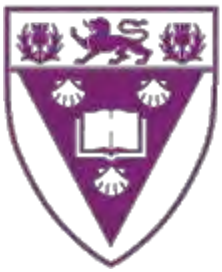


RHODES UNIVERSITY

Where leaders learn

An investigation on the effects of Afrocentric missense variations on the structure and function of CYP2A6 protein

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

MASTER OF SCIENCE

By

Coursework/Thesis

In

Bioinformatics and Computational Molecular Biology

Department of Biochemistry, Microbiology and Bioinformatics

Rhodes University

Chipo P. Makombe

2024

ABSTRACT

Pharmacogenomics, the foundation of personalized medicine distinguishes patients into different categories based on their response to the risk of a disease. Cytochrome P450 (CYPs) proteins are a family of enzymes critical in the metabolism of drugs and other substances. Genetic polymorphisms in CYPs can result in different enzymatic activity in individuals influencing the efficacy and toxicity of drugs. One of the CYPs which primarily metabolizes nicotine and other pharmaceutical drugs such as Artemisinin and Artesunate, Pilocarpine, Valproic Acid and Letrozole is CYP2A6. The gene encoding the protein is highly polymorphic and this can affect the rate of metabolism of drugs in individuals. Previously most studies unveiled connections between CYP2A6 variants and nicotine. Implications concerning the effects of specific missense variations in CYP2A6 drug metabolism have deficiencies. This study aimed to critically examine the structural and functional implications of 13 CYP2A6 allele variations on CYP2A6 protein using Bioinformatics techniques. Methods used were template selection, mutagenesis, parameter assignment and protonation. Molecular Dynamics to get insights regarding protein behavior at an atomic level, clustering to identify conformations during a simulation and DSSP for secondary structure analysis to monitor how secondary structures evolve. Berendsen and Parinello-Rahman barostats at production run were used for comparison. A global analysis was conducted to identify structural transitions (RMSD, RMSF, and Rg), clustering, and secondary structure prediction. Results from Berendsen barostat were inconsistent compared to Parrinello-Rahman barostat implying that CYP2A6 is sensitive to the pressure coupling parameter for precise and accurate results. Our clustering results showed each system in one conformation, fluctuations and shifts on the C-D, H-I loops and F, G, and L helices on variants I149M, F118l, K476R, and E390K_N418D_E419D. This indicated a potential loss of function limiting the protein's ability to conformational flexibility for catalysis and substrate recognition. Certain regions of CYP2A6 became more rigid due to variations, which could have a negative impact on the catalytic activity, regulatory interactions, and general function of the enzyme in metabolism. Globally the variations did not cause large changes to the protein, there is need for a local analysis using Dynamic Residue Networks to study how residue interactions affect the function of CYP2A6.

DECLARATION

I, Chipo Perpetual Makombe, declare that the thesis submitted to Rhodes University is my own work and has not previously been submitted for a degree or diploma at this or any other institution.

Chipo Perpetual Makombe

Signature:  .....

Date: ...November 2024

ACKNOWLEDGEMENTS

I am very much grateful to my supervisor Professor Özlem Tastan Bishop for her valuable tremendous compassion and humble support. The journey would not be fulfilled if it were not for your understanding, kind knowledge, and expertise in supervision, and science.

I appreciate all the efforts from my co-supervisor Dr Allan Sanyanga for his mentorship making it happen, taking his valuable time to read this thesis, and providing guidelines for the success of this study.

I would like to thank the Centre for High Performance Computing (CHPC), Cape Town, South Africa for providing computing resources. Also not forgetting the generous financial support from Novartis and GSK R&D Grant #GSKNVS1/202101/002 under the Project Africa Genomic Research Approach for Diversity and Optimizing Therapeutics program.

Chichi and Rehema thanks for all the help in this study. Thank you Shakes, Queenie, Rabelani, and Curtis for sharing a combination of valuable intellectual knowledge from our different academic backgrounds. I am so grateful to the RUBI family for their support, advice, and inspiration.

Finally, great thanks to my family, “Alice Chidembo" and the late “Ian Muwunde” for the motivation and spiritual support that made me confident. Keynes, Anotipa, and Tomuda my special thanks to you for your understanding.

DEDICATION

To Simbarashe Makombe, my greatest supporter, you walked every step of this journey, pushing me to shoot the stars with love at every turn.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

Abbreviation	Definition
2D	Two-dimensional
DSSP	Define Secondary Structure of Proteins
FF	Force fields
GROMACS	GROningen MAchine for Chemical Simulations
MD	Molecular dynamics
NADP	Nicotinamide Adenine Dinucleotide Phosphate
PDB	Protein Data Bank
RCSB	Research Collaboratory for Structural Bioinformatics
Rg	Radius of gyration
RMSD	Root mean square deviation
RMSF	Root mean square fluctuation
SI	Supplementary Information
SNP	Single Nucleotide Polymorphism
SRS	Substrate Recognition Site
VMD	Visual Molecular Dynamics
WT	Wild type

LIST OF FIGURES

LIST OF TABLES

LIST OF SUPPLEMENTARY FIGURES

LIST OF SUPPLEMENTARY TABLES

LIST OF AMINO ACIDS

Amino Acid	**Letter code single/three letter**
Alanine	A / ALA
Cysteine	C/CYS
Aspartic acid	D/ASP
Glutamic acid	E/GLU
Phenylalanine	F/PHE
Glycine	G/GLY
Histidine	H/HIS
Isoleucine	I/ILE
Lysine	K/LYS
Leucine	L/LEU
Methionine	M/MET
Asparagine	N/ASN
Proline	P/PRO
Glutamine	Q/GLN
Arginine	R/ARG
Serine	S/SER
Threonine	T/THR
Valine	V/VAL
Tryptophan	W/TRP
Tyrosine	Y/TYR

LIST OF WEB SERVERS AND SOFTWARE

Web server	Software
AMBER	https://ambermd.org/
Discovery Studio Visualizer	https://discover.3ds.com/discovery-studio-visualizer-down
GROMACS	https://manual.gromacs.org/5.1/user-guide/mdp-options.html
H++	http://newbiophysics.cs.vt.edu/H++/
OPM	https://opm.phar.umich.edu/
PharmVar	https://www.pharmvar.org/gene/CYP2A6
PyMOL	https://pymol.org/2/
RCSB	https://www.rcsb.org/
UniProt	https://www.uniprot.org/
VMD	https://www.ks.uiuc.edu/Research/vmd/
Xmgrace	https://plasma-gate.weizmann.ac.il/Grace/

CHAPTER 1: INTRODUCTION

1.1 Background

Pharmacogenomics is the pillar of personalized medicine [1] which distinguishes patients into different categories depending on their response to drugs resulting from one's genetic form. This extends to different populations having different variants (pharmacodynamics effects) [2,3] that affects drug metabolizing enzymes and transporters. How drugs are metabolized is very critical, it serves a crucial role in determining drug targets for suitable and effective therapies (precision medicine).

Gonzalez and Nebert [4] postulated that there are different drug metabolizers. Poor drug metabolizers are individuals with two variant alleles, which fail to have an adequate operational enzyme (decreased metabolism of drugs) and normal doses will result in drug toxicity. Most people are normal metabolizers with alleles from two wild types (WT) and take the normal drug dose. The Intermediate metabolizers possess both a WT allele and a variant allele. Lastly are the fast metabolizers where increased metabolism results in drugs being disposed of faster than expected and therefore these individuals need higher levels of dosages [5-9]. Reducing side effects and increasing pharmacological efficiency can be achieved by guaranteeing that the therapy is directly related to an individual's genetic makeup [10,11]. This takes us to drug potency and its efficacy which is explained as an indication of strength [12] of the drug's impact upon binding to its target.

The family of cytochrome P450 enzymes (CYPs) is largely responsible for the disintegration and elimination of most natural, foreign, and therapeutics (xenobiotic) through interaction with a wide range of chemically varied small compounds that form unique rates of metabolic processes [13]. Variations of human CYPs can cause metabolic abnormalities [5]. Proteins rely on the structural properties for proper biological functioning making it very important to study the effects of these variations. Understanding the structural components of proteins can guarantee safe therapies and effective metabolism.

CYP enzymes involved in human metabolism are categorized into 18 families and 43 subfamilies [14] as illustrated in Figure 1.1. Approximately 80% of FDA accepted drugs and over 50% of CYP related drug metabolism are metabolized by CYP families 1, 2, and 3 [15]. These families comprise CYP1A1, CYP1A2, CYP1B1, CYP1D1, CYP2A6, CYP2A7, CYP2A13, CYP2B6, CYP2C8, CYP2C9, CYP2C18, CYP2C19, CYP2D6, CYP2D7, CYP2E1, CYP2F1, CYP3A4, CYP3A5 and CYP3A7. In humans, the largest subfamily is CYP2 and is the most abundant member. It contains key enzymes involved in the metabolism of several medicines, such as CYP2D6, CYP2C9, CYP2C19 and CYP2B6. Furthermore, CYP2A6, CYP2B6, and CYP3A4 are extremely diverse [16]**.** CYP2A6 enzyme, the largest member of the class 2 CYP family, from subfamily A, was the subject of this investigation [17]. It consists of more closely related CYP2s than any other CYP2 subfamily, CYP2A7 (not functional) and CYP2A13 [18].

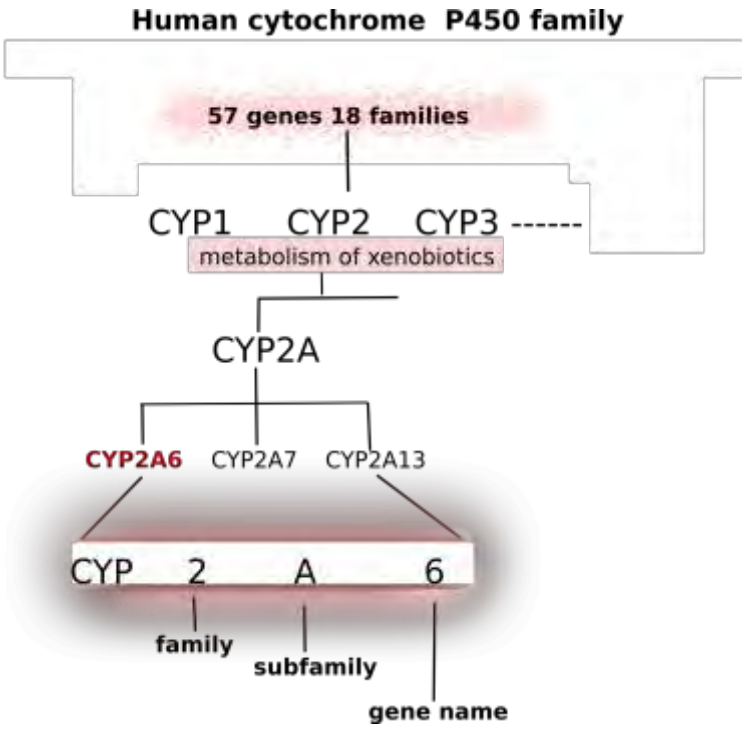


Figure 1.1. The CYP450 superfamily. In the family 2 subfamily A is CYP2A6, CYP2A7 and CYP2A13. The protein of interest CYP2A6 is marked in red.

CYP2A6 is a monooxygenase transmembrane enzyme dominant in the liver [19,20]. The CYP2A6 enzyme is one of the most active members in the 2A subfamily however it has been described as one of the most poorly understood CYP enzymes [21]. CYP2A6 lacks a comprehensive understanding regarding its complex genetic diversity, broad range of substrates, and intricate regulatory processes. The broader pharmacological and physiological functions have not been thoroughly examined. This narrow focus has led to significant gaps in our knowledge of this enzyme's activity in various populations and situations. In contrast to CYP2D6 and CYP3A4, CYP2A6 metabolizes only a handful of clinically utilized medications to a substantial degree [22]. CYP2A6 mainly metabolizes [23] about 3% of clinically authorized medications, which include treatments for cancer, tuberculosis, and malaria which are extremely prevalent in Africa. Variations in CYP2A6 function are an important clinical aspect since this enzyme takes part in the metabolism or bioactivation of medicinal therapies [24,25].

1.2 Drug metabolism

Metabolism, medicine potency, dosage, and mode of administration are among the variables that affect pharmacological effects. Drug metabolism has been defined as a very important process in the human body and an aspect of medical practice and pharmacology [26,27]. To enable the elimination of drugs via urine, the drug's highly lipophilic centres are changed to hydrophilic centres, making the substance water soluble [28]. Metabolism can convert prodrugs as illustrated in Figure 1.2 into active metabolites and this biotransformation primarily occurs in hepatocytes cells. Drug metabolism is through phase I (modifications) and phase II (conjugation) [22,29].

Phase I metabolism is described as a modified set of predominantly reactive non-synthetic processes mediated by CYPs [30,31]. Important subfamilies of CYP in Phase I include CYP1, CYP2, and CYP3, particularly CYP1A2, CYP2D6, CYP2C9, CYP2C19, CYP2A6 and CYP3A4. The purpose of these reactions is to transform water soluble lipophilic metabolites into polarized substances by adding a polar functional group like hydroxyl R-OH [26]. Furthermore, these interactions usually result in the formation of functional metabolites which allow prodrugs into

potent original drugs [32]. Phase I oxidation provides a functional reaction which renders the ability to conjugate with a polar chemical in phase II [32].

Phase II employs the following conjugation reactions: glucuronidation by glucuronic acid, acetylation by acetate, and sulfation by sulfate to create an inactive polar conjugate that will be expelled from the body [33,34]. Phase I reactions allow conjugation reactions with an endogenous substance of Phase II to take place [33].

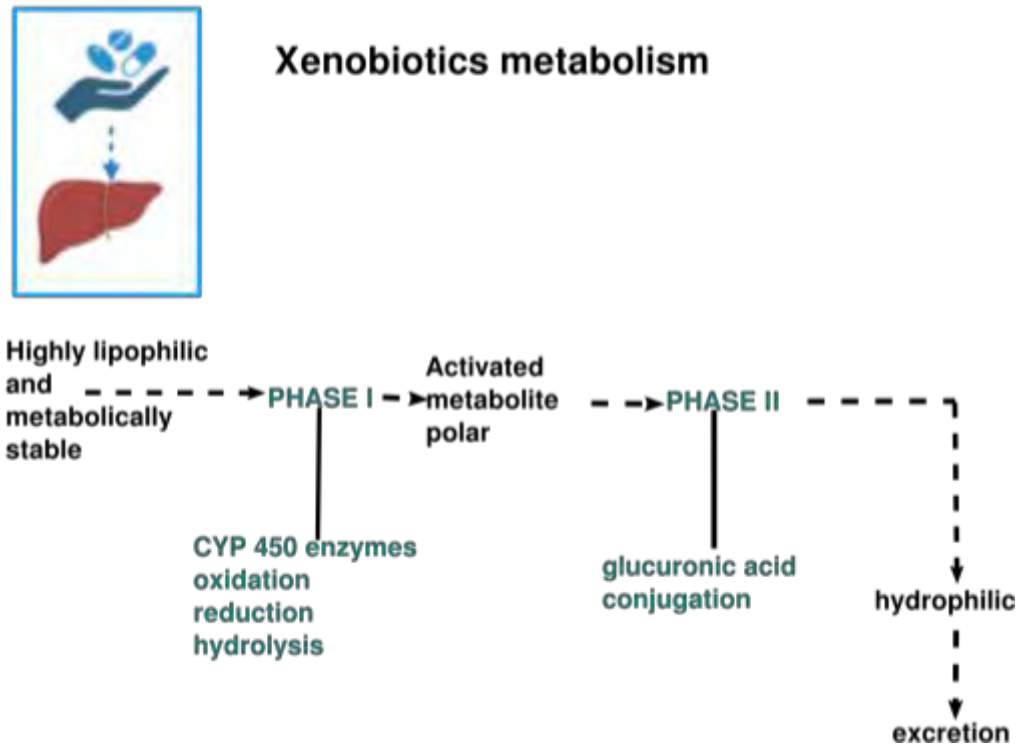


Figure 1.2. Metabolism of xenobiotics. Phase I involves the metabolism of drugs to a water-loving metabolite through reactions like oxidation, reduction, and hydrolysis. In the phase II stage, the polar metabolites are further conjugated to become more hydrophilic for easy excretion

An extended period of a drug in the body makes it toxic and should a drug leave the body rapidly, it may not be effective thus phase I metabolism might have been restricted. The rate of metabolism of substrates is hindered by the massive personalized variations in CYP2A6 enzymatic activity. It

was reported that other aspects like age changes, gender differences, and relationships with other hepatic enzymes, pharmacological, naturally occurring, and dietary compounds, cofactors, and coenzymes contribute to the difference in enzyme reaction [35]. Mwenifumbo et al [36] revealed that allele variations have different frequencies with the bulk of them having a clinical significance to population genomics. The genetic polymorphisms within the CYP2A6 allele perform a significant part in varying expression [23]. Thus, an inactive drug can be metabolically activated and this generally affects the therapeutic impact. Understanding the enzymatic activity that is in a seven-stage catalytic cycle is very important in drug metabolism for optimal pharmacological intervention in individualized treatment.

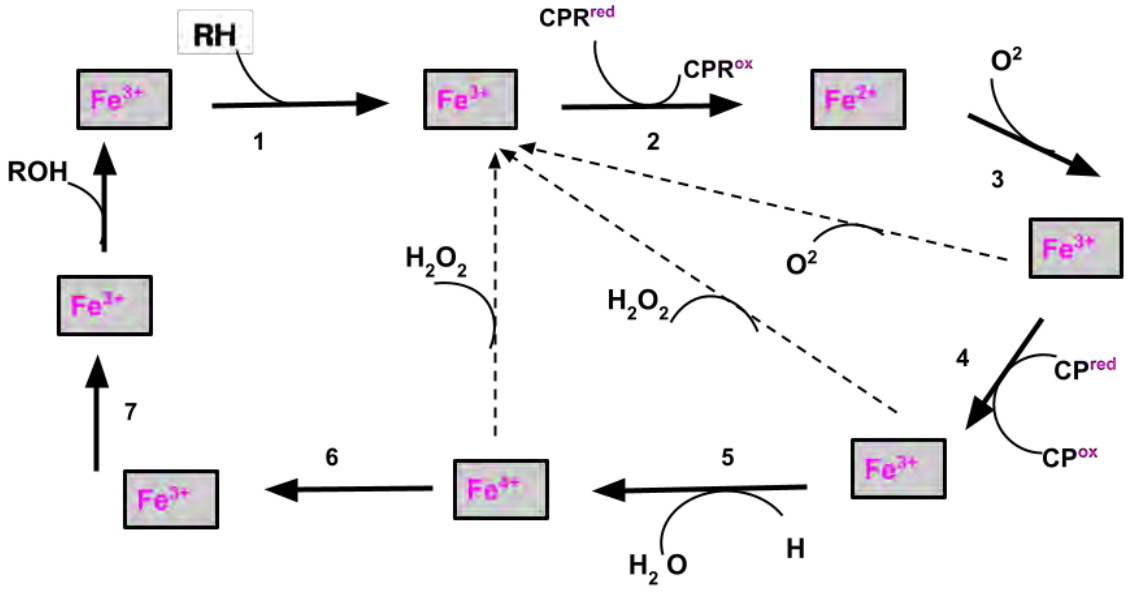


Figure 1.3. The seven stages of the P450. From stages one to seven: substrate binding, iron reduction, oxygen binding, activation, catalysis, product release, and iron oxidation respectively then back to its original state for another cycle.

Modified efficacy, decreased metabolism, heightened formation of harmful metabolites, and unfavorable drug interactions all contribute to increased toxicity, thus different drugs have different pharmacokinetic statuses [37]. The CYP enzyme has a heme with an iron center whose role is to insert molecular oxygen into activation. This is followed by the seven stages of the catalytic cycle [38] illustrated in Figure 1.3. The primary function is to put the oxygen molecule into a stable and hydrophobic compound [39]. All CYPs have a reduction role awaiting substrate binding just before oxygen binding. First step; the CYP enzyme opens and the substrate binds to it [40]. At this point, iron is often in the ferric form. The first iron reduction occurs in the second stage, where redox partners provide the first two electrons. Fe^{2+} is created when NAD(P)H uses the electron transfer chain to transfer an electron to the heme. At this point, oxygen from the lungs attaches itself to the Fe^{2+} heme group, converting $Fe^{2+}O_2$ to $Fe^{3+}O_2$. The second reduction process, which forms $Fe^{3+}O_2^{2-}$, is the fourth stage. Fifth stage; O_2^{2-} splits the oxygen link with two protons to generate ($Fe^{3+}O$). In the sixth step, heme-bound oxygen atoms are moved to the substrate and lastly, the product is released from the enzyme and returns to its original condition.

1.3 CYP2A6 structure and function

The oxidoreductase class CYP2A6 is a monomer with four common beta sheets and twelve common helices (designated A–L). The structure of CYP2A6 PDB ID: 2FDV (UniProt ID: P11509) displayed in Figure 1.4 is found on the RCSB Protein Data Bank (PDB) database. The catalytic portion of CYP450 enzymes contain heme iron centers which are important in the catalytic cycle [41].

