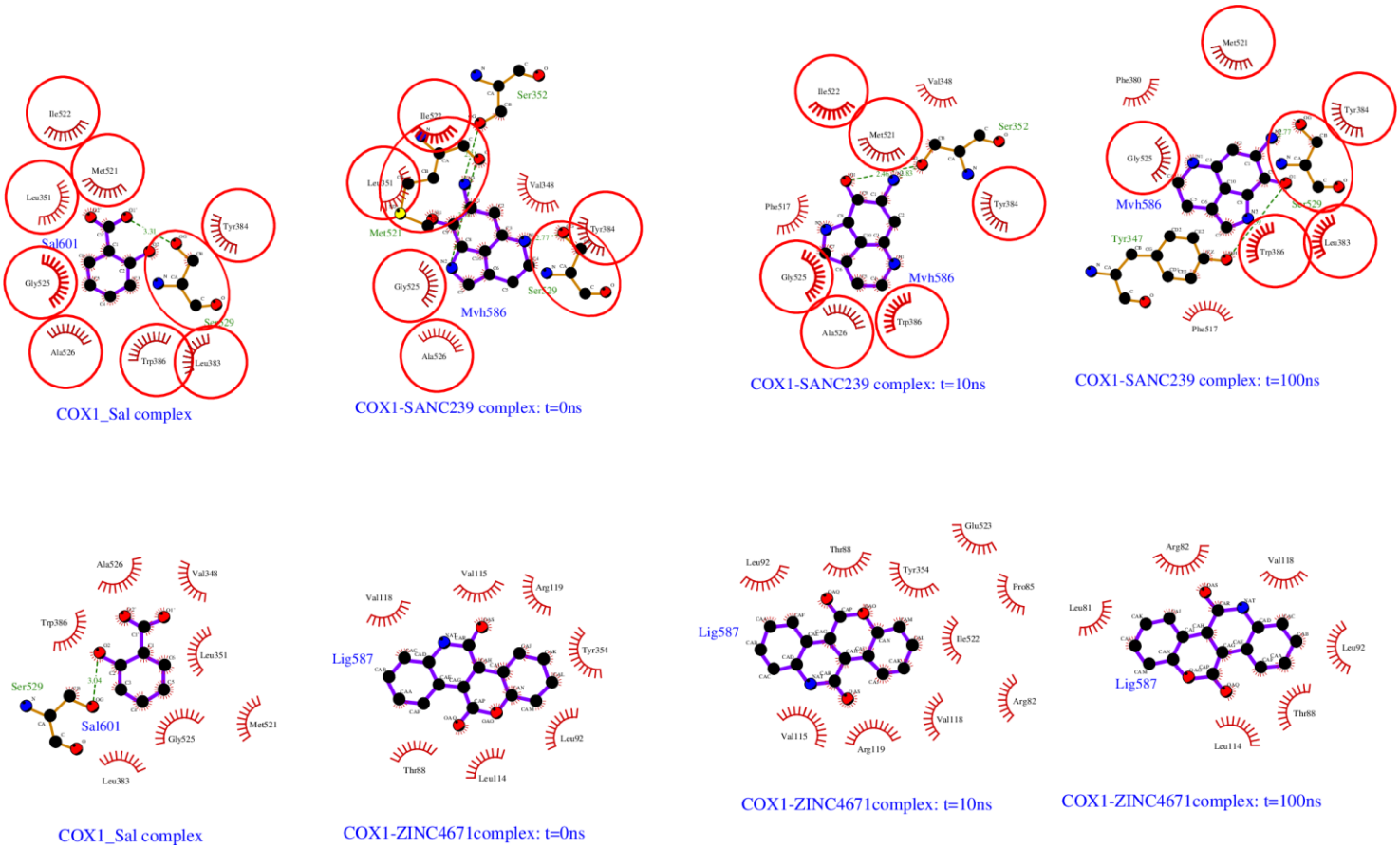


Supplemental Figure 8: RSMD of SANC 239 protein-ligand complex.