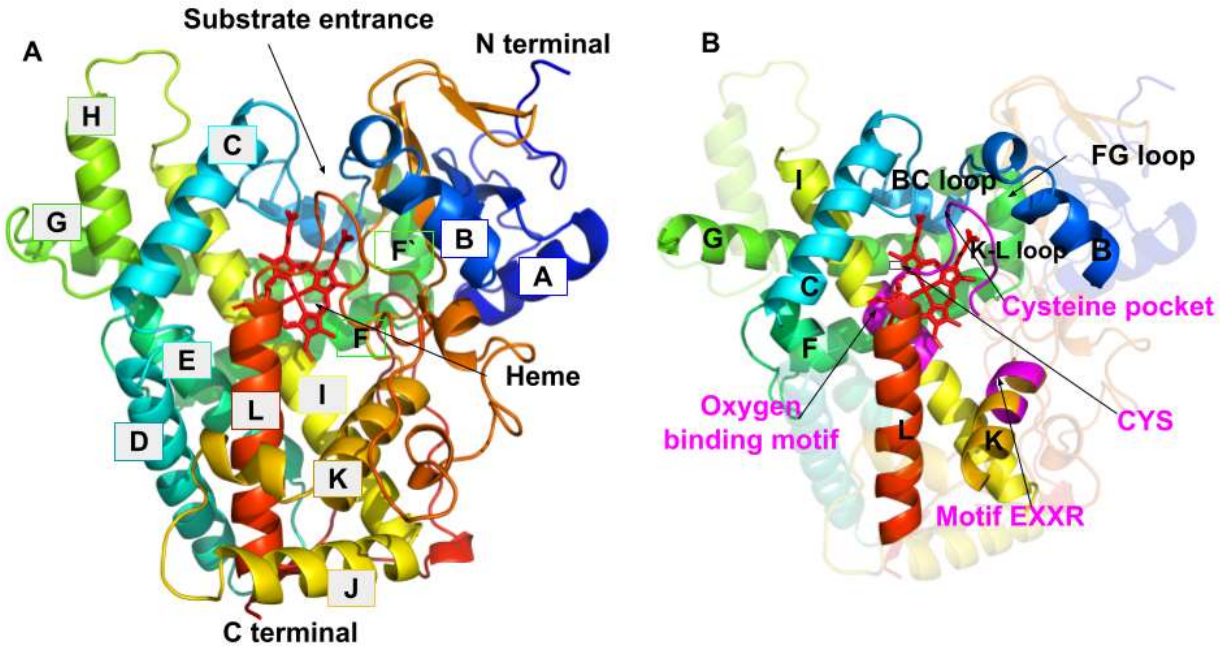


Figure 1.4. Crystal structure of CYP2A6 PDB ID: 2FDV presented in cartoon. Labelled A-L are helices and the heme iron center is in red (A). Important secondary structures (B), for substrate entrance is the B-C loop (royal blue) and F-G loop (green), F helix (dirty pink) and G helix (green), the oxygen motif (magenta) in the K helix, I helix (yellow), signature sequence (magenta), the EXXR motif (magenta) in the K helix and cysteine (CYS) (magenta) in the K-L loop. structures were generated using PyMol software.

The CYP2A6 crystal structure has residues starting from 30 to 494 residues. The N-terminal initial 29 residues were removed from the protein since the crystal structure of these residues did not exhibit any systematic secondary structure. Each of the secondary structures displayed in Figure 1.4 B plays a role in the functioning of the CYP2A6 enzyme however the BC loop, the L - K loop, helices F, G, C, K, L, and I have important functions summarized in Table 1.1.

TABLE 1.1.

IMPORTANT SECONDARY STRUCTURES IN CYP2A6 (PDB ID: 2FDV)

Secondary structure	**Function and position**
I and L helices	Helix I contains the oxygen binding motif [42] and the catalytic site is positioned between helix I and L.
B-C loop	Substrate access to the catalytic site (residue 92-120) and are more flexible
F-G helices	Flexible helices that comprise the top portion of the catalytic site, holds CYPs in the membrane, they can alter conformation together with the F` G` helices to facilitate substrate entrance for enzymatic action (F residues: 196-212, G residues: 233-256).
C, K, L helices	Interconnect with the redox partner (residues C: 121–137, L: 442-459) K: 350–363 with glutamic acid, arginine motif EXXR on 358-361 [43]
I helix	I helix residues 288-319 have a greatly conserved motif (A/G)XXXT important in CYPs catalytic cycle (activates and binds oxygen) [44,45]
L-K loop	accommodates the cysteine residue 439 [45]
Signature sequence/ bulge region (L-K loop)	FXXGXXXCXG [45] cysteine pocket FSIGKRNCFG residues 432-441 and the most conserved region

CYP2A6 has six Substrate Recognition Sites (SRS1- SRS6) [46-48], depicted in Figure 1.5 A. The SRS sites are located within specific regions in the important functional sites (domains), SRS1 (B` helix), SRS2 (F helix), SRS3 (G helix), SRS4 (I helix), SRS5 (Beta 4 sheet turn) and SRS6 (KL

loop). The SRS sites (SRS1, SRS4, SRS5, SRS6) attract and, houses the substrates allowing them to gain recognition and interaction with the protein to bind and SRS sites SRS2 and SRS3 create the access channel keeping the protein structure steady [49].

CYP2A6's catalytic domain in Figure 1.5 B consists of a small catalytic site that binds to tiny substrates [50]. The catalytic site has a volume of 2.65 $Å^3$ and is in a closed conformation [50]. Maestro sitemap was used to predict the catalytic site residues: F107, F111, V117, F118, I208, F209, L241, N297, I300, G301, T305, L370, and F480, displayed in Figure 1.5 B. The catalytic site is hydrophobic except for residues T305 and N297 whose major role is hydrogen bonding for substrate binding [50,51].

To depict the membrane-bound residues for CYP2A6 structure, the Orientations of Proteins in Membranes (OPM) database was used for Figure 1.5 C. CYP enzymes have contact with the membrane surface through their N-terminal domain as demonstrated in Figure 1.5 C, allowing liposoluble substrates to bind directly [24,25] to the catalytic site within the protein, however, the nature of this association is still poorly known. For CYPs to catalyze reactions, they need electrons from NADPH-cytochrome P450 reductase which is encased in the membrane protein [52]. For CYP2A6 to remain active while upholding its structural integrity, the membrane domain environment is important for keeping protein conformations open and providing an access tunnel [53,54].

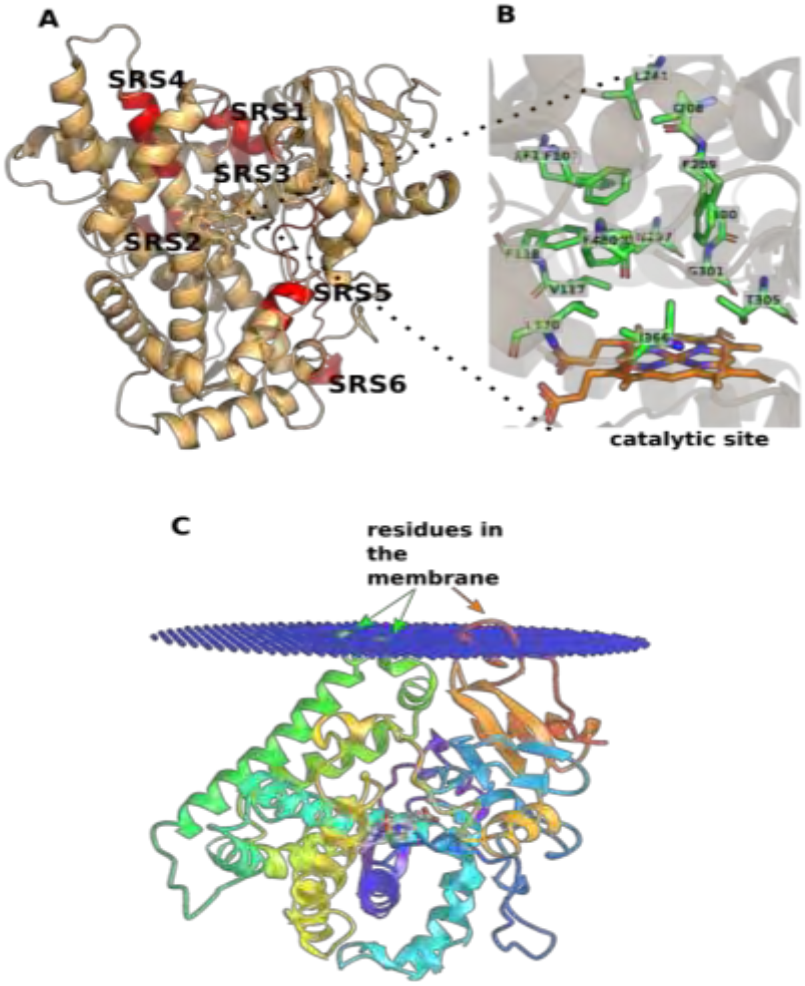


Figure 1.5. CYP2A6 important sites. Substrate Recognition Sites (SRS) SRS 1-6 which are important for substrate specificity and binding are marked in red in the CYP2A6 structure (A). Residues that form the catalytic site (B) are in the green stick model and in brown is the heme with an iron center, nitrogen (blue), and oxygen (red). In Figure C are the residues anchored in the membrane.

The structural stability, interactions with substrates, and catalytic efficacy of membrane-bound enzymes are significantly influenced by the membrane environment. The regions in the membrane are hydrophobic: 39-43 in the N terminal loop and 226 and 229 in the G` and F` regions [48,49]. Sequences of atoms exhibiting extreme temperature factors connected to the protein membrane and the catalytic region are recognized in transmembrane protein analysis as potential sites for channel openings [55].

1.4 CYP2A6 alleles and genetic polymorphisms

CYP2A6 is highly polymorphic, indicating a high degree of inter-ethnic variation; to date, there are more than 48 known allelic variants for CYP2A6 enzyme including their sub-alleles from PharmVar database (https://www.pharmvar.org/gene/CYP2A6, accessed 13 March 2024). The majority of these CYP2A6 polymorphisms are non-synonymous SNP alleles with decreased enzyme activity as summarized in a review for sources of variations including *5, *7, *17, *18, *21, *23, *25, and *35 [56]. These genetic polymorphisms have a significant impact on drug metabolism. Since most drugs are not tested on diverse individuals this leads to limited efficacy and drug toxicity [57,58].

The African population variants have recently proved that they have an important role to play [59] in affecting the metabolism of antimalarial as well as antituberculosis. Tanner and colleagues [60] demonstrated the significance of African population variants in modifying the metabolism of medications like antituberculosis and antimalarials. CYP2A6*1B variations, where the active metabolite dihydroartemisinin was considered to be the reason, results indicated greater deleterious effects in prodrug artesunate [61]. Variations with the highest frequencies in the African population for CYP2A6 alleles (CYP2A6*17, CYP2A6*23, CYP2A6*25, and CYP2A6*28) metabolized antimalarial and anti-tuberculosis drugs [62]. CYP2A6*17 was found in African populations with 11.9% higher than the American and Asian populations with less than 0.6% [36,15]. According to Ho et al [62], CYP2A6*23 was only detected in the African population, with a frequency of 2%.

It seems the genetic variety and population success has been well investigated in European and Asian populations. There is an underrepresentation and lack of evidence concerning genetic

variations and population accomplishments among the African population [7] resulting in more vulnerability to adverse drug effects, rare and infectious diseases [8]. These previous studies have shown that populations have distinct variants; medications cannot be one size fits all.

Individualized treatment plans might enhance the effectiveness of treating illnesses in distinct populations. The global community could benefit from personalized healthcare if genetic differences in different population groups are better understood. This would bridge the gap in making knowledgeable decisions about the appropriate dosage and tailor-made therapies that can suit individual patients following their genetic makeup. This study's focus is on investigating how the novel CYP2A6 variants affect the structure and function of the CYP2A6 protein.

1.5 Problem Statement

Africa has been found to contain a higher diversity of CYP enzymes than the rest of the globe [63], with novel alleles, which is consistent with the continent's high genetic diversity [59,63-65]. Most therapeutic drugs have been tested on European populations but distinct populations have different allele variants. The high polymorphic nature of CYP2A6 and the rate of drug metabolism give different clinical consequences increasing the risk of adverse effects or death [66,67]. Variations have been reported to affect metabolic activity. Understanding how the variations affect the structure and function of the CYP2A6 gene is important.

1.6 Aims, Objectives and Hypothesis

1.6.1 Aims

The study aimed to critically examine the structural and functional implications of 13 CYP2A6 allele variations on CYP2A6 by carrying out Molecular Dynamics simulations (MDs) which reveal the structural dynamics to help us get more understanding of variation effects on the enzyme's behaviour. The aim was broken down into objectives to make it attainable.

1.6.2 Objectives:

- Template identification: Retrieval of CYP2A6 identifier protein from the RCSB PDB database (Chapter 2).

- Allele identification: Selecting CYP2A6 star alleles from the PharmVar database according to a set selection criterion (Chapter 2).
- Mutagenesis: Introducing each missense variation to the WT to generate 13 allele mutations using Discovery Studio Visualizer (Chapter 2).
- Metal parameterization: to protonate and assign heme and protein parameters to the AMBER force field (Chapter 3).
- MD simulation: to study the structural changes regarding the initial structure using GROMACS (Chapter 3).
- Conduct global analysis: to analyze the equilibration status of the protein using Root Mean Square Deviations (RMSD), the region of fluctuation in the protein with Root Mean Square Fluctuations (RMSF) and to measure compactness using radius of gyration (Rg) (Chapter 3).
- Clustering: to investigate conformational sampling of the variant and the reference CYP2A6 protein with AMBERTools (Chapter 3).
- Prediction of secondary structure: to investigate the changes to the secondary structure as a result of variant presence on the protein using a DSSP script (Chapter 3)

1.6.3 Knowledge gap and the Hypothesis

- Implications concerning the effects of specific missense variations in CYP2A6 drug metabolism are not yet understood.
- Missense variations affect the structure and function of the CYP2A6 enzyme. If variations are introduced to the reference structure, structural and functional changes are expected in the mutant structure.

CHAPTER 2: TEMPLATE SELECTION AND MUTAGENESIS

2.1 Introduction

This chapter covers how we identified the initial structure to use as a template in this study. Getting a good quality structure was more important for accurate Molecular Dynamics results which are interpretable and reproducible. The RCSB Protein Data Bank (PDB) database was used to get the initial structure. Next was the selection of the alleles from the PharmVar database and then finally the introduction of the variations to the reference structure using BIOVIA Discovery Studio. The results show the selected structure in 2D and the site for the variations introduced.

2.2 Methodology

2.2.1 Identifying the reference structure

A query was performed on the freely accessible PDB database using the UniProt accession number P11509 to find a list of structures similarly connected to the CYP2A6 enzyme [64]. The query generated a match of eleven structures: 2FDW, 1Z10, 1Z11, 2FDVU, 2FDV, 2FDY, 3EBS, 3T3Q, 3T3R, 4EJJ and 4RUI. The search was limited to scientific names: oxidoreductase, source of organisms to be *Homo Sapiens*, experimental method: x-ray diffraction, and a higher resolution (1.5 Å - 2.0 Å). Four structures out of eleven matched the search: 1Z10, 2FDU, 2FDV and 2FDY. Further analysis was performed to check for missing residues in the non-terminal regions using PyMOL (The PyMOL Molecular Graphics System, Version 1.2r3pre, Schrödinger, LLC) which is a common protein molecular viewer [68] .

The four structures were aligned in one PyMOL session and they all superimposed. Of the four structures, PDB ID: 2FDV had the highest resolution of 1.65 Å, and was selected for this study. Visual inspection of the template on PyMOL revealed missing residues on the N-terminal and the C-terminal and to validate this we checked the PDB file REMARKS section which showed the same missing residues. Missing residues can be a result of a disordered region that is highly flexible and mobile which gives a challenge in X-ray crystallography in determining the exact

positions of residues in the protein [7,69,70]. Since the missing residues were on the terminals there was no need to model the structure. A study on missing residues reported that the resolution of a protein structure is related to the percentage of missing residues, a resolution ranging between 1.50 Å - 1.75 Å was predicted to have ~5.5% missing residues [69,70] which in this case gives ~27 missing residues. Considering a standard error of ±2 this was in agreement with the 29 initial missing residues and four last residues for CYP2A6. The N-terminal and the C-terminal are anchored in the transmembrane and the structure is not regular. The next step was allele identification.

2.2.2 Allele identification

To make sure we identify the appropriate alleles from the PharmVar database that address our biological question and the African population alleles under investigation, stringent screening procedures were use [68]. Not every allele variation would result in noticeable changes to the structure and functionality of the CYP2A6 enzyme. The selection criteria disregarded deleterious genes, pseudogenes (non-functional genes), frame shifts (alters frame reading), 3′-Untranslated Region (3`UTR) conversions (changes the expression) and favored missense variations which result in decreased or increased activity [71,72]. Alleles that were still in preparation and variations not within range (residues 30-494) were not included in the selection. PharmVar is a curated database and also articles that present in vitro and in vivo experiments on the novel alleles serve as the validation for the identified alleles. Results of the selected alleles were presented in Table 2.2 with their respective enzymatic activities and the population frequencies.

2.2.3 Preparation of CYP2A6 allele structures

The amino acid modifications were added to the reference structure via site-directed mutagenesis using BIOVIA Discovery Studio 2021. The selected structure PDB ID: 2FDV was opened in Discovery studio and a hierarchical table with all the residues in the structure was selected. The residues of interest were selected using the Macromolecule tool and an option to choose the required residue. To mutate residues, the Build and Edit Protein tool was used to choose the required amino acid. Validation of mutagenesis was done by checking if the sequence has the

mutated amino acid at the correct position as well as the correct amino acid abbreviation. This was done for the alleles identified for the study.

2.3 Results and discussion

The study aimed to investigate how the missense variations affect the structure and function of CYP2A6. Table 2.1 presents the alleles that were not considered for this study because they did not have missense variations. Most of these alleles were in the 3`UTR conversion category.

TABLE 2.1

THE DISREGARDED ALLELES

Nature of allele	**CYP2A6 variants**
No function and deleterious	CYP2A6*2, CYP2A6*4, CYP2A6*5, CYP2A6*12, CYP2A6*20, CYP2A6*34
Frame shifts, 3'UTR conversion, and alleles in preparation	cyp2A6*5, CYP2A6*7, CYP2A6*8, CYP2A6*10, CYP2A6*19, CYP2A6*24, CYP2A6*28, CYP2A6*35, CYP2A6*36, CYP2A6*37, CYP2A6*36, CYP2A6*42, CYP2A6*46, CYP2A6*48, CYP2A6*49, CYP2A6*51, CYP2A6*52, CYP2A6*54, CYP2A6*56, CYPA6*50
Allele variation off-range	CYP2A6*13 (G5R), CYP2A6*14 (S29N), CYP2A6*31 (M6L)

From the search 13 alleles out of more than 43 alleles were identified, retrieved from PharmVar, and presented in Table 2.2. The following is a list of the selected alleles in Table 2.2: *6, *11, *16, *17, *18, *21, *22, *23, *25, *38, *40, *41, and *44. Two alleles (*16 and *21) have a normal activity and the rest have a decreased activity.

Alleles with a decreased function and only found in the African population were: CYP2A6*17 with the highest frequency of 7.3%-10.5%, CYP2A6*25, and CYP2A6*23 [36, 56, 65]. Other alleles found in the African population are CYP2A6*16 and CYP2A6*21, both with a normal function. Most of the alleles belong to other populations like Caucasians and Asians. The clinical consequences were based on nicotine metabolism experiments with substrates like coumarin

[73,74] so there might be a probability that the clinical consequences may differ with specific substrates [75].

TABLE 2.2

SELECTED CYP2A6 ALLELES

CYP2A6* Alleles	Amino acid variation	Clinical consequence	Frequency populations	References
CYP2A6*6	R128Q	decreased	0.4% Japanese	[76]
CYP2A6*11	S224P	decreased	0.5-0.7% Japanese and Korean	[77]
CYP2A6*16	R203S	normal	0.3-3.6% Caucasian	[78]
			0-1.7% African	
CYP2A6*17	V365M	decreased	7.3-10.5% African	[79]
CYP2A6*18	Y392F	decreased	0.3-2.3% Caucasian	[80]
			0.3-0.5% Korean	
CYP2A6*21	K476R	normal	0.5-2.3% Caucasian	[62,81]
			0.6-0.7% African	
CYP2A6*22	D158E, L160I	decreased	<0.3% Caucasian	[81]
CYP2A6*23	R203C	decreased	1-2% African only	[62]

CYP2A6*25	F118L	decreased	Only in African	[36]
CYP2A6*38	Y351H	decreased	European	[82]
CYP2A6*39	V68M	decreased	0.6% African	[83]
CYP2A6*40	I149M	decreased		[83]
CYP2A6*44	E390K N418D, E419D	decreased		[83]

The location of the variations introduced to the reference structure is shown in Figure 2.1. The variants F118L and K476R are in the catalytic site, R128Q, R203C, R203S, V365M and S224P are located close to the catalytic site and E390K, Y392F, V68M, Y476R, N418D, E419D, Y351H, D158E, L160I, I149M are around the outer active state region. The location of variations: R203C and R203S on helix F, S224P; helix G', D158E_L160I and I149M; D helix, Y392F; Beta sheet 1, V68M; Beta sheet 2, V365M; K K` loop and Y351H; K helix [50].

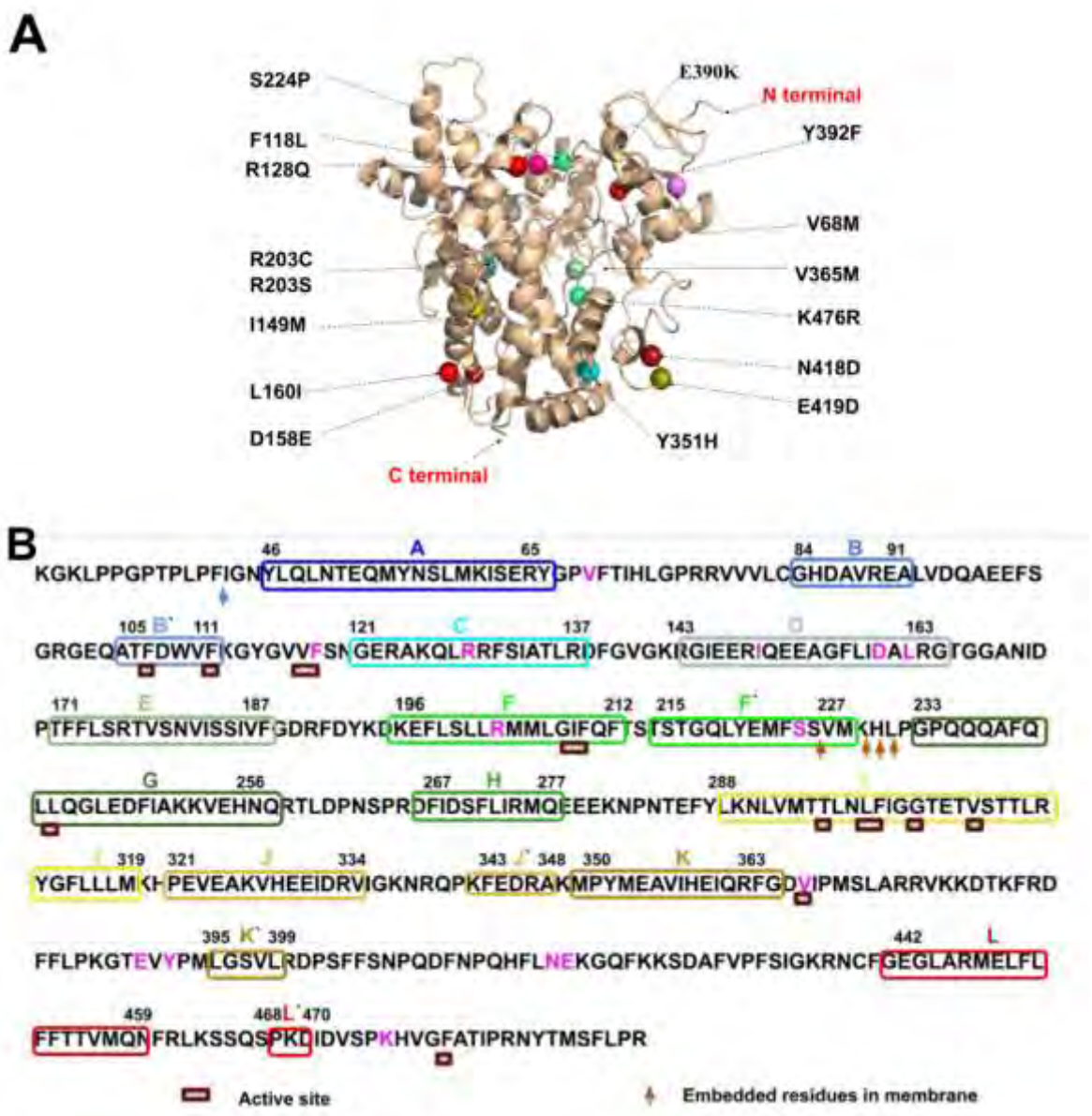


Figure 2.1. Location of CYP2A6 (PDB ID: 2FDV) variants. The colorful spheres mark the position of each variant on the structure (A). Residues that make the active site are in brown rectangles, letters A-L represent the helices' position in the sequence and the orange arrows show residues embedded in the membrane (B)

2.4 Expected reaction of variations with regards to amino acids properties

CYP2A6 ionizable residues can determine the protein structure, functions, and catalytic site by altering the stability, interactions, and solubility of the protein [84]. Amino acids have amino, carboxyl, and the R group (side chain) [85] except for Glycine (G) which is achiral. In this study we aimed at investigating how CYP2A6 missense variations R203C, R128Q, D158E_L160I, E390D, E419D, Y351H, K476R, S224P, R203S, Y392F, V68M, F118L and E390K_N418D_E419D affect the structure and function of CYP2A6. In the presence of variations, amino acids can result in alteration of physicochemical properties, size, charges, residue interactions, hydrogen bonding patterns, and protein stability [86].

The log of a negative Hydronium ion ($-H^+$) measures the hydrogen ion concentration and that is the pH [87]. The chemical structure of an acid determines its probability of being ionized and this differs across acids ranging from strong to weak. That measure of strength is its pK_a value, where *'K'* is the dissociation constant and 'a' is the molecule group. A higher pK_a value represents a weak acid and the lower the pK_a value the stronger the acid. There is a connection between pK_a and pH as illustrated in Equation 2.1:

$$\text{pH} = \text{pK}_a + \log \frac{[A-]}{[HA]}$$

Equation 2.1. The relationship between pK_a and pH demonstrated by the Henderson Hasselbalch equation [88, 89], where: $\text{pH} = -\log[H^+]$, $\text{pK}_a = -\log[k_a]$, $[A-]$: conjugate base, $[HA]$: acid concentration

An increase in pH increases the chances of acid deprotonating. A pH lower than the pK_a value will have a greater than half protonated acid (inequality (i)), if pK_a is less than the pH value then protonated acid will be less than half (inequality (ii)) and if the two are equal then it implies half protonated and half deprotonated (inequality (iii)). The relationship between pK_a and the pH has a great influence on the degree of deprotonation.

$$\text{pH} < \text{pK}_a \rightarrow \frac{1}{2} < \text{protonated acid (i)}$$

$$pH > pK_a \rightarrow \frac{1}{2} > \text{protonated acid (ii)}$$

$$pH = pK_a \rightarrow \text{protonated acid} = \frac{1}{2} = \text{deprotonated acid (iii)}$$

A pH shift, below or above 7.45 can impact the body's acid-base balance in patients which may cause metabolic disorders [90,91].

Hydrophilic amino acids Serine (S), Arginine (R), Aspartic acid (D), Glutamic acid (E), Histidine (H), Lysine (K), and Threonine (T) are expected to interact with solvents on the surface of the protein structure maintaining protein stability [92-94]. Impacts on the stability and function of the protein may be a result of surface alterations [95]. According to Spyrakis et al [96], water has many roles in protein dynamics which include hydrogen bonding being modulated by water to potentially accessible ranges. There may be possibilities of getting impacts on structure and function from variants R203C, R128Q, D158E, E390D, E419D, Y351H, K476R, and S224P which had a hydrophilic status before substitution.

On the other hand, hydrophobic residues are in charge of enzymatic reactions in the catalytic site and establishing a binding site for non-polar molecules [96,97]. In the core of a protein are hydrophobic residues, Glycine (G), Phenylalanine (F), Tyrosine (T), Valine (V), Alanine (A), Isoleucine (I), Methionine (M), and special cases Proline (P) and Cysteine (C) [98]. A possible impact is expected following the substitution of F118L located in the hydrophobic core and SRS1. But because they have similar properties, both are hydrophobic, non-polar and have a neutral charge, the substitution may be tolerated. Phenylalanine contains a phenyl group while Leucine has an isobutyl group and this suggests that both are less likely to form Hydrogen bonds but hydrophobic interactions are expected.

Protein stability is expected to be altered in R203C, Cysteine is more hydrophobic than Arginine which is involved in ionic interaction. The replacement with Cysteine can create hydrogen bonds because it contains a sulphyl group (-SH). Hydrogen bonding patterns are expected to change. In general variation R203C may affect the protein structure in terms of protein stability and folding.

In the substitution of R203S, R is positively charged and S is neutral it is reasonable to expect a charge loss because the substitution can alter substrate binding and enzymatic activity. While R

has a guanidium group that can form ionic contacts and hydrogen bonds, there may be changes to those connections since S has a hydroxyl group that forms hydrogen bonds.

Opening and closing confirmation on the substrate entrance F and G helices is mainly controlled by the physicochemical properties of residues and the structural orientation of the F` and G` region [99]. S224P is located in the F` helix, R203C, and R203S are in the F helix region and this suggests that these substitutions may affect the function and structure of the secondary structures at the substrate access channel. Interaction changes in the residues, stability of the protein, and clashes due to changes in size can be expected.

When Proline (P), which has a non-polar cyclic structure, replaces Serine (S224P) which is polar [100], hydrogen bonds may be destroyed, which may also have an impact on the stability of the protein. The substitution can bring a loss in hydrogen bonds suggesting that substrate interactions (recognition and binding) can be affected. S224P substitution also changes S from small to P medium size. Due to Proline's rigidity [86], changes in residue-residue interactions are expected to impact the protein conformation.

Glutamic acid and Lysine are some of the polar-charged amino acids important in keeping the protein stable [101]. Glutamic acid (E) substitution with Lysine (K) (E390K) located in the beta sheets surface may influence protein folding. Glutamic acid is negatively charged and can form hydrogen bonds while Lysine can form both hydrogen bonds and ionic interactions, a salt bridge is likely to be created. A change in the hydrogen patterns may disrupt the formation of a salt bridge. Glutamic acid is medium size and Lysine is large, the substitution may result in steric conflicts that can change protein conformations.

Glutamic acid (E) and Aspartate (D) (E390D) have the same physicochemical properties, both are hydrophilic, acidic, same size, and have the same charges. This suggests that the impact of this substitution may be tolerant. The same applies to K476R variation, Lysine (K) to Arginine (R) substitution they have similar properties, hydrophilic, basic, positively charged, and a small difference in the pKa (Lysine (~10.5), Arginine (~12.5)). The substitution may not impact many changes in the structure and function of the protein. However, Arginine can form more hydrogen bonds because of the guanidino and the carboxyl group than Lysine with -NH2 group [102].

Valine to Methionine substitution V365M (SRS5) and V68M (beta sheets surface) is classified as a radical replacement. Methionine is a large amino acid with sulfur in the side chain while Valine is a medium size aliphatic amino acid and both are hydrophobic and have neutral charges. The size change may denature the protein structure and with similar properties, this may make the substitution withstand the consequences.

Amino acids influence the preference of effective interactions needed for the protein to function properly [103]. Certain amino acids exhibit a like towards a specific secondary structure and this excludes Arginine, Glutamate, and Lysine which appear stable in most conformations. Most Glutamate and Alanine molecules are helices than any other type of secondary structure, Threonine, Cysteine, Phenylalanine, and Valine favor strands while Glycine, Proline, and Serine have a significant preference for turns or coils [104].

2.5 Conclusion

A total of 13 CYP2A6 alleles with designated clinical significance were identified from the PharmVar database. Of these, eleven were pathogenic, two were benign (R203S, K476R). Variations were then added to the reference structure CYP2A6 protein (PDB ID: 2FDV). The behavior of amino acid substitution was linked to their properties. Due to different physicochemical properties and functions in amino acids, substitutions can result in changes in the confirmation, substrate interactions, loss of charges, protein stability, hydrogen bonding, enzymatic activities, steric clashes and no significant changes if similar properties. Conclusively, amino acid properties play a major role in the effects of variations on the structure and function of a protein. Since CYP2A6 has a heme metal iron at its core, the next chapter looks at protonation, assignment of heme parameters to the force field, running Molecular Dynamics simulations (MDs) and the global analysis to address the research question.

CHAPTER 3: PARAMETER ASSIGNMENT AND MD SIMULATIONS

3.1 Introduction

Previous investigations have demonstrated that metal ions actively regulate the amount of heme present in cells [89]. This chapter explains the rationale for having metal parameterization, indicates the importance of setting up the correct parameters, and how force field parameters were set in this study. In this chapter, we show how AMBER techniques were used to pass parameters to protein, heme (HEM), and iron (Fe) centers. This step was required to maintain the Fe^{2+} in the heme in the catalytic site to fulfil the catalytic purpose. MDs technique was employed to investigate the changes in the structure and the conformations in CYP2A6 concerning the reference structure. MD simulations were done using the GROMACS protocol. Calculations for Root Mean Square Deviation (RMSD), Root Mean Square Fluctuations (RMSF), and Radius of Gyration (Rg) were done for the global analysis followed by clustering and prediction of secondary structures.

3.2 Assignment of heme parameters to the force field

3.2.1 The Amber Force Field

The term "force field" (FF) refers to a collection of mathematical equations (Hamiltonian) constructed from a variety of parameters to characterize the energy of the protein systems according to the location of its atoms in MD simulations [105-107]. Commonly utilized biomolecular FF include OPLS-AA, CHARMM, GROMOS, and AMBER [108]. Researchers regard Assisted Model Building with Energy Refinement (AMBER) as a dependable stable software that is powerful and distinctive [108]. This is because previous studies demonstrated that AMBER FF in MDs faithfully mimics the kinetic and structural characteristics of a wide range of canonical and non-canonical nucleic acids in water [109-111].

Wang et al [112] developed and experimented with General Amber FF (GAFF) which is an extension of Amber FF. Included in AMBER is GAFF and an antechamber tool enabling the

generation of AMBER FF template for any given molecule and the MCPB.py (Python-based metal parameter builder) [113] which allows for the development of reliable FF for simulating many different metal ions. This makes AMBER a powerful tool in pharmaceuticals with parameters for small molecules [113]. Quantum mechanics can be adjusted to determine bond distance [114], dihedral angles [115,116], partial charges [117], and van der Waals [118] parameters for individual atoms using metal properties [119,120]. AMBER FF14SB parameters are more adaptable to diverse protein simulations that incorporate heme proteins and are compatible with capturing the distinctive characteristics of heme proteins [112,121,122,105,106] compared to other FF.

3.2.2 The heme domain and the metal parameters

CYPs contain heme-b type which is a cytotoxic, hydrophobic, prosthetic element of proteins that function in xenobiotic metabolism, oxygen pathways, and preservation in CYP enzymes [123,125]. The heme group illustrated in Figure 3.1, presents a complex molecule with a porphyrin ring and an iron atom center found in CYP450 enzymes [126]. The catalytic process of CYP450 enzymes is influenced by the heme compound and a cysteine (CYS) residue which are crucial to the molecular makeup and operation of these enzymes. Varfaj's work [127], however, found an indirect connection between CYS residues and the heme binding. Fe is useful in systems of life because of its capacity to transition between the ferrous (Fe^{2+}) and ferric (Fe^{3+}) forms, while excessive levels can be adverse, [128]. The Fe^{2+} center coordination displayed by the CYP2A6 enzyme is deficiently described by the current AMBER FF parameters.

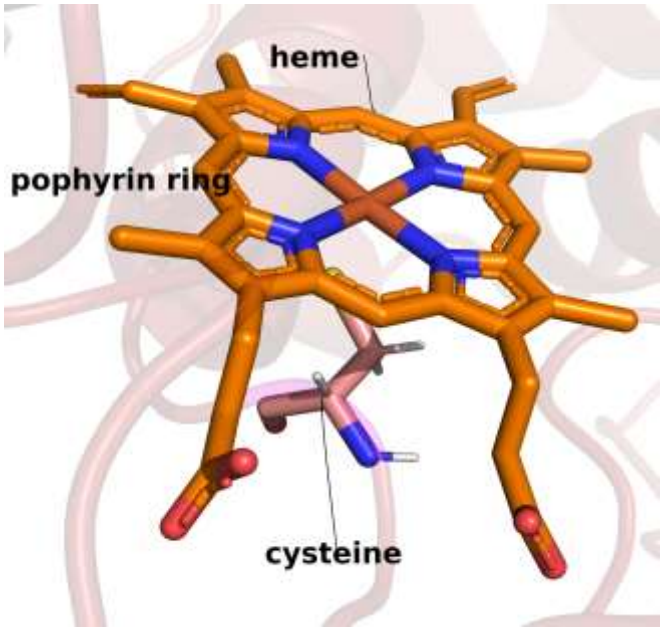


Figure 3.1. The structure of the heme. Illustrated is the pophyrin ring (orange), iron center (brown) is held by nitrogen atoms in blue. The CYS residue in sticks (pink) connection to the heme for the catalytic process.

Chebon-Bore and colleagues [129] investigated the effects of resistant variations on Cytb-ISP in the bilayer membrane. They validated the derived force field parameters for the heme using MD simulations and their results were consistent and reliable for usage in membrane MD simulations. In this current study, Fe^{2+} parameters were adopted from Chebon-Bore et al [129] discussed in the methodology section.

3.2.3 Methodology

3.2.3.1 Protein structure protonation

Figure 3.2 shows all steps followed in the protonation process. Coordinates of chain A, the heme, and iron (Fe) of the reference protein (PDB ID: 2FDV) for the selected 13 variations were separated into different files. The reduce command was used to remove hydrogens and the ATOM sequence of all protein structures was extracted and added to new folders. The Heme group was extracted and added to a new file. Fe was separated from the heme group and pasted into a new

file. The heme was protonated using the reduce command and added to a new folder. The protonated heme was cleaned and renumbered using pdb4amber

Protonation of the unconjugated protein was done using H++ software at pH 7.4 which is the average pH of blood in humans and mimics the effective environment for metabolic processes. H++ results gave the protonated protein coordinates (*.crd) and topologies (*.top) which were downloaded for further processing.

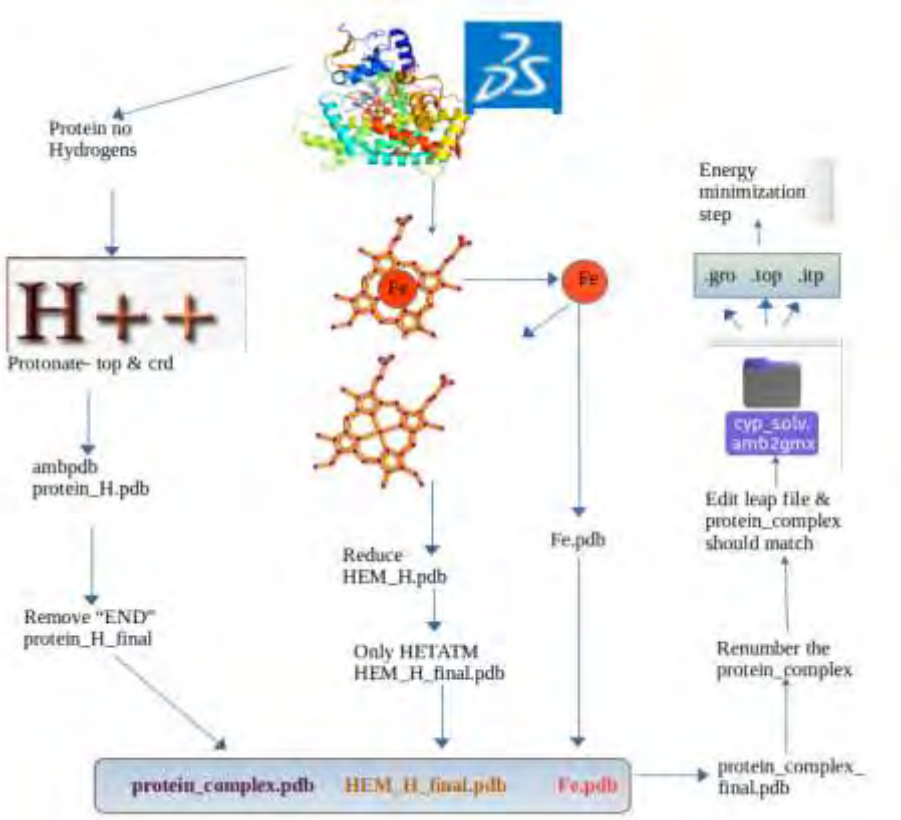


Figure 3.2. Metal parameterization flow diagram. The initial structure is prepared, protonated using H++ server at ph7.4 and force field parameters are fitted into the final protein complex. The parameters are validated using MD simulations.

Using the *ambpdb* command, the topology and coordinate protein files were concatenated into a single file. The line with "END" was removed from the protonated protein. The three files, protein complex, HEM, and Fe were concatenated into a new complex. This complex file was cleaned and renumbered using *pdb4amber* and added to a new final complex file. The parameterized residues

were given custom names edited as depicted in Table 3.1 to avoid confusing the names and prevent conflicts with the built-in FF so that they match with the *cyp_tleap.in* file parameter names. Residue numbers were renumbered to include the correct parameters in the leap file.

TABLE 3.1

RESIDUE CUSTOM NAMES

Residue	Custom name
CYS	CM1
HEM	HM1
Fe	FE1

Next was the solvation step to explain the integrity, sustainability, and electrostatics of the aqueous environment as well as the interactions between macromolecules [15]. The distance utilized to create a periodic border box has a significant impact on both the number of atoms and the simulation's duration [129]. The t-leap file was modified to include the final protein complex pdb file, correct coordinates of Fe and parameters, box shape: box, box distance: 12.0 Å [130], protein: ff14SB, heme: GAFF and water model: TIP3P.

AnteChamber Python Parser interface (ACPYPE) [131] tool was used to change the resulting system files to be compatible with GROMACS generating a *.gro file for the structure, *.top for the topology and *.itp. The mdp files were downloaded from GROMACS tutorial https://manual.gromacs.org/5.1/user-guide/mdp-options.html and edited to suit the required time of simulation and barostats (Berendsen and Parinello-Rahman) for the production run in MDs.

3.3 MD simulations

3.3.1 GROMACS-Based Molecular Dynamics

Research needs, user preferences, and the type of structure on investigation may control the

decision to choose between the software [132]. Although there are many open-source tools [133] for MD simulations, including AMBER, CHARMM (Chemistry at HARvard Molecular Mechanics), NAMD (NAnoscale Molecular Dynamics), the GROMACS (GROningen MAchine for Chemical Simulation) [134] software was our choice which best suite goals of this study. GROMACS is a highly preferred software involving all types of MD based on pair possibilities since calculations for non-bonded contacts may be performed with the selected MD tool [121-123]. GROMACS is typically employed on biological molecules with observable as well as complicated bonded interactions. GROMACS's ability to work with multiple FFs and proteins displayed its compatible relevance to this study [134].