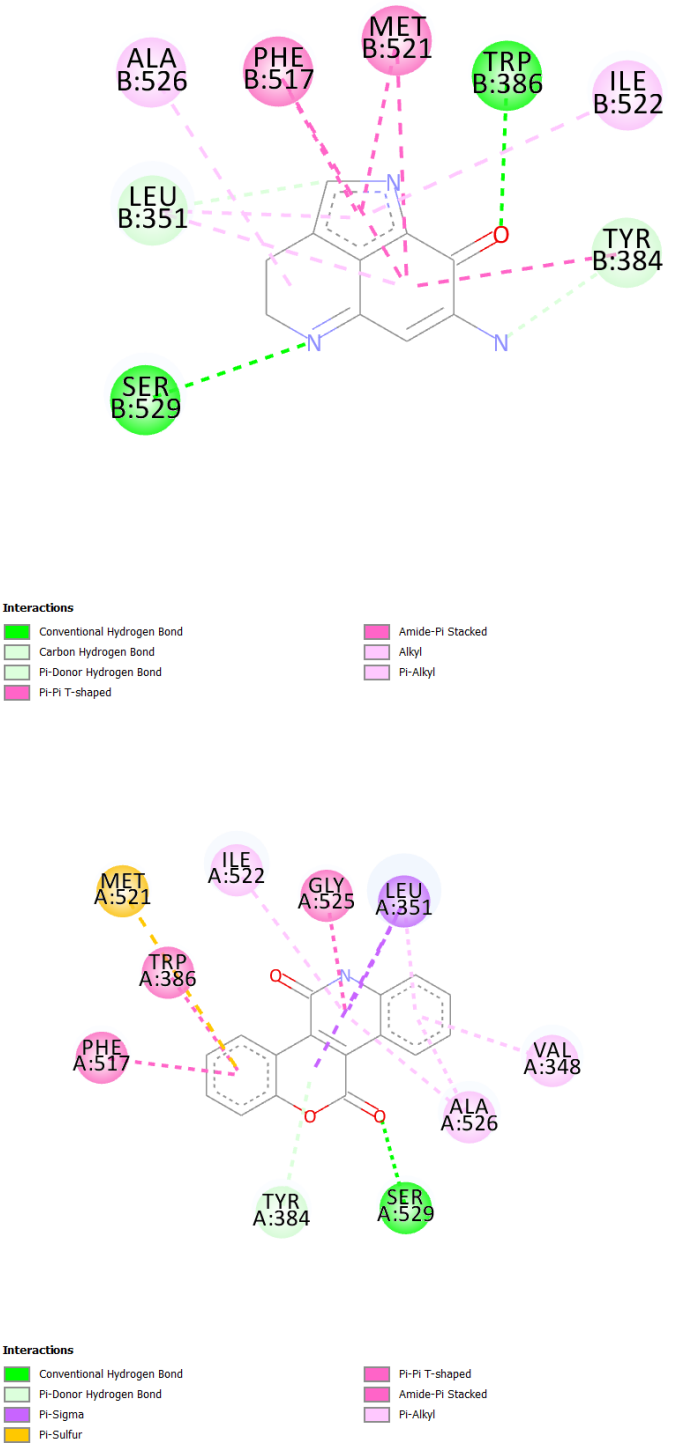


Supplemental Figure 9: RSMD of ZINC4671 protein-ligand complex.

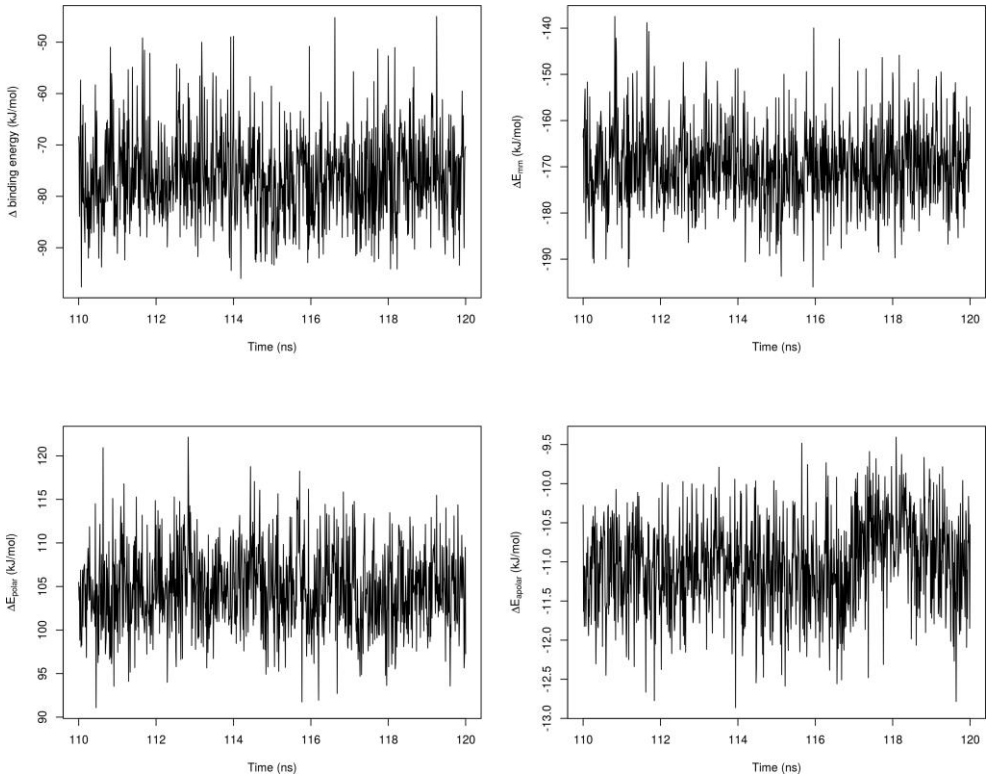
.