3.3.2 Rationale for MD simulations

MDs are effective in modeling, binding free energy prediction, and providing detailed information on the structure and function of CYPs regardless of their ligands [137]. Previous studies demonstrated the applicability of MDs [138] and how critical they are to structural dynamic investigations for further understanding of protein conformational behavior, flexibility, interactions, and stability [138-142].

Following an extensive comprehension of the physics directing interatomic interactions, MDs are powerful in predicting the motion of each atom in a protein complex over time [140]. Based on interatomic connections, we use MDs to determine the atom's motion and behavior in time and space [140]. This method allows for the clarification of the protein's dynamics and stability. Its applicability to various systems from liquids to biomolecules makes them diverse in studying an appropriate analysis.

3.3.3 Steps in MDs

The preparation step, solvation was done in the previous chapter, where the topology, structure, and position restraint files were generated [132]. During solvation, there was an addition of solvent molecules to residues and that experience imposed unfavorable interconnections between molecules. In this chapter, energy minimization, equilibration then production run [143] were followed in the respective order. Energy minimization was done to relax the system stabilizing the protein. During energy minimization, the solute (protein) molecules are placed in the simulation

box but this placement can be randomly done causing instability of atoms. At this stage, the system is not stable and is not a natural representation anymore [144]. Atoms are capable of clashing resulting in bond distortions. To predict the arrangement of molecular atoms the total energy of the system is minimized. This step uses the steepest descent algorithm (gradient descent). A minimum point is selected, set initial iterative counting at zero (m=0), a gradient (steepest) is calculated at a point for each iteration (m), and then descend (step down the gradient opposing the direction) based on the step size [145].

During equilibration, the system needs to be brought to the required temperature [146] for a well-equilibrated starting point for the production MD simulation. This is done by continuously increasing the temperature of the system. The aim was to observe stable predictable properties across time, these include variations in energy, pressure, and temperature. This allows the system to adapt to the temperature change and prevents sudden shock to the system. Lastly is the production run.

3.3.4 Methodology

MDs were carried out utilizing the GROMACS 2018 version 6 program on the GPU cluster of the Center for High Performance Computing (CHPC), Cape Town, South Africa. A total of 13 proteins and the WT were simulated using uniform parameters to examine the effects of variations on CYP2A6 [147]. Minimization was done at an energy step size of 0.0001 and 50000 steps for each mutant and the holo WT up to a maximum force of 1000 kj/mol/nm until convergence. Convergence was verified, whether the gradient approached zero, if not converged the algorithm would repeat the process. The model was then simulated under desired conditions, constant temperature constant volume ensemble (NVT ensemble) [148] with a reference temperature of 300K for each group (protein, cysteine, iron, and heme) and constant temperature constant pressure (NPT ensemble) at 1 bar pressure [149]. This was to ensure that the solute molecules adjust to the new environment. The NPT ensemble production run used the leapfrog integrator, the neighbor searching algorithm, and Berendsen barostat [149] t-coupling to ensure that the system progressed toward a stable state. The systems had a rcoulomb cut-off of 1.2 nm and a rvdw cut-off of 1.2 nm. The MDs were done using different barostats at the production run (Berendsen and Parrinello-Rahman) to determine the standards of our results. MDs were run for 300ns, a total number of 150

000 000 steps and a time step of 0.002 ps with 8 core-nodes. A gmx trjconv command was used to center and remove jumps and rotations before making calculations.

3.3.5 Post MD local analysis of trajectories

GROMACS analysis tools gmx energy, gmx rmsf, gmx rms, and gmx gyrate were used for RMSD, RMSF, and Rg calculations [150] from the MD simulations to generate xvg files used to create graphs. Visualization tools; xmgrace, matplotlib for the line plots, seaborn for the violin plots, and the heatmap were used to create the graphs. Visual Molecular Dynamics (VMD) software was used to view the conformational changes for all the systems and as preparation for the analysis of the proteins after MD simulations [151].

To study the structural similarity between the holo WT and mutants, RMSD was used [152]. The calculated RMSD values measure the average distance between matching atoms in a simulated trajectory [153]. First was a check for consistency on the WT runs in both the Berendsen barostat and Parrinello barostat sample results. The line plots display how much the mutant system structure had deviated from the reference structure. Structural deviations were expected because of variations introduced to the structure as discussed in Chapter 2. Because line plots do not show the RMSD distributions, violin plots were utilized to exhibit the distribution of RMSD values and the summary statistics between the WT and the mutants [154]. These violin plots combine density charts which show each peak reflecting the frequency of the data points in each region, with a box plot inscribed in the peak to indicate the mean and maximum quartiles [155].

RMSF is how much a residue moves from the expected average thus the flexibility of a residue, how much it has moved from its initial position with regards to C-alpha [152]. It calculates a quantity's fluctuations around its mean value across an allocated duration of the simulation. An unstable residue has higher mobility and, thus flexibility and the RMSF value is high.

Clustering is an unsupervised machine learning model, in a molecular simulation context, it means grouping the data according to a similar neighborhood showing the same confirmations at a specified cut-off [156]. To resolve an imbalance in the data which usually causes analysis and interpretation of large systems like MD simulations hard to do, clustering was done. In clustering, different algorithms like K-Means [157], density based (DBSCAN) [158], and the hierarchical

agglomerative algorithm [159] can be used. In this study, clustering was done using AMBER tools, *cpptraj* script, and then aligning the clusters using PyMOL. The bottom-up approach (hierarchical agglomerative algorithm) [159,160] regards each data point as its cluster, clusters of shortest distance merge to form another cluster, repeating this to form an overall cluster with all observations. It uses an average-linkage between two clusters based on distance to create sub-clusters. Confirmations were set using a distance between the L helix (442-459), and the CYS438 and was set to get a default of three conformational clusters.

We hypothesized that there would be structural changes because of the variations. We used an important Bioinformatics tool, the DSSP program for predicting secondary structures [161] to create plots for visualizing secondary structural changes. DSSP states were assigned to the coordinates according to their probabilities (in percentages) of occurring at respective residues in the protein: Loop (L), 3_{10}-helix (G), helix-turn (T), α-helix (H), beta-sheet (extended (E)), beta bridge (B), bend (S) and the π-helix (I) [162,163]. DSSP allocates every residue to one out of the eight states largely mostly on hydrogen bonding arrangements and certain geometric restrictions. A DSSP script was used to run an interactive mode using gromacs 2023.2 and a csv file was generated to make graphs for the assigned structures.

3.3.6 Validation of parameters

It was important to test the success of the heme parameters assigned to the force fields which were aimed at holding Fe in the heme. To validate the heme parameters set, 100ns MDs were run for the reference structure using the GPU and CPU clusters. The purpose of this validation was to check if the heme stayed in place during the MD simulations before extending the time for MDs to 300ns. Results showed consistency in all runs however the GPU cluster was faster than the CPU cluster so an extension of the MDs to 300ns was done using the GPU discussed in the next section.

3.3.7 Barostats used at production run (Berendsen vs Parrinello-Rahman results)

A comparison of RMSD violin plots for Parrinello–Rahman (A) and Berendsen (B) results is illustrated in Figure 3.3.1. Both results show that WT runs were consistent for both Berendsen and Parinello throughout the simulation showing stability. Most systems like I149M, S224P, Y351H,

K476R, and F118L in Berendsen results exhibited different distributions across the respective runs compared to Parinello results which showed similar distributions in most systems. This suggested the inconsistency of Berendsen barostat findings. RMSD values for Parinello-Rahman systems ranged from 0.075nm (0.7 Å) to 0.25nm (0.2 Å) and was the same for Berendsen. Parrinello results were consistent compared to Berendsen results.

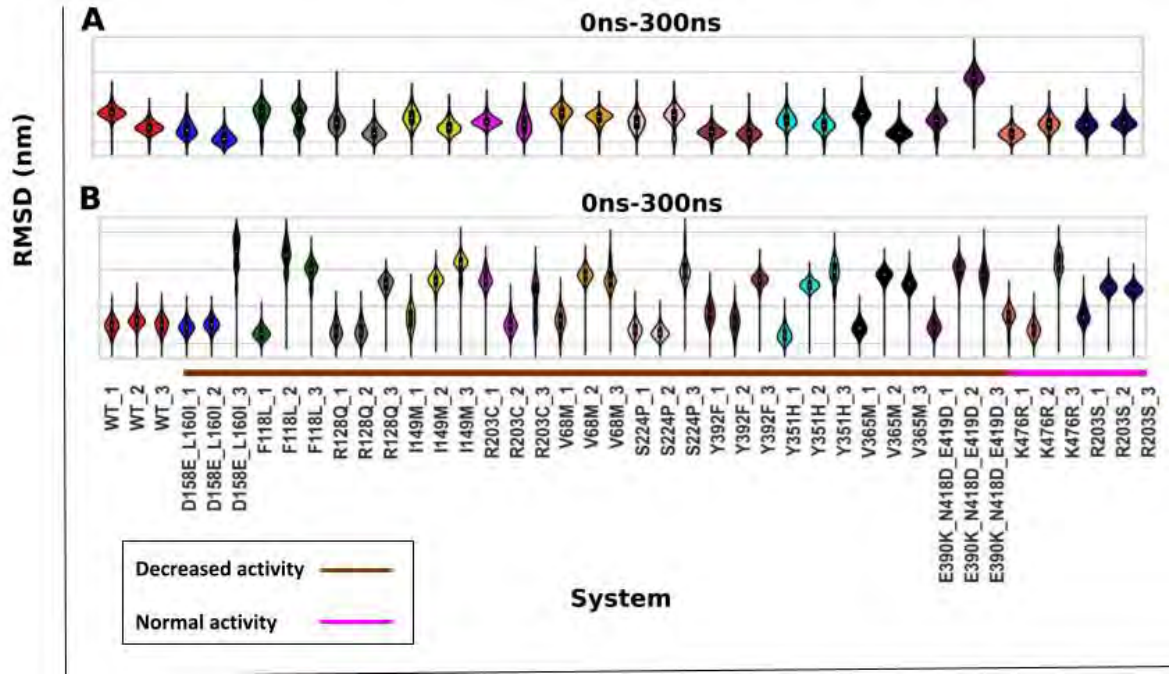


Figure 3.3.1. Comparison of RMSD distribution. Different barostats were used during the production run, Parinello-Rahman barostat (A) and Berendsen barostat (B). In both systems marked in red are the reference structures and the runs for each variant are presented in the same color. The clinical consequence for the variations is decreased activity (brown) and normal activity (magenta).

The heatmap in Figure 3.3.2 displayed information on the fluctuating regions. The Berendsen heat map (B) showed many darker regions which suggest more fluctuations as compared to the Parinello barostat results (A).

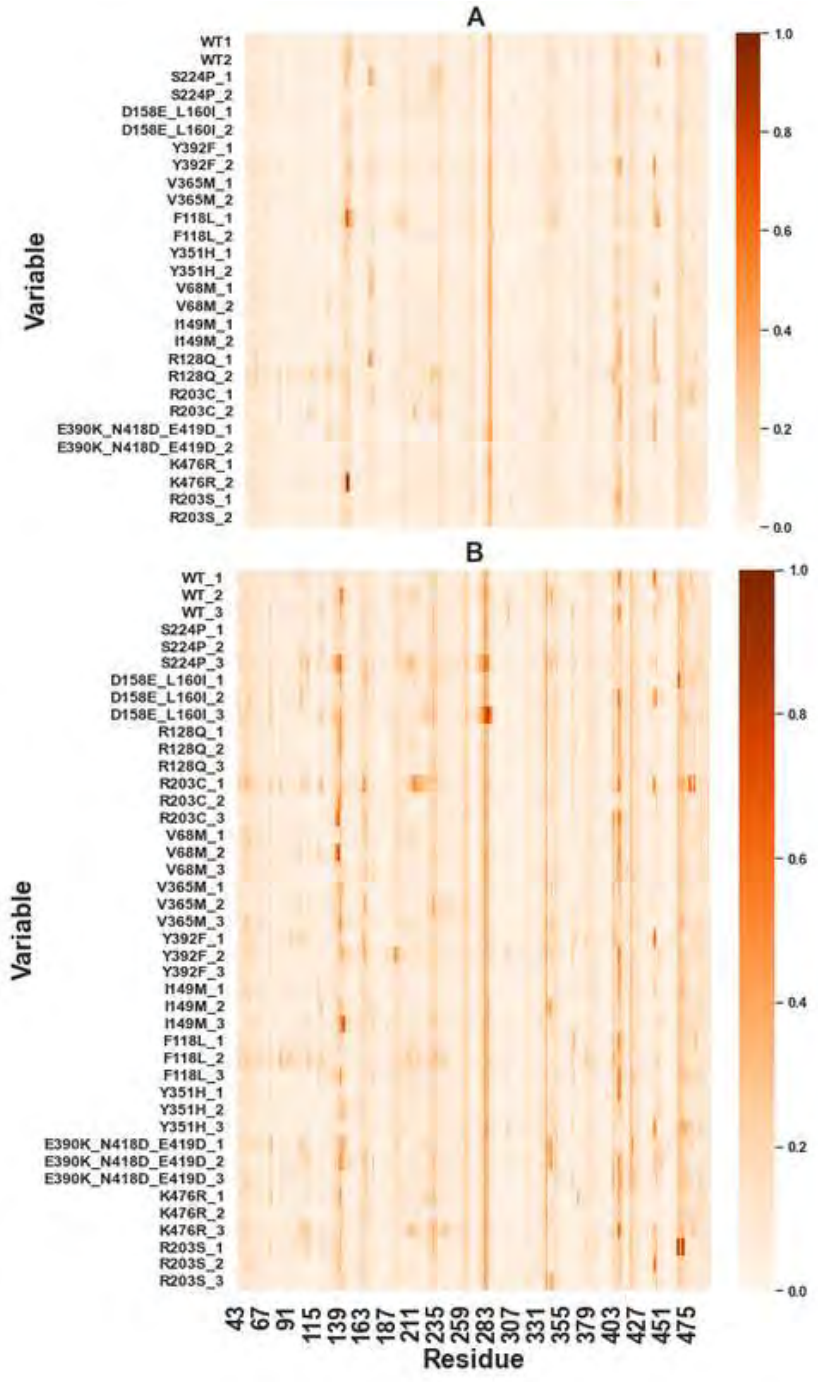


Figure 3.3.2. Normalized heatmap fluctuations at each residue. Berendsen at production run results (A). Parinello-Rahman at production run, RMSF results (B). The color scale is from 0-1 depicting that the darker the color the more fluctuating the residue is.

This suggested that Berendsen results in most regions were not stable. The C-D loop, H-I loop, J-K loop, and I-K loop were the most fluctuating in both barostat results as expected since loop regions have the highest mobility due to their biological roles. Measuring consistency, fluctuating regions are supposed to be uniform across all the runs for each system. However, it was observed that Berendsen heatmap showed more inconsistent fluctuations making it hard to analyze. The use of Berendsen barostat may be the reason for too many fluctuations in the systems.

Figure 3.3.3 shows the gyration plots illustrating the spread of mass distribution from the center of mass. From our results, both barostats indicate compact distributions which suggests stability in the protein structures. Rg values were small ranging from 2.26nm to 2.30nm indicating rigidity and that the protein structures were more folded and ordered. There were not many differences noted in Rg between the two barostats.

The comparison gave evidence that MD simulation results can be affected by simulation settings such as barostats [164]. In this study, using Berendsen at the production run was not appropriate as it affected the MD results and analysis negatively. The structural analysis is a bit biased because of the inconsistency of the results making reproducibility of these results unguaranteed. According to the GROMACS 5.1.1 guide on MD parameters, because it is a weak coupling method, Berendsen barostat is good at relaxing systems rapidly which makes it very effective in equilibration and not recommended for production run. No further analysis was done with the Berendsen barostat results in this study; instead, we continued with the global analysis of Parrinello-Rahman barostat [139] results.

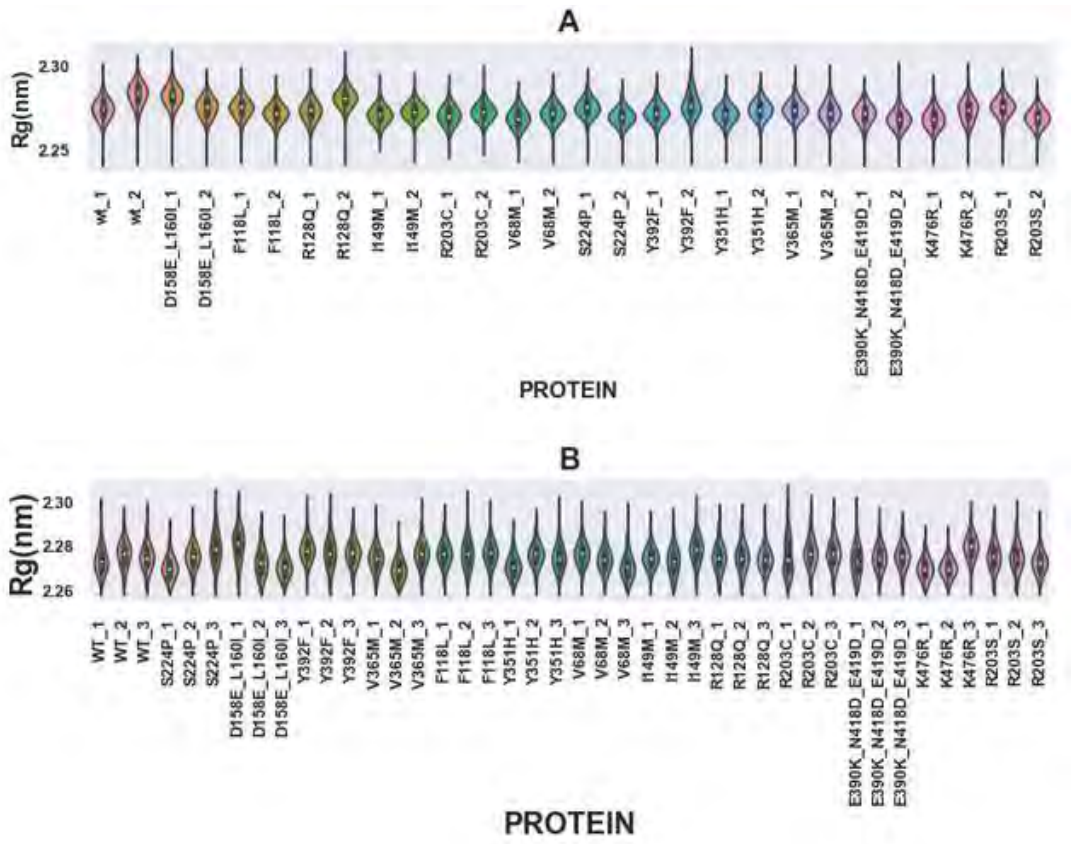


Figure 3.3.3. Rg plots comparison of Parinello-Rahman (A) and Berendsen barostat results (B). The plots show the compactness of the residues.

3.4 Results and Discussion

3.4.1 Structural alterations exhibited as a result of variations

3.4.1.1 RMSD analysis

Results in Figure 3.4 were grouped into three-time steps: 0ns-300ns, 50ns-300ns, and 150ns-300ns to see when the systems equilibrated since at the beginning of the simulation the system may not be stable. The analysis was based on 150ns-300ns as most systems converged after 150ns as depicted on the violin and line plots in the supplementary Figure S2 which was an indication that the systems had reached a stable state. Results also showed satisfactory parameterization that Fe^{2+} remained anchored in the heme in the catalytic site

during the entire MD simulation.

The line plots did not show much detail on how the RMSD values are distributed on each mutant system so to observe the small changes in the variant structures with time we analyzed the Kernel distribution of the RMSD values using the violin plots in Figure 3.4. RMSD values ranged from 1Å-2Å, the difference was not significant suggesting small changes in the variant structures. From the outcome, most variant systems remained in the same conformation throughout the MDs.

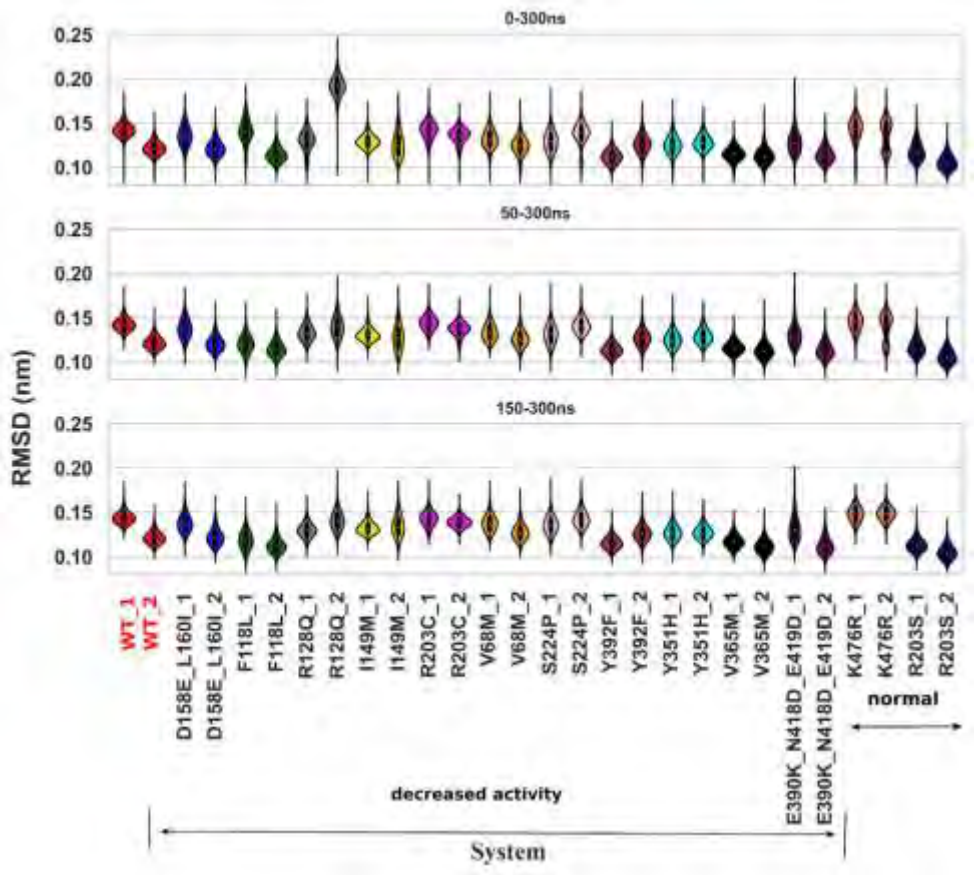


Figure 3.4. The Kernel distribution of RMSD values. The plots are for different time steps. In red are the reference structures and the variant replicate runs are in the same colors. The left arrow shows systems with decreased activity and on the right are the normal activity systems.

When compared to the reference structure, these variants displayed minor variations, either a slight increase or decrease in RMSD values and the mean distribution. Both the increased and decreased function systems were stable and the RMSD distributions were unimodal for all the systems after 150ns. R203S with a normal activity, located in the F helix's SRS1, results showed stability of the system and the RMSD distribution looked more similar to the WT system. This suggests that they may have the same function. This appears to be consistent with R203S in vivo experimental test results reported by Tong et al [165], that variant R203S had no alteration in the catalytic activity it had a normal function. Our results showed that K476R tagged normal activity, had the highest RMSD values across the whole system but they were stable. In vivo, results from a previous study [166], showed that nicotine metabolism decreased function systems however our results did not show similar results of normal activity as suggested. An explanation for this is that different substrates may have different catalytic activity, in this study we used a holo system which can give us different results. I149M and E390K_N418D_E419D had slightly different distributions.

All the systems were in one conformation except for K476R's second run with two bumps which disappear when it equilibrates after 150ns. Decreased function systems such as R203C and Y392F displayed similar RMSD distributions as the reference structure which contradicts the in vivo experiments where R203C and Y393F are said to have a decreased function and not a normal function. In general, RMSD values showed small structural deviations in the variant structure when compared with the reference structure. This was in agreement with a previous study by Kato et al. [48] who asserted that there were no major differences observed in a short time simulation but after 400 ns. It would be interesting to run MDs over 400ns to see if there would be any significant changes. RMSD alone gave information about the equilibration and RMSD values distribution, RMSF calculations allowed for analysis of the residues in fluctuation and the results are depicted in Figure 3.5 in the next section.

3.4.1.2 RMSF Analysis

Generally, residues in the regions 390-400 (K` region), 278-290 (H-I loop), and 135-145 (C-D loop) showed higher fluctuations (Figure 3.5) and this was expected as loops are always moving carrying out their biological roles. R128Q displayed more fluctuations around residue 50 and 120 when compared to all the systems suggesting that they were not stable. Other decreased function variations, F118L, and I149M also showed more flexibility in the residues. These fluctuations may have caused other alterations like shifts in other functional secondary structures. Normal activity R203S showed the least flexible regions as compared to the reference structure and other variants. Most of the decreased function systems showed more flexibility as compared to the normal function systems.

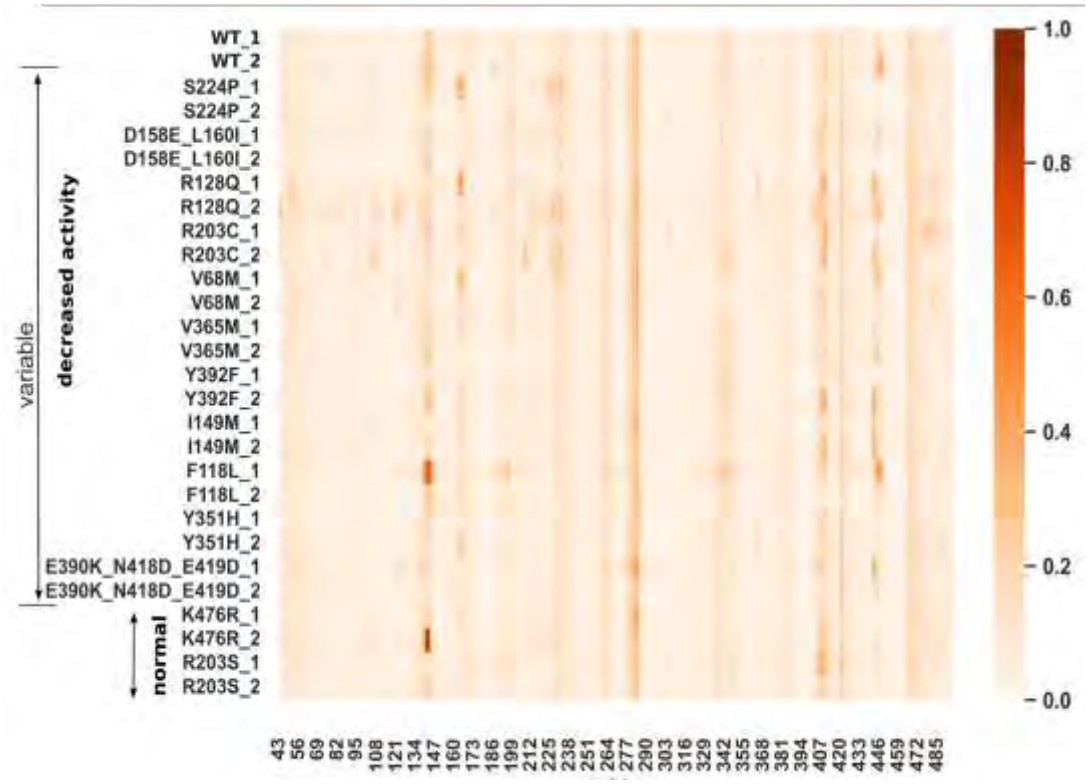


Figure 3.5**.** A normalized heatmap showing fluctuations in residues during MD simulations. The darker regions in orange color represent the most fluctuating residues concerning the system. The color scale measures how much the residue fluctuates.

Yes, RMSD violin plots showed unimodal distributions but clustering would confirm more about the fluctuations and conformations, to get true representatives of the allele variant systems.

3.4.2 Clustering Analysis

Results from clustering in Figure 3.6 showed three clusters for each of the systems, however quality analysis of results was important. Careful evaluation of each cluster percentage was considered, the other two clusters for each system had very few frames and the fraction percentages were very low as displayed in Table S1 (Supplementary material). It was not ethical to consider the other two clusters, rather we considered clusters with ≥ 30% and disregarded clusters with lower percentages. From our results, only one cluster dominated in each of the systems and was used as the representative cluster. The WTs were aligned and results showed an RMSD of 1.248 Å showing no significant differences and suggesting that the conformation was similar. After aligning the variant clusters with the WTs, the RMSD values ranged from 0.968Å - 1.801Å as in the supplementary material Table S2, which showed little difference.

Results depicted in Figure 3.6, decreased activity systems I149M (A), F118L (C), E390K_418D_419D (D), and normal activity variation K476R (B) showed some differences in the representative cluster structures compared to the WT. These differences were on the C-D loops, H-I loops, K' helix, and shifts in the F and G helices as illustrated in Figure 3.6. Variation E390K_N418D_E419D showed differences in the C-D loop and H-I loop and shifts in the 'B' helix. The C-D loop in variations N418D and E419D promotes a structural alteration in the substrate binding location and the substrate entry channel [48].

Shifts were noticed on K476R on F and, G helices and the H-I loop. Residue F118L is one of the residues that form the catalytic site and I149M is close to the catalytic site; they both had shifts on the F and G helices. Flexibility on these helices which are connected to the F` and G` helices allows the opening and closing of the substrate entrance allowing movements that may end up in shifts on the helices. This was also noted in a previous related study where F118L was said to have side chain shifts away from the heme during MD simulation, having an impact on secondary structure formation and interactions with substrates and heme [48]. Most of the aligned structures (Figure 3.7) had no significant differences in structure regarding the WT.

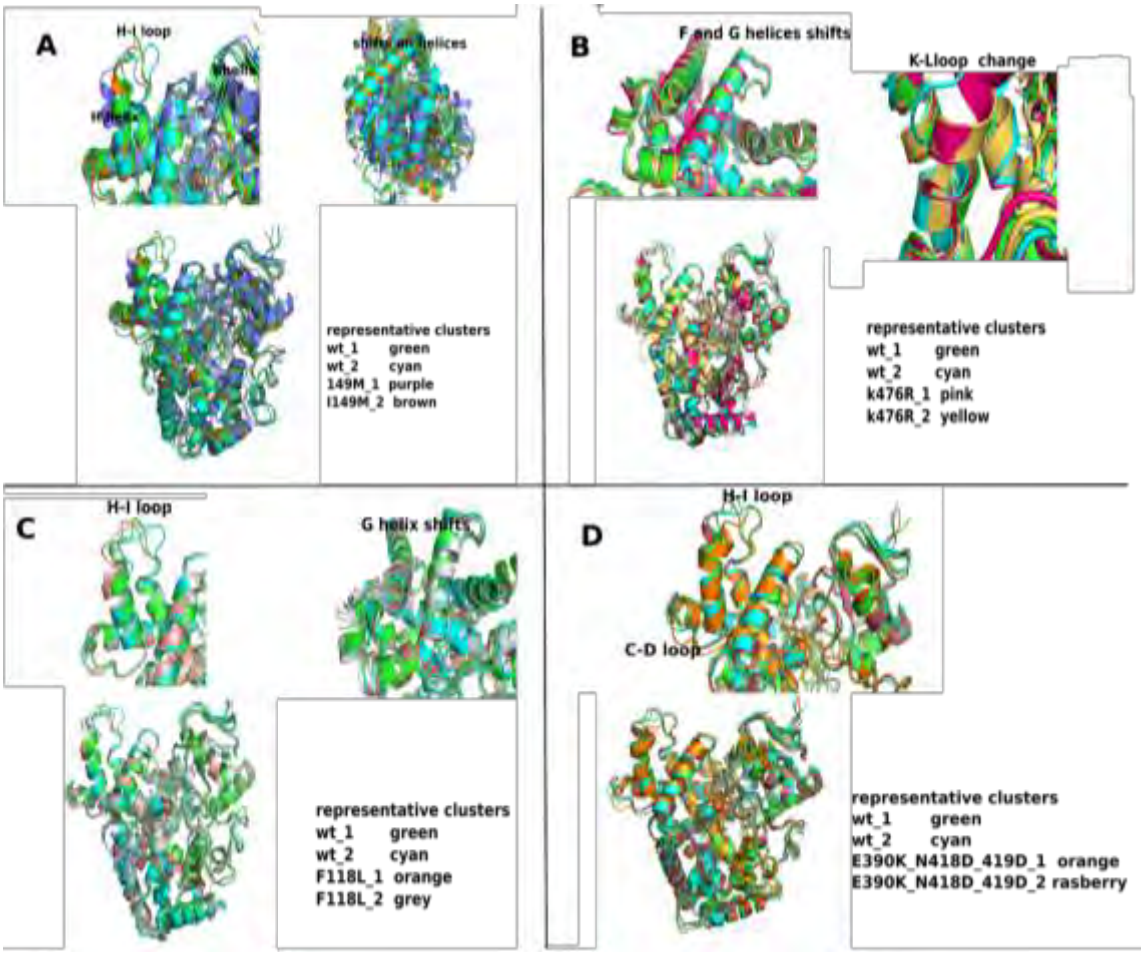


Figure 3.6. Representative clusters that showed differences. Reference structures (green and cyan) and the variants were aligned using PyMoL to show the differences in structures I149M, F118L, E390K_N418D_E419D and K476R.

Conclusive to clustering results all the systems were in one conformation; a few systems showed some differences in structure while in most systems there were no significant differences noted. However, no matter how slight a change in the protein structure is, it may also result in modifications to the catalytic site [13] and other secondary structures.

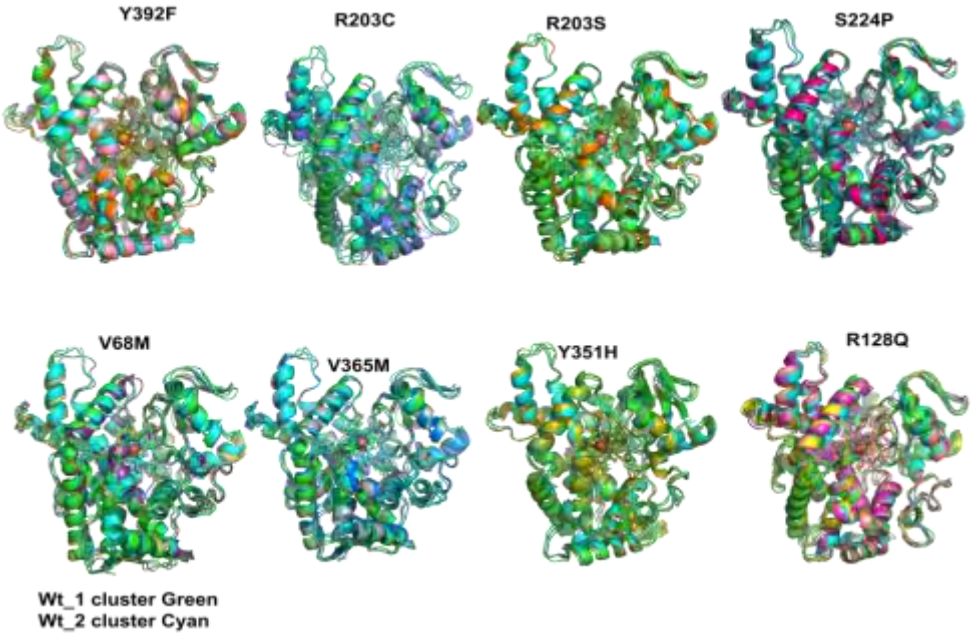
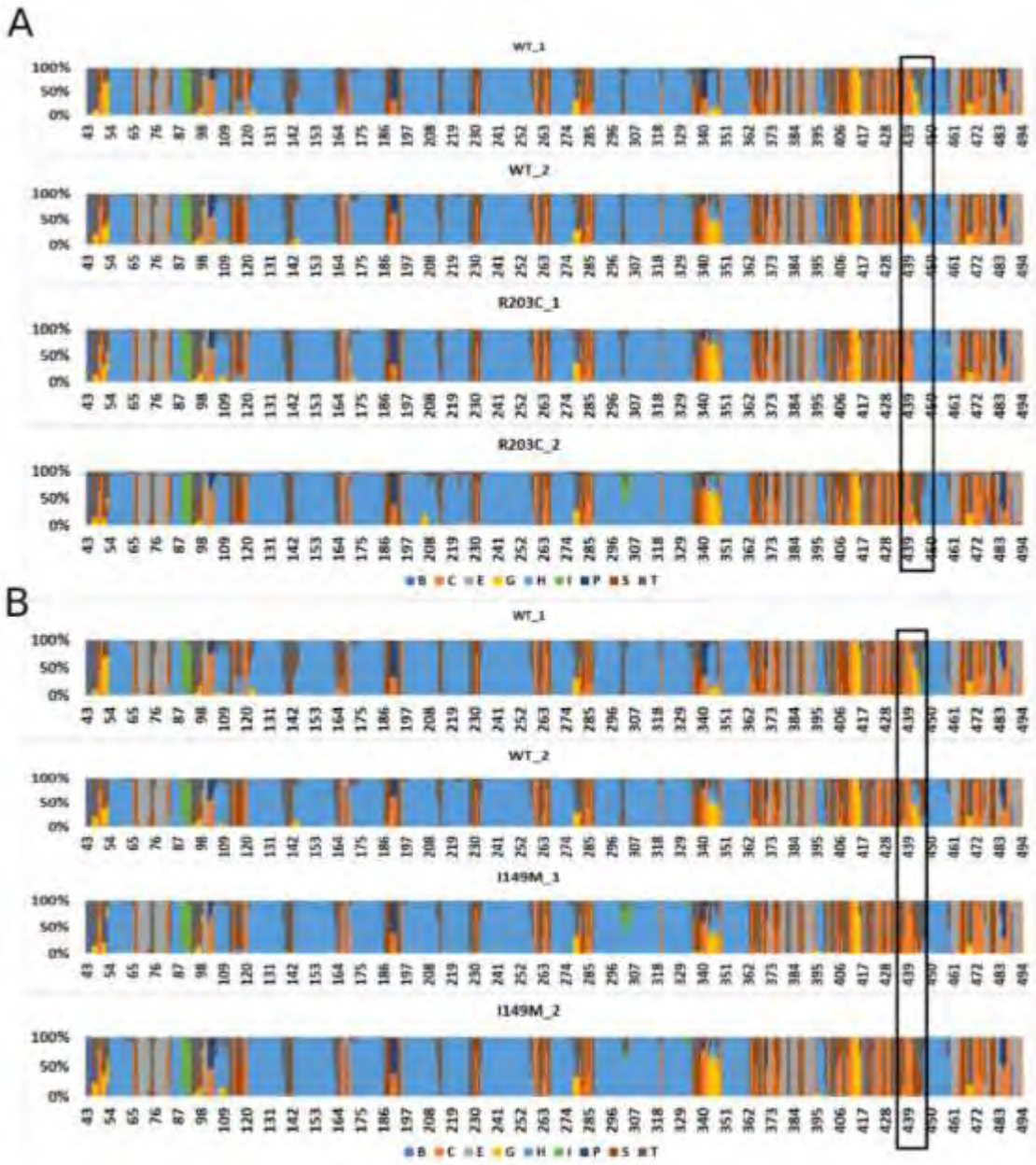
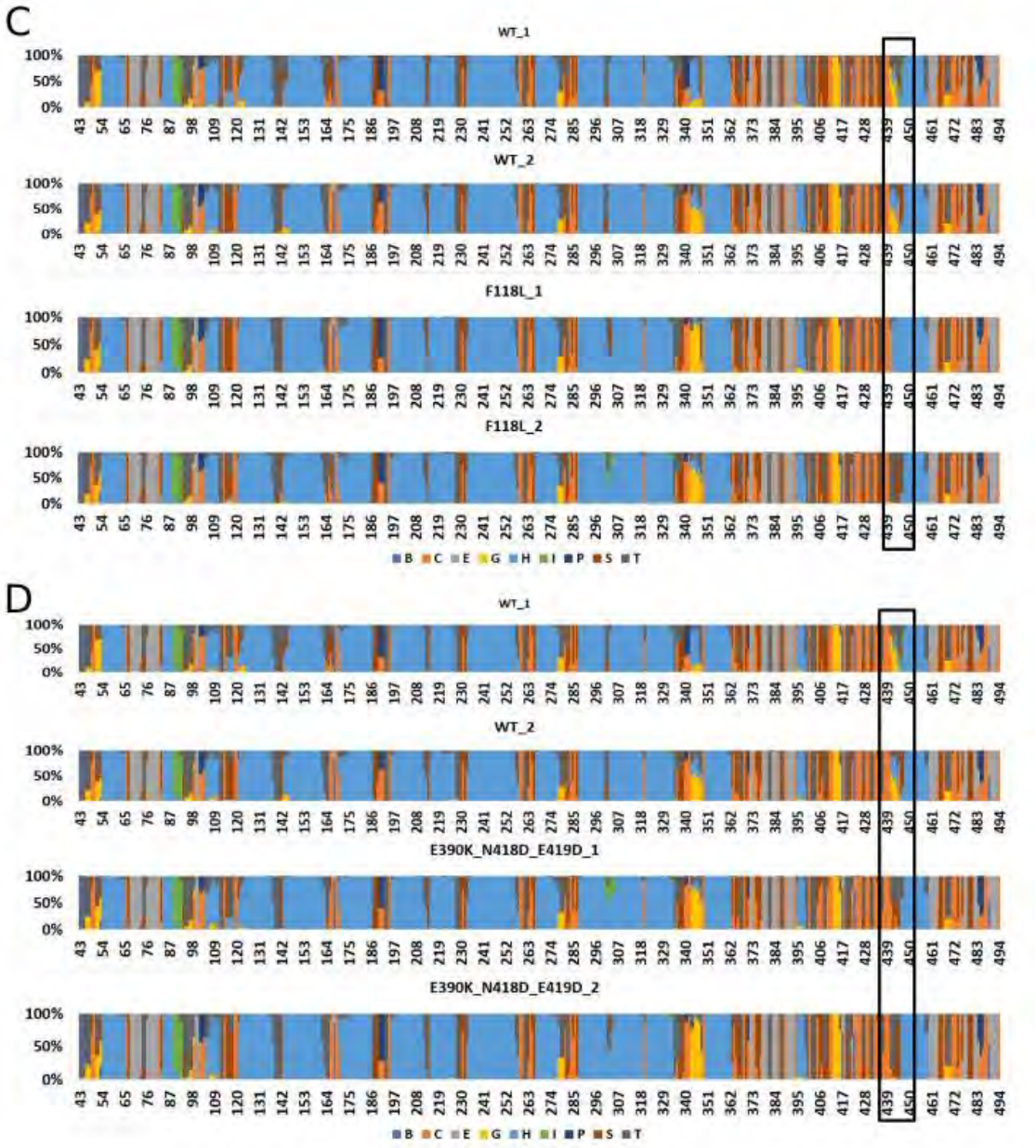


Figure 3.7. Representative clusters with no significant differences. Superimposed variants with the reference structures, Wt_1 and Wt_2 are in green and cyan for all the systems.