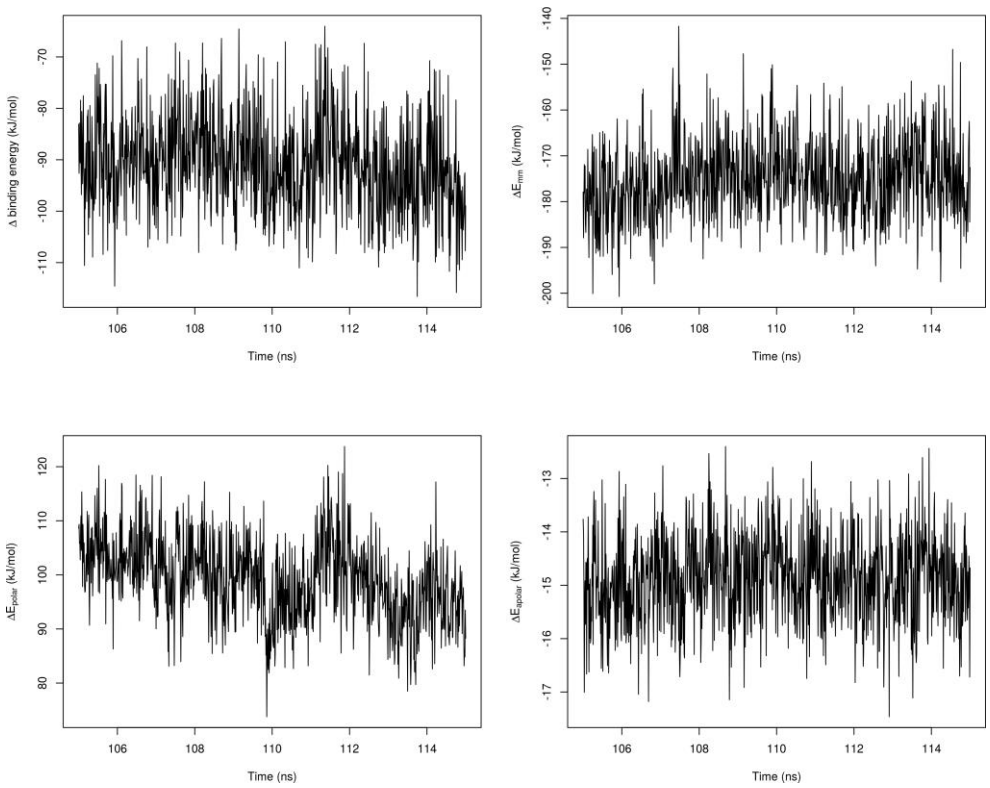
Supplemental Figure 10:: LigPlot diagrams showing interactions of SANC 239 in chain A (above) and ZINC 4671 in chain B (below).



Supplemental Figure 11: DS generated ligand interaction diagrams showing interactions of SANC 239 in chain B (above) and ZINC 4671 in chain A (below).



Supplemental Figure 12: SANC 239 B MMPBSA full energy analysis showing energy fluctuations over the 10ns, used for analysis.



Supplemental Figure 13: ZINC 4671A MMPBSA full energy analysis showing energy fluctuations over the 10ns, used for analysis.