3.4.3 Structural change using Define Secondary Structure Prediction (DSSP)

Comparisons of variants against the reference structural elements were done for all 13 systems and results are displayed in Figure 3.8 (A-E): R203C, I149M, F118L, E390K_N418D_E419D, K476R, and those with no significant differences are in the supplementary material Figures S5 – S12.


A
WT_1
WT_2
R203C_1
R203C_2
B
WT_1
WT_2
I149M_1
I149M_2
100%
50%
0%
B C E G H I P S T


C
WT_1
WT_2
F118L_1
F118L_2
100%
50%
0%
43 54 65 76 87 98 109 120 131 142 153 164 175 186 197 208 219 230 241 252 263 274 285 296 307 318 329 340 351 362 373 384 395 406 417 428 439 450 461 472 483 494
B C E G H I P S T
D
WT_1
WT_2
E390K_N418D_E419D_1
E390K_N418D_E419D_2
100%
50%
0%
43 54 65 76 87 98 109 120 131 142 153 164 175 186 197 208 219 230 241 252 263 274 285 296 307 318 329 340 351 362 373 384 395 406 417 428 439 450 461 472 483 494
B C E G H I P S T

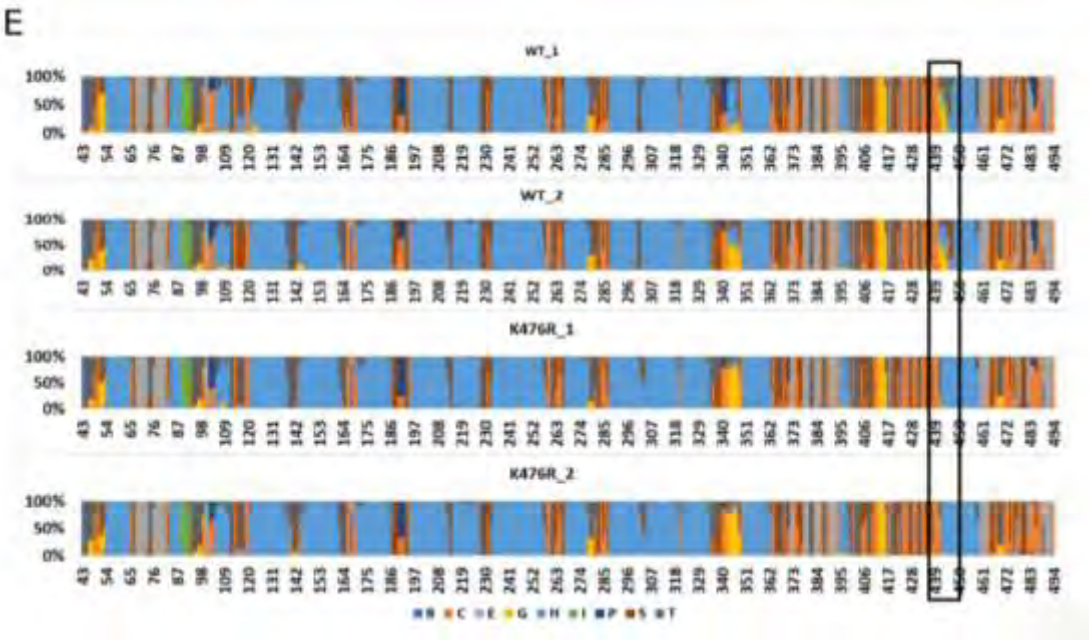


Figure 3.8. Differences in secondary structures. The reference structures (WT_1, WT_2) and the variants R203C, I149M, F118L, E390K_N418D_E419D, and K476R are presented in plots A-E respectively. The colors show secondary structure conformations at each residue (x-axis) and the percentages (y-axis) depict the probability of a secondary structure forming at a specific residue. On the legends, the letters represent amino acid states, B (residue in isolate), C (beta bridge), E (extended strand), G (3_{10} helix), H (alpha helix), I (pi helix), P (polyproline helix), S (bend) and T (hydrogen bonded turn). The black rectangle highlights the region with differences.

Our results show that the reference structure at residues around 439-445 (L helix and KL loop region) there was a secondary structure conformation of ~50% 3_{10} helix (yellow), ~15% hydrogen bond (jet black), and ~35% alpha helix (sky blue). On the same region in variations A-E (Figure 3.8), there was a different orientation in the secondary structure conformations. The variant structures showed a loss of the 3_{10} helix. The substitution of Arginine with Cysteine, R203C, as predicted in Chapter 2 can result in structural changes, agreed with the DSSP analysis. The 3_{10} helix could have disrupted the likelihood of Cysteine forming a hydrogen bond resulting in an altered hydrogen bonding pattern thereby impacting a structural change and loss of the 3_{10} helix.

Variant I149M (Figure 3.8 B) had a higher percentage of a hydrogen bond turn forming (~70%) and at some point, the secondary structure existed as a bend (~30%). Again, there was a loss in the 3_{10} helix and the alpha helix which existed in the reference structure. F118L displayed a higher percent of an alpha helix and a loss in the 3_{10} helix. In variant E390K_N418D_E419D there was a loss of the 3_{10} helix and it showed more of a hydrogen-bonded turn. The presence of a Cysteine residue (439) might be the reason for a hydrogen bond turn formed in these variants. Besides Cysteine being polar, and having the ability to form hydrogen bonds, its small size and flexibility properties contribute to a high probability of a turn conformation. On the sequence, on residues 438-443 (NCFGEG), there are turn-loving residues Asparagine (N), Cysteine (C) and Glycine (G) which increase the likelihood of a structure conformation to a hydrogen bond turn.

In the same region the variant K476R with normal activity and close to the catalytic site, presented in Figure 3.8 E displayed a very high percent of an alpha helix (blue) dominating in that region and there was a loss of the 3_{10} helix and conformation to a hydrogen-bonded turn (grey). As predicted in Chapter 2, Arginine substitution can form hydrogen bonds which can participate in the turn formation by imposing rigidness while stabilizing the turn conformation. A related study demonstrated that variation CYP2A6*21: K476R has a minimum effect on the catalytic functions [167] and their results showed non-significant structural changes in K476R variation and a loss of H bond.

All five systems have a decreased activity except for K476R with a normal function. The location of the five systems: I149M (D helix, SRS1), R203C (F helix, SRS2), F118L (B`C loop, SRS1), E390K_N418D_E419D (K`L loop, beta-sheet surface) and K476R (L` loop) play a role in the successful functioning of the protein. The B`C loop is close to the F F` loop and this region opens and closes the access channel for the substrates to interact with the catalytic site. Around the B`C loop, the F and FF` loop could be more channels since the region is part of SRS1 and SRS2. The DSSP approach was in agreement with the results from the heatmap (Figure 3.9) where these five systems showed higher fluctuations in that same region. This suggests that residues around 439-443 were not stable. The fluctuations might have been caused by the movements at the access channel opening or closing orientation impacting the variations located in such regions. The region which showed differences is a section of the L helix and I loop that borders the catalytic site and

contains the oxygen binding motif. F118L is in the catalytic site, K476R is close to the catalytic site and all the other systems are scattered around outside the catalytic site.

There may be something common linking these five variations. We need to understand more about the changes exhibited by the five systems on the access channels for this protein. The secondary structure prediction has shown structural changes but there is a need to further investigate these five variants at residue level.

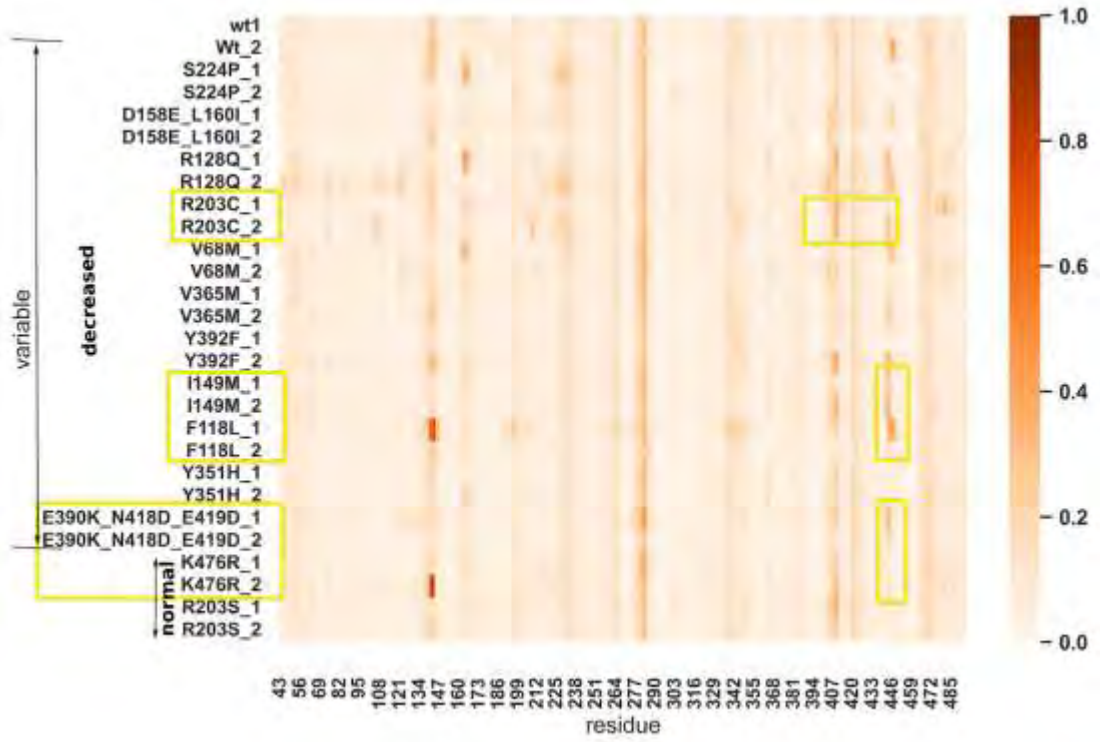


Figure 3.9. The fluctuations exhibited by the five systems with secondary structural changes from DSSP results. The yellow rectangles highlight the variation and the region of fluctuation.

3.4.4 Radius of Gyration (Rg) Analysis

The Radius of Gyration measures the distance between the atoms to their center of mass [167]. Assessing how a protein structure is tightly confined (compactness) or how loose it is in a three-dimensional arrangement during MD simulation defines Rg. A conformational change [168] can give a change in Rg distance after ligand binding [169]. The results for Rg are shown in Figure

3.9.1. The findings demonstrate the compactness of the residues. From the violin plots the Kernel distributions look similar to the wild types. None of the Rg distances exceeded 2nm showing little changes in the Rg distance therefore there may not be much conformational changes in the structures, maybe because there is no ligand binding to the protein.

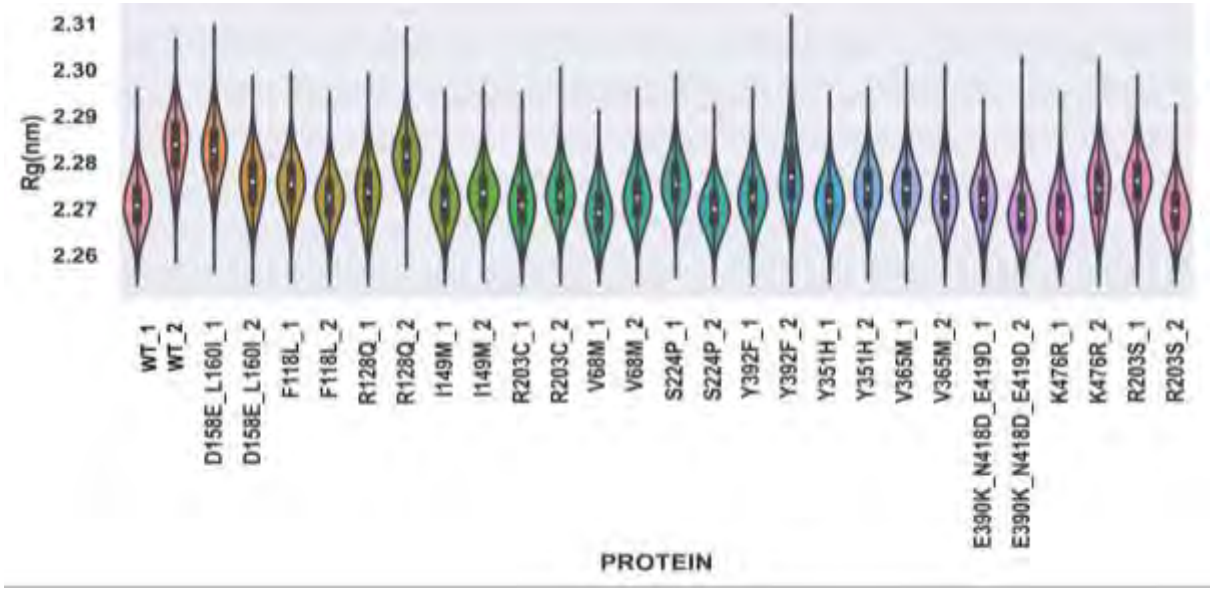


Figure 3.9.1. Radius of gyration violin plots showing the compactness of the residues

3.5 Conclusion

Parameterization was done to include the parameters for both the protein and the heme in the FF so that the heme does not leave the protein during MDs. The protonation condition of the protein, the FF, parameters utilized, and equilibration of the system have a significant impact on the quality of MDs. MD simulations were successful in providing information on the global structural alterations, how packed the residues are, and their flexibility. MD simulation also validated our parameters as the heme remained intact in the catalytic site. RMSD results showed very small deviations between the reference structure and most mutant systems. Loop regions showed the highest fluctuations as expected. All the mutants were in one conformation and showed a few differences after clustering. DSSP results revealed changes in secondary structure in residues close

to the cysteine pocket. The global analysis did not give the functional effects, there is need for a further investigation at the residue level (local analysis) for the five systems that showed structural differences.

CHAPTER 4: CONCLUSION

4.1 General conclusion and forthcoming initiatives

This study aimed to investigate the effects of variations in the CYP2A6 allele focusing on the protein structure and function. The study hypothesized that the presence of the variations would cause structural changes in the CYP2A6. Little is understood about how CYP2A6 missense variations affect the structure and function of the CYP2A6 enzyme. There is also inadequate knowledge of how Afrocentric missense variations impact the structure of CYP2A6 which is mainly involved in the catalysis of drugs. The research was worth studying as CYP2A6‘s polymorphic state is known to affect the metabolism of drugs such as cancer, malaria, and TB which may cause side effects due to the rate of enzymatic activity differing from each individual and each population[77, 170-172]. It would greatly give more understanding of conformational dynamics and the functions of CYP2A6 so that therapies are tailor-made to reduce the risk of drug effects. Conformational dynamics in CYP2A6 are linked to personalized medicine in that the variations in CYP2A6 activity affect the rate of metabolism differently in individuals because of distinct genetic profiles that influence therapeutic responses. This helps the individuals get the right dose specifically following the rate of metabolism for effective therapy.

4.1.2 Chapter 2

Chapter 2 was a search for a good quality reference structure and selected PDB ID: 2FDV. A list of 13 missense variations was accessed from PharmVar where two of them (K476R and R203S) had a normal function and the rest had a decreased function. With BIOVIA, mutagenesis of 13 allele variants was constructed.

4.1.3 Chapter 3

In Chapter 3, the main idea was to assign heme and protein parameters to the FF where all the systems were protonated and mimicked a human CYP enzyme environment important for the MD simulations to get reliable results. MD simulations run for 300ns were used to observe the structural changes. At the production run, Berendsen and Parinello-Rahman parameters were used separately on each system. The Berendsen results gave unstable results as reflected on RMSD

violin plots in Figure 3.3.1 with a weird third or second run. Going forth, Parinello-Rahman's results were used for analysis. RMSD, RMSF, and Rg were calculated to check for deviations, fluctuations, and compactness respectively comparing the reference structure to each of the 13 systems.

4.1.4 Main Outcomes

Outcomes from MD simulations may help predict modifications to structure in CYP2A6 mutants and offer important insights into pharmacogenetics. RMSD results in general did not show many deviations between the reference structure and the mutant structures except for a few. Fluctuations were noted on loop regions which was expected and five systems had higher fluctuations around residues 439-443 suggesting instability. Variations might have caused the fluctuations noted on other residues. Clustering results showed one cluster for each mutant and changes were noted mainly on the GI and HI loop and shifts on the C, F, and G helices. Clustering did not give much detail on structural transition as compared to DSSP analysis for secondary structure prediction. Variations at K476R, F118L, I149M E390K_N418D_E419D, and R203C may have affected the L helix (442-459) and K-L loop (432-441) secondary structure. The L helix is very important in maintaining the stability of the structure and variations can alter and destabilize the structure. According to our results, we can report that the variations may have an impact on the structure of CYP2A6 as hypothesized in this study. Further research is needed to investigate if there is any link between residue interactions for the five systems which exhibited structural changes.

This study was able to unveil structural changes imposed by variations on CYP2A6 and no matter how small they are, they can have a big impact on the catalytic site. From our findings the five systems residue interactions need to be investigated, why they behaved the same in the same region. Further investigation on variation impacts on CYP2A6 function could give a deeper clarification on the residue interactions during MD simulation in the catalytic site. To provide enough information on the effect of variations on the function of the CYP2A6 gene, the Dynamic Residue Network (DRN) approach [173] is recommended for future study. In the future MDs might need to be extended to 400 or 500 ns to investigate any significant change.

References

[1] M. Hasanzad, G. Patrinos, N. Sarhangi, B. Sarrami, and B. Larijani, 'Using ChatGPT to Predict the Future of Personalized Medicine', In Review, preprint, Apr. 2023. doi: 10.21203/rs.3.rs-2799531/v1.

[2] *Clinical Biochemistry: Metabolic and Clinical Aspects*. Elsevier, 2014. doi: 10.1016/C2011-0-07946-2.

[3] *Nonclinical Development of Novel Biologics, Biosimilars, Vaccines and Specialty Biologics*. Elsevier, 2013. doi: 10.1016/C2011-0-07530-0.

[4] F. J. Gonzalez and D. W. Nebert, 'Evolution of the P450 gene superfamily': *Trends in Genetics*, vol. 6, pp. 182–186, 1990, doi: 10.1016/0168-9525(90)90174-5.

[5] D. W. Nebert and T. P. Dalton, 'The role of cytochrome P450 enzymes in endogenous signalling pathways and environmental carcinogenesis', *Nat Rev Cancer*, vol. 6, no. 12, pp. 947–960, Dec. 2006, doi: 10.1038/nrc2015.

[6] E. Topic, 'The Role of Pharmacogenetics in Management of Cardiovascular Disease', *EJIFCC*, vol. 14, no. 2, pp. 78–88, Jul. 2003.

[7] G. Sirugo, S. M. Williams, and S. A. Tishkoff, 'The Missing Diversity in Human Genetic Studies', *Cell*, vol. 177, no. 1, pp. 26–31, Mar. 2019, doi: 10.1016/j.cell.2019.02.048.

[8] R. A. Wilke *et al.*, 'Identifying genetic risk factors for serious adverse drug reactions: current progress and challenges', *Nat Rev Drug Discov*, vol. 6, no. 11, pp. 904–916, Nov. 2007, doi: 10.1038/nrd2423.

[9] R. Weinshilboum, 'Inheritance and Drug Response', *N Engl J Med*, vol. 348, no. 6, pp. 529–537, Feb. 2003, doi: 10.1056/NEJMra020021.

[10] A. G. Madian, H. E. Wheeler, R. B. Jones, and M. E. Dolan, 'Relating human genetic variation to variation in drug responses', *Trends Genet*, vol. 28, no. 10, pp. 487–495, Oct. 2012, doi: 10.1016/j.tig.2012.06.008.

[11] M. Dolsten and M. Søgaard, ‘Precision medicine: an approach to R&D for delivering superior medicines to patients’, *Clinical & Translational Med*, vol. 1, no. 1, p. e7, Dec. 2012, doi: 10.1186/2001-1326-1-7.

[12] M. J. Brown, ‘A RATIONAL BASIS FOR SELECTION AMONG DRUGS OF THE SAME CLASS’, *Heart*, vol. 89, no. 6, pp. 687–694, Jun. 2003, doi: 10.1136/heart.89.6.687.

[13] S. C. Gay, A. G. Roberts, and J. R. Halpert, ‘Structural features of cytochromes P450 and ligands that affect drug metabolism as revealed by X-ray crystallography and NMR’, *Future Med Chem*, vol. 2, no. 9, pp. 1451–1468, Sep. 2010, doi: 10.4155/fmc.10.229.

[14] S. P. Rendic and F. Peter Guengerich, ‘Human cytochrome P450 enzymes 5-51 as targets of drugs and natural and environmental compounds: mechanisms, induction, and inhibition - toxic effects and benefits’, *Drug Metab Rev*, vol. 50, no. 3, pp. 256–342, Aug. 2018, doi: 10.1080/03602532.2018.1483401.

[15] M. J. De Groot, ‘Designing better drugs: predicting cytochrome P450 metabolism’, *Drug Discovery Today*, vol. 11, no. 13–14, pp. 601–606, Jul. 2006, doi: 10.1016/j.drudis.2006.05.001.

[16] G. R. Ford, A. Niehaus, F. Joubert, and M. S. Pepper, ‘Pharmacogenetics of CYP2A6, CYP2B6, and UGT2B7 in the Context of HIV Treatments in African Populations’, *J Pers Med*, vol. 12, no. 12, p. 2013, Dec. 2022, doi: 10.3390/jpm12122013.

[17] H. Raunio, A. Rautio, and O. Pelkonen, ‘The CYP2A subfamily: function, expression and genetic polymorphism’, *IARC Sci Publ*, no. 148, pp. 197–207, 1999.

[18] P. Fernandez-Salguero *et al.*, ‘A genetic polymorphism in coumarin 7-hydroxylation: sequence of the human CYP2A genes and identification of variant CYP2A6 alleles’, *Am J Hum Genet*, vol. 57, no. 3, pp. 651–660, Sep. 1995.

[19] L. Dong, H. Wang, K. Chen, and Y. Li, ‘Roles of hydroxyeicosatetraenoic acids in diabetes (HETEs and diabetes)’, *Biomedicine & Pharmacotherapy*, vol. 156, p. 113981, Dec. 2022, doi: 10.1016/j.biopha.2022.113981.

[20] I. A. Pikuleva and N. Cartier, ‘Cholesterol Hydroxylating Cytochrome P450 46A1: From Mechanisms of Action to Clinical Applications’, *Front. Aging Neurosci.*, vol. 13, p. 696778, Jul. 2021, doi: 10.3389/fnagi.2021.696778.

[21] O. Pelkonen, A. Rautio, H. Raunio, and M. Pasanen, 'CYP2A6: a human coumarin 7-hydroxylase', *Toxicology*, vol. 144, no. 1–3, pp. 139–147, Apr. 2000, doi: 10.1016/S0300-483X(99)00200-0.

[22] *Handbook of Pharmacogenomics and Stratified Medicine*. Elsevier, 2014. doi: 10.1016/C2010-0-67325-1.

[23] B. Luo, D. Yan, H. Yan, and J. Yuan, 'Cytochrome P450: Implications for human breast cancer (Review)', *Oncol Lett*, vol. 22, no. 1, p. 548, May 2021, doi: 10.3892/ol.2021.12809.

[24] F. P. Guengerich, 'Cytochrome P450 and Chemical Toxicology', *Chem. Res. Toxicol.*, vol. 21, no. 1, pp. 70–83, Jan. 2008, doi: 10.1021/tx700079z.

[25] A. Gaedigk *et al.*, 'The Pharmacogene Variation (PharmVar) Consortium: Incorporation of the Human Cytochrome P450 (CYP) Allele Nomenclature Database', *Clin Pharma and Therapeutics*, vol. 103, no. 3, pp. 399–401, Mar. 2018, doi: 10.1002/cpt.910.

[26] A. Judge and M. S. Dodd, 'Metabolism', *Essays in Biochemistry*, vol. 64, no. 4, pp. 607–647, Oct. 2020, doi: 10.1042/EBC20190041.

[27] U. M. Zanger and M. Schwab, 'Cytochrome P450 enzymes in drug metabolism: Regulation of gene expression, enzyme activities, and impact of genetic variation', *Pharmacology & Therapeutics*, vol. 138, no. 1, pp. 103–141, Apr. 2013, doi: 10.1016/j.pharmthera.2012.12.007.

[28] E. Climent, D. Benaiges, and J. Pedro-Botet, 'Hydrophilic or Lipophilic Statins?', *Front. Cardiovasc. Med.*, vol. 8, p. 687585, May 2021, doi: 10.3389/fcvm.2021.687585.

[29] *Pharmacogenomics*. Elsevier, 2013. doi: 10.1016/C2010-0-69430-2.

[30] M. Wilde *et al.*, 'Metabolic Pathways and Potencies of New Fentanyl Analogs', *Front Pharmacol*, vol. 10, p. 238, 2019, doi: 10.3389/fphar.2019.00238.

[31] D. Rizzieri, B. Paul, and Y. Kang, 'Metabolic alterations and the potential for targeting metabolic pathways in the treatment of multiple myeloma', *J Cancer Metastasis Treat*, vol. 5, p. 26, 2019, doi: 10.20517/2394-4722.2019.05.

[32] V. J. Stella, 'Prodrugs as therapeutics', *Expert Opinion on Therapeutic Patents*, vol. 14, no. 3, pp. 277–280, Mar. 2004, doi: 10.1517/13543776.14.3.277.

[33] P. Janov and M. Iller, ‘Phase II Drug Metabolism’, in *Topics on Drug Metabolism*, J. Paxton, Ed., InTech, 2012. doi: 10.5772/29996.

[34] P. Jancova, P. Anzenbacher, and E. Anzenbacherova, ‘PHASE II DRUG METABOLIZING ENZYMES’, *Biomed Pap Med Fac Univ Palacky Olomouc Czech Repub*, vol. 154, no. 2, pp. 103–116, Jun. 2010, doi: 10.5507/bp.2010.017.

[35] N. Hakooz and I. Hamdan, ‘Effects of Dietary Broccoli on Human in Vivo Caffeine Metabolism: A Pilot Study on a Group of Jordanian Volunteers’, *CDM*, vol. 8, no. 1, pp. 9–15, Jan. 2007, doi: 10.2174/138920007779315080.

[36] J. C. Mwenifumbo *et al.*, ‘Novel and established CYP2A6 alleles impair in vivo nicotine metabolism in a population of Black African descent’, *Hum. Mutat.*, vol. 29, no. 5, pp. 679–688, May 2008, doi: 10.1002/humu.20698.

[37] E. E. V. Bezirtzoglou, ‘Intestinal cytochromes P450 regulating the intestinal microbiota and its probiotic profile’, *Microb Ecol Health Dis*, vol. 23, 2012, doi: 10.3402/mehd.v23i0.18370.

[38] M. Sarparast, D. Dattmore, J. Alan, and K. S. S. Lee, ‘Cytochrome P450 Metabolism of Polyunsaturated Fatty Acids and Neurodegeneration’, *Nutrients*, vol. 12, no. 11, p. 3523, Nov. 2020, doi: 10.3390/nu12113523.

[39] C. Barnaba, K. Gentry, N. Sumangala, and A. Ramamoorthy, ‘The catalytic function of cytochrome P450 is entwined with its membrane-bound nature’, *F1000Res*, vol. 6, p. 662, May 2017, doi: 10.12688/f1000research.11015.1.

[40] L. S. Kaminsky, R. S. Obach, and M. J. Fasco, ‘Cytochrome P450: Probes of Active Site Residues’, in *Cytochrome P450*, vol. 105, J. B. Schenkman and H. Greim, Eds., in Handbook of Experimental Pharmacology, vol. 105. , Berlin, Heidelberg: Springer Berlin Heidelberg, 1993, pp. 183–194. doi: 10.1007/978-3-642-77763-9_12.

[41] L. B. Poole, ‘The basics of thiols and cysteines in redox biology and chemistry’, *Free Radic Biol Med*, vol. 80, pp. 148–157, Mar. 2015, doi: 10.1016/j.freeradbiomed.2014.11.013.

[42] J. Zhang *et al.*, ‘The “outsized” role of the I-helix kink in human Cytochrome P450s’, *Clin Transl Med*, vol. 13, no. 9, p. e1378, Sep. 2023, doi: 10.1002/ctm2.1378.

[43] K.-T. Nguyen *et al.*, 'A Novel Thermostable Cytochrome P450 from Sequence-Based Metagenomics of Binh Chau Hot Spring as a Promising Catalyst for Testosterone Conversion', *Catalysts*, vol. 10, no. 9, p. 1083, Sep. 2020, doi: 10.3390/catal10091083.

[44] B. S. P. P.B. Danielson, 'The Cytochrome P450 Superfamily: Biochemistry, Evolution and Drug Metabolism in Humans', *CDM*, vol. 3, no. 6, pp. 561–597, Dec. 2002, doi: 10.2174/1389200023337054.

[45] M. Otyepka, J. Skopalík, E. Anzenbacherová, and P. Anzenbacher, 'What common structural features and variations of mammalian P450s are known to date?', *Biochimica et Biophysica Acta (BBA) - General Subjects*, vol. 1770, no. 3, pp. 376–389, Mar. 2007, doi: 10.1016/j.bbagen.2006.09.013.

[46] D. F. V. Lewis, 'Essential requirements for substrate binding affinity and selectivity toward human CYP2 family enzymes', *Archives of Biochemistry and Biophysics*, vol. 409, no. 1, pp. 32–44, Jan. 2003, doi: 10.1016/S0003-9861(02)00349-1.

[47] O. Gotoh, 'Substrate recognition sites in cytochrome P450 family 2 (CYP2) proteins inferred from comparative analyses of amino acid and coding nucleotide sequences.', *Journal of Biological Chemistry*, vol. 267, no. 1, pp. 83–90, Jan. 1992, doi: 10.1016/S0021-9258(18)48462-1.

[48] K. Kato *et al.*, 'Deciphering Structural Alterations Associated with Activity Reductions of Genetic Polymorphisms in Cytochrome P450 2A6 Using Molecular Dynamics Simulations', *IJMS*, vol. 22, no. 18, p. 10119, Sep. 2021, doi: 10.3390/ijms221810119.

[49] M. A. Schuler and M. R. Berenbaum, 'Structure and Function of Cytochrome P450S in Insect Adaptation to Natural and Synthetic Toxins: Insights Gained from Molecular Modeling', *J Chem Ecol*, vol. 39, no. 9, pp. 1232–1245, Sep. 2013, doi: 10.1007/s10886-013-0335-7.

[50] J. K. Yano, M.-H. Hsu, K. J. Griffin, C. D. Stout, and E. F. Johnson, 'Structures of human microsomal cytochrome P450 2A6 complexed with coumarin and methoxsalen', *Nat Struct Mol Biol*, vol. 12, no. 9, pp. 822–823, Sep. 2005, doi: 10.1038/nsmb971.

[51] R. Davydov *et al.*, 'Role of the Proximal Cysteine Hydrogen Bonding Interaction in Cytochrome P450 2B4 Studied by Cryoreduction, Electron Paramagnetic Resonance, and Electron–Nuclear Double Resonance Spectroscopy', *Biochemistry*, vol. 55, no. 6, pp. 869–883, Feb. 2016, doi: 10.1021/acs.biochem.5b00744.

[52] P. A. Hubbard, A. L. Shen, R. Paschke, C. B. Kasper, and J.-J. P. Kim, 'NADPH-Cytochrome P450 Oxidoreductase', *Journal of Biological Chemistry*, vol. 276, no. 31, pp. 29163–29170, Aug. 2001, doi: 10.1074/jbc.M101731200.

[53] A. S. Ladokhin and S. H. White, 'Protein Chemistry at Membrane Interfaces: Non-additivity of Electrostatic and Hydrophobic Interactions', *Journal of Molecular Biology*, vol. 309, no. 3, pp. 543–552, Jun. 2001, doi: 10.1006/jmbi.2001.4684.

[54] M. A. Lomize, A. L. Lomize, I. D. Pogozheva, and H. I. Mosberg, 'OPM: Orientations of Proteins in Membranes database', *Bioinformatics*, vol. 22, no. 5, pp. 623–625, Mar. 2006, doi: 10.1093/bioinformatics/btk023.

[55] O. Carugo and P. Argos, 'Accessibility to internal cavities and ligand binding sites monitored by protein crystallographic thermal factors', *Proteins*, vol. 31, no. 2, pp. 201–213, May 1998, doi: 10.1002/(SICI)1097-0134(19980501)31:2<201::AID-PROT9>3.0.CO;2-O.

[56] B. C. Monk *et al.*, 'Architecture of a single membrane spanning cytochrome P450 suggests constraints that orient the catalytic domain relative to a bilayer', *Proc. Natl. Acad. Sci. U.S.A.*, vol. 111, no. 10, pp. 3865–3870, Mar. 2014, doi: 10.1073/pnas.1324245111.

[57] A. W. R. Langlois *et al.*, 'Genotyping, characterization, and imputation of known and novel CYP2A6 structural variants using SNP array data', *J Hum Genet*, vol. 68, no. 8, pp. 533–541, Aug. 2023, doi: 10.1038/s10038-023-01148-y.

[58] S. C. Khojasteh-Bakht, L. L. Koenigs, R. M. Peter, W. F. Trager, and S. D. Nelson, '(R)-(+)-Menthofuran is a potent, mechanism-based inactivator of human liver cytochrome P450 2A6', *Drug Metabolism and Disposition*, vol. 26, no. 7, pp. 701–704, 1998.

[59] C. Dandara, M. Swart, B. Mpeta, A. Wonkam, and C. Masimirembwa, 'Cytochrome P450 pharmacogenetics in African populations: implications for public health', *Expert Opinion on Drug Metabolism & Toxicology*, vol. 10, no. 6, pp. 769–785, Jun. 2014, doi: 10.1517/17425255.2014.894020.

[60] J.-A. Tanner *et al.*, 'Predictors of Variation in CYP2A6 mRNA, Protein, and Enzyme Activity in a Human Liver Bank: Influence of Genetic and Nongenetic Factors', *J Pharmacol Exp Ther*, vol. 360, no. 1, pp. 129–139, Jan. 2017, doi: 10.1124/jpet.116.237594.

[61] W. Yusof and G. S. Hua, ‘Gene, ethnic and gender influences predisposition of adverse drug reactions to artesunate among Malaysians’, *Toxicology Mechanisms and Methods*, vol. 22, no. 3, pp. 184–192, Apr. 2012, doi: 10.3109/15376516.2011.623331.

[62] M. K. Ho, J. C. Mwenifumbo, B. Zhao, E. M. J. Gillam, and R. F. Tyndale, ‘A novel CYP2A6 allele, CYP2A6*23, impairs enzyme function in vitro and in vivo and decreases smoking in a population of Black-African descent’, *Pharmacogenet Genomics*, vol. 18, no. 1, pp. 67–75, Jan. 2008, doi: 10.1097/FPC.0b013e3282f3606e.

[63] I. Rajman, L. Knapp, T. Morgan, and C. Masimirembwa, ‘African Genetic Diversity: Implications for Cytochrome P450-mediated Drug Metabolism and Drug Development’, *EBioMedicine*, vol. 17, pp. 67–74, Mar. 2017, doi: 10.1016/j.ebiom.2017.02.017.

[64] S. Ramachandran, O. Deshpande, C. C. Roseman, N. A. Rosenberg, M. W. Feldman, and L. L. Cavalli-Sforza, ‘Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa’, *Proc. Natl. Acad. Sci. U.S.A.*, vol. 102, no. 44, pp. 15942–15947, Nov. 2005, doi: 10.1073/pnas.0507611102.

[65] R. K. Bains, ‘African variation at Cytochrome P450 genes’, *Evolution, Medicine, and Public Health*, vol. 2013, no. 1, pp. 118–134, 2013, doi: 10.1093/emph/eot010.

[66] M. Eichelbaum, M. Ingelman-Sundberg, and W. E. Evans, ‘Pharmacogenomics and Individualized Drug Therapy’, *Annu. Rev. Med.*, vol. 57, no. 1, pp. 119–137, Feb. 2006, doi: 10.1146/annurev.med.56.082103.104724.

[67] D. Stefanicka-Wojtas and D. Kurpas, ‘Personalised Medicine—Implementation to the Healthcare System in Europe (Focus Group Discussions)’, *JPM*, vol. 13, no. 3, p. 380, Feb. 2023, doi: 10.3390/jpm13030380.

[68] A. Gaedigk *et al.*, ‘The Pharmacogene Variation (PharmVar) Consortium: Incorporation of the Human Cytochrome P450 (*CYP*) Allele Nomenclature Database’, *Clin Pharma and Therapeutics*, vol. 103, no. 3, pp. 399–401, Mar. 2018, doi: 10.1002/cpt.910.

[69] ‘Crystallography: Protein Data Bank’, *Nature New Biology*, vol. 233, no. 42, pp. 223–223, Oct. 1971, doi: 10.1038/newbio233223b0.

[70] K. Djinovic-Carugo and O. Carugo, ‘Missing strings of residues in protein crystal structures’, *Intrinsically Disord Proteins*, vol. 3, no. 1, p. e1095697, 2015, doi: 10.1080/21690707.2015.1095697.

[71] D. Zheng, Z. Zhang, P. M. Harrison, J. Karro, N. Carriero, and M. Gerstein, 'Integrated Pseudogene Annotation for Human Chromosome 22: Evidence for Transcription', *Journal of Molecular Biology*, vol. 349, no. 1, pp. 27–45, May 2005, doi: 10.1016/j.jmb.2005.02.072.

[72] C. Mayr, 'What Are 3′ UTRs Doing?', *Cold Spring Harb Perspect Biol*, vol. 11, no. 10, p. a034728, Oct. 2019, doi: 10.1101/cshperspect.a034728.

[73] S. Cholerton, M. E. Idle, A. Vas, F. J. Gonzalez, and J. R. Idle, 'Comparison of a novel thin-layer chromatographic—fluorescence detection method with a spectrofluorometric method for the determination of 7-hydroxycoumarin in human urine', *Journal of Chromatography B: Biomedical Sciences and Applications*, vol. 575, no. 2, pp. 325–330, Mar. 1992, doi: 10.1016/0378-4347(92)80166-N.

[74] S. K. Jones, B. J. Wolf, B. Froeliger, K. Wallace, M. J. Carpenter, and A. J. Alberg, 'Nicotine Metabolism Predicted by *CYP2A6* Genotypes in Relation to Smoking Cessation: A Systematic Review', *Nicotine & Tobacco Research*, vol. 24, no. 5, pp. 633–642, Mar. 2022, doi: 10.1093/ntr/ntab175.

[75] C. Xu, S. Goodz, E. M. Sellers, and R. F. Tyndale, 'CYP2A6 genetic variation and potential consequences', *Advanced Drug Delivery Reviews*, vol. 54, no. 10, pp. 1245–1256, Nov. 2002, doi: 10.1016/S0169-409X(02)00065-0.

[76] K. Kitagawa, N. Kunugita, M. Kitagawa, and T. Kawamoto, 'CYP2A6*6, a Novel Polymorphism in Cytochrome P450 2A6, Has a Single Amino Acid Substitution (R128Q) That Inactivates Enzymatic Activity', *Journal of Biological Chemistry*, vol. 276, no. 21, pp. 17830–17835, May 2001, doi: 10.1074/jbc.M009432200.

[77] S. Daigo *et al.*, 'A novel mutant allele of the CYP2A6 gene (CYP2A6*11) found in a cancer patient who showed poor metabolic phenotype towards tegafur':, *Pharmacogenetics*, vol. 12, no. 4, pp. 299–306, Jun. 2002, doi: 10.1097/00008571-200206000-00005.

[78] K. Kiyotani *et al.*, 'Twenty One Novel Single Nucleotide Polymorphisms (SNPs) of the CYP2A6 Gene in Japanese and Caucasians', *Drug Metabolism and Pharmacokinetics*, vol. 17, no. 5, pp. 482–487, 2002, doi: 10.2133/dmpk.17.482.

[79] T. Fukami *et al.*, 'A novel polymorphism of human gene has an amino acid substitution (V365M) that decreases enzymatic activity in vitro and in vivo', *Clinical Pharmacology & Therapeutics*, vol. 76, no. 6, pp. 519–527, Dec. 2004, doi: 10.1016/j.clpt.2004.08.014.

[80] T. Fukami *et al.*, 'CHARACTERIZATION OF NOVEL *CYP2A6* POLYMORPHIC ALLELES (*CYP2A6*18* AND *CYP2A6*19*) THAT AFFECT ENZYMATIC ACTIVITY', *Drug Metab Dispos*, vol. 33, no. 8, pp. 1202–1210, Aug. 2005, doi: 10.1124/dmd.105.004994.

[81] M. Haberl *et al.*, 'Three haplotypes associated with CYP2A6 phenotypes in Caucasians', *Pharmacogenetics and Genomics*, vol. 15, no. 9, pp. 609–624, Sep. 2005, doi: 10.1097/01.fpc.0000171517.22258.f1.

[82] J. Bloom *et al.*, 'The contribution of common CYP2A6 alleles to variation in nicotine metabolism among European-Americans', *Pharmacogenet Genomics*, vol. 21, no. 7, pp. 403–416, Jul. 2011, doi: 10.1097/FPC.0b013e328346e8c0.

[83] M. Piliguian *et al.*, 'Novel CYP2A6 variants identified in African Americans are associated with slow nicotine metabolism in vitro and in vivo', *Pharmacogenet Genomics*, vol. 24, no. 2, pp. 118–128, Feb. 2014, doi: 10.1097/FPC.0000000000000026.

[84] A. H. Mao, S. L. Crick, A. Vitalis, C. L. Chicoine, and R. V. Pappu, 'Net charge per residue modulates conformational ensembles of intrinsically disordered proteins', *Proc. Natl. Acad. Sci. U.S.A.*, vol. 107, no. 18, pp. 8183–8188, May 2010, doi: 10.1073/pnas.0911107107.

[85] T. Ubuka, 'Amino acids', in *Handbook of Hormones*, Elsevier, 2021, pp. 1061–1062. doi: 10.1016/B978-0-12-820649-2.00295-3.

[86] M. J. Betts and R. B. Russell, 'Amino Acid Properties and Consequences of Substitutions', in *Bioinformatics for Geneticists*, 1st ed., M. R. Barnes and I. C. Gray, Eds., Wiley, 2003, pp. 289–316. doi: 10.1002/0470867302.ch14.

[87] D. Williams, A. Kenyon, and D. Adamson, 'Physiology', in *Basic Science in Obstetrics and Gynaecology*, Elsevier, 2010, pp. 173–230. doi: 10.1016/B978-0-443-10281-3.00014-2.

[88] M Blétry, 'Henderson Hasselbalch relationship and weak acid titration', 2018, doi: 10.13140/RG.2.2.11836.13445.

[89] H. N. Po and N. M. Senozan, 'The Henderson-Hasselbalch Equation: Its History and Limitations', *J. Chem. Educ.*, vol. 78, no. 11, p. 1499, Nov. 2001, doi: 10.1021/ed078p1499.

[90] Y. Cao, M. Wang, Y. Yuan, C. Li, Q. Bai, and M. Li, 'Arterial blood gas and acid-base balance in patients with pregnancy-induced hypertension syndrome', *Exp Ther Med*, vol. 17, no. 1, pp. 349–353, Jan. 2019, doi: 10.3892/etm.2018.6893.

[91] D. Castro, S. M. Patil, M. Zubair, and M. Keenaghan, 'Arterial Blood Gas', in *StatPearls*, Treasure Island (FL): StatPearls Publishing, 2024. Accessed: Mar. 09, 2024. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK536919/

[92] H. Li and T. Poulos, 'Crystallization of Cytochromes P450 and Substrate-Enzyme Interactions', *CTMC*, vol. 4, no. 16, pp. 1789–1802, Dec. 2004, doi: 10.2174/1568026043387205.

[93] P. Urban, T. Lautier, D. Pompon, and G. Truan, 'Ligand Access Channels in Cytochrome P450 Enzymes: A Review', *IJMS*, vol. 19, no. 6, p. 1617, May 2018, doi: 10.3390/ijms19061617.

[94] P. Chakrabarti and J. Janin, 'Dissecting protein–protein recognition sites', *Proteins*, vol. 47, no. 3, pp. 334–343, May 2002, doi: 10.1002/prot.10085.

[95] Y. Wang, L. Y. Geer, C. Chappey, J. A. Kans, and S. H. Bryant, 'Cn3D: sequence and structure views for Entrez', *Trends in Biochemical Sciences*, vol. 25, no. 6, pp. 300–302, Jun. 2000, doi: 10.1016/S0968-0004(00)01561-9.

[96] F. Spyrakis, M. H. Ahmed, A. S. Bayden, P. Cozzini, A. Mozzarelli, and G. E. Kellogg, 'The Roles of Water in the Protein Matrix: A Largely Untapped Resource for Drug Discovery', *J. Med. Chem.*, vol. 60, no. 16, pp. 6781–6827, Aug. 2017, doi: 10.1021/acs.jmedchem.7b00057.

[97] Q. Sun, 'The Hydrophobic Effects: Our Current Understanding', *Molecules*, vol. 27, no. 20, p. 7009, Oct. 2022, doi: 10.3390/molecules27207009.

[98] M. Eilers, S. C. Shekar, T. Shieh, S. O. Smith, and P. J. Fleming, 'Internal packing of helical membrane proteins', *Proc. Natl. Acad. Sci. U.S.A.*, vol. 97, no. 11, pp. 5796–5801, May 2000, doi: 10.1073/pnas.97.11.5796.

[99] V. Cojocaru, P. J. Winn, and R. C. Wade, 'The ins and outs of cytochrome P450s', *Biochimica et Biophysica Acta (BBA) - General Subjects*, vol. 1770, no. 3, pp. 390–401, Mar. 2007, doi: 10.1016/j.bbagen.2006.07.005.

[100] J. F. Solus *et al.*, 'Genetic variation in eleven phase I drug metabolism genes in an ethnically diverse population', *pgs*, vol. 5, no. 7, pp. 895–931, Oct. 2004, doi: 10.1517/14622416.5.7.895.

[101] X.-X. Zhou, Y.-B. Wang, Y.-J. Pan, and W.-F. Li, 'Differences in amino acids composition and coupling patterns between mesophilic and thermophilic proteins', *Amino Acids*, vol. 34, no. 1, pp. 25–33, Jan. 2008, doi: 10.1007/s00726-007-0589-x.

[102] J.-M. Yang *et al.*, 'A Novel Glutamic Acid to Aspartic Acid Mutation Near the End of the 2B Rod Domain in the Keratin 1 Chain in Epidermolytic Hyperkeratosis', *Journal of Investigative Dermatology*, vol. 112, no. 3, pp. 376–379, Mar. 1999, doi: 10.1038/sj.jid.5600439.

[103] D. P. Dulebohn, H. J. Cho, and A. W. Karzai, 'Role of Conserved Surface Amino Acids in Binding of SmpB Protein to SsrA RNA', *Journal of Biological Chemistry*, vol. 281, no. 39, pp. 28536–28545, Sep. 2006, doi: 10.1074/jbc.M605137200.

[104] D. S. Dwyer, 'Amino Acids: Chemical Properties', in *Wiley Encyclopedia of Chemical Biology*, 1st ed., Wiley, 2008, pp. 1–11. doi: 10.1002/9780470048672.wecb007.

[105] O. Guvench and A. D. MacKerell, 'Comparison of Protein Force Fields for Molecular Dynamics Simulations', in *Molecular Modeling of Proteins*, vol. 443, A. Kukol, Ed., in Methods in Molecular Biology, vol. 443. , Totowa, NJ: Humana Press, 2008, pp. 63–88. doi: 10.1007/978-1-59745-177-2_4.

[106] M. A. González, 'Force fields and molecular dynamics simulations', *JDN*, vol. 12, pp. 169–200, 2011, doi: 10.1051/sfn/201112009.

[107] J. Huang *et al.*, 'CHARMM36m: an improved force field for folded and intrinsically disordered proteins', *Nat Methods*, vol. 14, no. 1, pp. 71–73, Jan. 2017, doi: 10.1038/nmeth.4067.

[108] D. A. C. P.A. Kollman, 'Amber 2023', in *The Amber Project*.

[109] T. E. Cheatham, 'Simulation and modeling of nucleic acid structure, dynamics and interactions', *Current Opinion in Structural Biology*, vol. 14, no. 3, pp. 360–367, Jun. 2004, doi: 10.1016/j.sbi.2004.05.001.

[110] Giudice Emmanuel and R. Lavery, 'Simulations of Nucleic Acids and Their Complexes', *Acc. Chem. Res.*, vol. 35, no. 6, pp. 350–357, Jun. 2002, doi: 10.1021/ar010023y.

[111] M. Orozco, A. Pérez, A. Noy, and F. J. Luque, 'Theoretical methods for the simulation of nucleic acids', *Chem. Soc. Rev.*, vol. 32, no. 6, pp. 350–364, 2003, doi: 10.1039/B207226M.

[112] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, 'Development and testing of a general amber force field', *J. Comput. Chem.*, vol. 25, no. 9, pp. 1157–1174, Jul. 2004, doi: 10.1002/jcc.20035.

[113] P. Li and K. M. Merz, 'MCPB.py: A Python Based Metal Center Parameter Builder', *J. Chem. Inf. Model.*, vol. 56, no. 4, pp. 599–604, Apr. 2016, doi: 10.1021/acs.jcim.5b00674.

[114] J. J. Novoa, P. Lafuente, R. E. Del Sesto, and J. S. Miller, 'Exceptionally Long (≥2.9 Å) C−C Bonds between [TCNE]− Ions: Two-Electron, Four-Center π*-π* C−C Bonding in π-[TCNE]22−', *Angew. Chem. Int. Ed.*, vol. 40, no. 13, pp. 2540–2545, Jul. 2001, doi: 10.1002/1521-3773(20010702)40:13<2540::AID-ANIE2540>3.0.CO;2-O.

[115] J. Parsons, J. B. Holmes, J. M. Rojas, J. Tsai, and C. E. M. Strauss, 'Practical conversion from torsion space to Cartesian space for *in silico* protein synthesis', *J Comput Chem*, vol. 26, no. 10, pp. 1063–1068, Jul. 2005, doi: 10.1002/jcc.20237.

[116] V. Gold, Ed., *The IUPAC Compendium of Chemical Terminology: The Gold Book*, 4th ed. Research Triangle Park, NC: International Union of Pure and Applied Chemistry (IUPAC), 2019. doi: 10.1351/goldbook.

[117] R. J. MacDonell, S. Patchkovskii, and M. S. Schuurman, 'A Comparison of Partial Atomic Charges for Electronically Excited States', *J. Chem. Theory Comput.*, vol. 18, no. 2, pp. 1061–1071, Feb. 2022, doi: 10.1021/acs.jctc.1c01101.

[118] S. S. Batsanov, '[No title found]', *Inorganic Materials*, vol. 37, no. 9, pp. 871–885, 2001, doi: 10.1023/A:1011625728803.

[119] K. Shahrokh, A. Orendt, G. S. Yost, and T. E. Cheatham, 'Quantum mechanically derived AMBER-compatible heme parameters for various states of the cytochrome P450 catalytic cycle', *J. Comput. Chem.*, vol. 33, no. 2, pp. 119–133, Jan. 2012, doi: 10.1002/jcc.21922.

[120] D. E. Bikiel *et al.*, 'Modeling heme proteins using atomistic simulations', *Phys. Chem. Chem. Phys.*, vol. 8, no. 48, pp. 5611–5628, 2006, doi: 10.1039/B611741B.

[121] M. D. Maines and A. Kappas, ‘Metals as Regulators of Heme Metabolism’, *Science*, vol. 198, no. 4323, pp. 1215–1221, Dec. 1977, doi: 10.1126/science.337492.

[122] C. Tian *et al.*, ‘ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution’, *J. Chem. Theory Comput.*, vol. 16, no. 1, pp. 528–552, Jan. 2020, doi: 10.1021/acs.jctc.9b00591.

[123] Y. Xing, S. Gao, X. Zhang, and J. Zang, ‘Dietary Heme-Containing Proteins: Structures, Applications, and Challenges’, *Foods*, vol. 11, no. 22, p. 3594, Nov. 2022, doi: 10.3390/foods11223594.

[124] G. Gilardi and G. Di Nardo, ‘Heme iron centers in cytochrome P450: structure and catalytic activity’, *Rend. Fis. Acc. Lincei*, vol. 28, no. S1, pp. 159–167, Jul. 2017, doi: 10.1007/s12210-016-0565-z.

[125] P. Ponka, ‘Cell Biology of Heme’, *The American Journal of the Medical Sciences*, vol. 318, no. 4, pp. 241–256, Oct. 1999, doi: 10.1016/S0002-9629(15)40628-7.

[126] M. Paoli, J. Marles-Wright, and A. Smith, ‘Structure–Function Relationships in Heme-Proteins’, *DNA and Cell Biology*, vol. 21, no. 4, pp. 271–280, Apr. 2002, doi: 10.1089/104454902753759690.

[127] F. Varfaj, J. N. Lampe, and P. R. Ortiz de Montellano, ‘Role of cysteine residues in heme binding to human heme oxygenase-2 elucidated by two-dimensional NMR spectroscopy’, *J Biol Chem*, vol. 287, no. 42, pp. 35181–35191, Oct. 2012, doi: 10.1074/jbc.M112.378042.

[128] S. Dutt, I. Hamza, and T. B. Bartnikas, ‘Molecular Mechanisms of Iron and Heme Metabolism’, *Annu Rev Nutr*, vol. 42, pp. 311–335, Aug. 2022, doi: 10.1146/annurev-nutr-062320-112625.

[129] L. Chebon-Bore, T. A. Sanyanga, C. V. Manyumwa, A. Khairallah, and Ö. Tastan Bishop, ‘Decoding the Molecular Effects of Atovaquone Linked Resistant Mutations on Plasmodium falciparum Cytb-ISP Complex in the Phospholipid Bilayer Membrane’, *IJMS*, vol. 22, no. 4, p. 2138, Feb. 2021, doi: 10.3390/ijms22042138.

[130] T. A. Wassenaar and A. E. Mark, ‘The effect of box shape on the dynamic properties of proteins simulated under periodic boundary conditions’, *J Comput Chem*, vol. 27, no. 3, pp. 316–325, Feb. 2006, doi: 10.1002/jcc.20341.

[131] A. W. Sousa Da Silva and W. F. Vranken, ‘ACPYPE - AnteChamber PYthon Parser interfacE’, *BMC Res Notes*, vol. 5, no. 1, p. 367, Dec. 2012, doi: 10.1186/1756-0500-5-367.

[132] E. Lindahl, B. Hess, and D. Van Der Spoel, ‘GROMACS 3.0: a package for molecular simulation and trajectory analysis’, *J Mol Model*, vol. 7, no. 8, pp. 306–317, Aug. 2001, doi: 10.1007/s008940100045.

[133] M. Hernández-Rodríguez, M. C. Rosales-Hernández, J. E. Mendieta-Wejebe, M. Martínez-Archundia, and J. Correa Basurto, ‘Current Tools and Methods in Molecular Dynamics (MD) Simulations for Drug Design’, *CMC*, vol. 23, no. 34, pp. 3909–3924, Nov. 2016, doi: 10.2174/0929867323666160530144742.

[134] D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen, ‘GROMACS: Fast, flexible, and free’, *J. Comput. Chem.*, vol. 26, no. 16, pp. 1701–1718, Dec. 2005, doi: 10.1002/jcc.20291.

[135] M. Abraham *et al.*, ‘GROMACS 2023.2 Source code’. Zenodo, Jul. 12, 2023. doi: 10.5281/ZENODO.8134397.

[136] M. J. Abraham *et al.*, ‘GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers’, *SoftwareX*, vol. 1–2, pp. 19–25, Sep. 2015, doi: 10.1016/j.softx.2015.06.001.

[137] P. C. Nair, R. A. McKinnon, and J. O. Miners, ‘Cytochrome P450 structure–function: insights from molecular dynamics simulations’, *Drug Metabolism Reviews*, vol. 48, no. 3, pp. 434–452, Jul. 2016, doi: 10.1080/03602532.2016.1178771.

[138] N. Magalhães *et al.*, ‘Interactions between Rhodamine Dyes and Model Membrane Systems—Insights from Molecular Dynamics Simulations’, *Molecules*, vol. 27, no. 4, p. 1420, Feb. 2022, doi: 10.3390/molecules27041420.

[139] S. A. Hollingsworth and R. O. Dror, ‘Molecular Dynamics Simulation for All’, *Neuron*, vol. 99, no. 6, pp. 1129–1143, Sep. 2018, doi: 10.1016/j.neuron.2018.08.011.

[140] M. Karplus and J. A. McCammon, ‘Molecular dynamics simulations of biomolecules’, *Nat. Struct Biol.*, vol. 9, no. 9, pp. 646–652, Sep. 2002, doi: 10.1038/nsb0902-646.

[141] H. Liu, S. Xiang, H. Zhu, and L. Li, 'The Structural and Dynamical Properties of the Hydration of SNase Based on a Molecular Dynamics Simulation', *Molecules*, vol. 26, no. 17, p. 5403, Sep. 2021, doi: 10.3390/molecules26175403.

[142] S. Yao, B. Ma, Q. Yi, M.-X. Guan, and X. Cang, 'Investigating the Broad Matrix-Gate Network in the Mitochondrial ADP/ATP Carrier through Molecular Dynamics Simulations', *Molecules*, vol. 27, no. 3, p. 1071, Feb. 2022, doi: 10.3390/molecules27031071.

[143] H. Liu, Y. Jin, and H. Ding, 'MDBuilder: a PyMOL plugin for the preparation of molecular dynamics simulations', *Briefings in Bioinformatics*, vol. 24, no. 2, p. bbad057, Mar. 2023, doi: 10.1093/bib/bbad057.

[144] B. Gautam, 'Energy Minimization', in *Homology Molecular Modeling - Perspectives and Applications*, R. Trindade Maia, R. Maciel De Moraes Filho, and M. Campos, Eds., IntechOpen, 2021. doi: 10.5772/intechopen.94809.

[145] G. Romeo, 'Classical static nonlinear optimization theory', in *Elements of Numerical Mathematical Economics with Excel*, Elsevier, 2020, pp. 219–293. doi: 10.1016/B978-0-12-817648-1.00005-0.

[146] L. Fiedler *et al.*, 'Accelerating equilibration in first-principles molecular dynamics with orbital-free density functional theory', *Phys. Rev. Research*, vol. 4, no. 4, p. 043033, Oct. 2022, doi: 10.1103/PhysRevResearch.4.043033.

[147] L. Piao *et al.*, 'Molecular Dynamics Simulations of Wild Type and Mutants of SAPAP in Complexed with Shank3', *Int J Mol Sci*, vol. 20, no. 1, p. 224, Jan. 2019, doi: 10.3390/ijms20010224.

[148] G. Clavier, N. Desbiens, E. Bourasseau, V. Lachet, N. Brusselle-Dupend, and B. Rousseau, 'Computation of elastic constants of solids using molecular simulation: comparison of constant volume and constant pressure ensemble methods', *Molecular Simulation*, vol. 43, no. 17, pp. 1413–1422, Nov. 2017, doi: 10.1080/08927022.2017.1313418.

[149] S. Nosé and M. L. Klein, 'Constant pressure molecular dynamics for molecular systems', *Molecular Physics*, vol. 50, no. 5, pp. 1055–1076, Dec. 1983, doi: 10.1080/00268978300102851.

[150] S. Ishak *et al.*, 'Molecular Dynamic Simulation of Space and Earth-Grown Crystal Structures of Thermostable T1 Lipase Geobacillus zalihae Revealed a Better Structure', *Molecules*, vol. 22, no. 10, p. 1574, Sep. 2017, doi: 10.3390/molecules22101574.

[151] W. Humphrey, A. Dalke, and K. Schulten, 'VMD: Visual molecular dynamics', *Journal of Molecular Graphics*, vol. 14, no. 1, pp. 33–38, Feb. 1996, doi: 10.1016/0263-7855(96)00018-5.

[152] J. Farmer, F. Kanwal, N. Nikulsin, M. Tsilimigras, and D. Jacobs, 'Statistical Measures to Quantify Similarity between Molecular Dynamics Simulation Trajectories', *Entropy*, vol. 19, no. 12, p. 646, Nov. 2017, doi: 10.3390/e19120646.

[153] K. Sargsyan, C. Grauffel, and C. Lim, 'How Molecular Size Impacts RMSD Applications in Molecular Dynamics Simulations', *J. Chem. Theory Comput.*, vol. 13, no. 4, pp. 1518–1524, Apr. 2017, doi: 10.1021/acs.jctc.7b00028.

[154] R. Tanious and R. Manolov, 'Violin plots as visual tools in the meta-analysis of Single-Case Experimental Designs', *Methodology*, vol. 18, no. 3, pp. 221–238, Sep. 2022, doi: 10.5964/meth.9209.

[155] M. Kenny and I. Schoen, 'Violin SuperPlots: visualizing replicate heterogeneity in large data sets', *MBoC*, vol. 32, no. 15, pp. 1333–1334, Jul. 2021, doi: 10.1091/mbc.E21-03-0130.

[156] A. K. Jain, M. N. Murty, and P. J. Flynn, 'Data clustering: a review', *ACM Comput. Surv.*, vol. 31, no. 3, pp. 264–323, Sep. 1999, doi: 10.1145/331499.331504.

[157] C. Ordonez and E. Omiecinski, 'Efficient disk-based K-means clustering for relational databases', *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 8, pp. 909–921, Aug. 2004, doi: 10.1109/TKDE.2004.25.

[158] B. Ma, C. Yang, A. Li, Y. Chi, and L. Chen, 'A Faster DBSCAN Algorithm Based on Self-Adaptive Determination of Parameters', *Procedia Computer Science*, vol. 221, pp. 113–120, 2023, doi: 10.1016/j.procs.2023.07.017.

[159] A. Bouguettaya, Q. Yu, X. Liu, X. Zhou, and A. Song, 'Efficient agglomerative hierarchical clustering', *Expert Systems with Applications*, vol. 42, no. 5, pp. 2785–2797, Apr. 2015, doi: 10.1016/j.eswa.2014.09.054.

[160] J. Shao, S. W. Tanner, N. Thompson, and T. E. Cheatham, 'Clustering Molecular Dynamics Trajectories: 1. Characterizing the Performance of Different Clustering

Algorithms', *J. Chem. Theory Comput.*, vol. 3, no. 6, pp. 2312–2334, Nov. 2007, doi: 10.1021/ct700119m.

[161] P. Carter, C. A. F. Andersen, and B. Rost, 'DSSPcont: Continuous secondary structure assignments for proteins', *Nucleic Acids Res*, vol. 31, no. 13, pp. 3293–3295, Jul. 2003, doi: 10.1093/nar/gkg626.

[162] J. Reeb and B. Rost, 'Secondary Structure Prediction', in *Encyclopedia of Bioinformatics and Computational Biology*, Elsevier, 2019, pp. 488–496. doi: 10.1016/B978-0-12-809633-8.20267-7.

[163] S. Abeln, K. A. Feenstra, and J. Heringa, 'Protein Three-Dimensional Structure Prediction', in *Encyclopedia of Bioinformatics and Computational Biology*, Elsevier, 2019, pp. 497–511. doi: 10.1016/B978-0-12-809633-8.20505-0.

[164] Q. Ke, X. Gong, S. Liao, C. Duan, and L. Li, 'Effects of thermostats/barostats on physical properties of liquids by molecular dynamics simulations', *Journal of Molecular Liquids*, vol. 365, p. 120116, Nov. 2022, doi: 10.1016/j.molliq.2022.120116.

[165] K. H. Tiong, N. A. Mohammed Yunus, B. C. Yiap, E. L. Tan, R. Ismail, and C. E. Ong, 'Inhibitory Potency of 8-Methoxypsoralen on Cytochrome P450 2A6 (CYP2A6) Allelic Variants CYP2A6*15, CYP2A6*16, CYP2A6*21 and CYP2A6*22: Differential Susceptibility Due to Different Sequence Locations of the Mutations', *PLoS ONE*, vol. 9, no. 1, p. e86230, Jan. 2014, doi: 10.1371/journal.pone.0086230.

[166] K. H. Tiong, B. C. Yiap, E. L. Tan, R. Ismail, and C. E. Ong, 'Functional Characterization of Cytochrome P450 2A6 Allelic Variants *CYP2A6*15* , *CYP2A6*16* , *CYP2A6*21* , and *CYP2A6*22*', *Drug Metab Dispos*, vol. 38, no. 5, pp. 745–751, May 2010, doi: 10.1124/dmd.109.031054.

[167] M. Yu. Lobanov, N. S. Bogatyreva, and O. V. Galzitskaya, 'Radius of gyration as an indicator of protein structure compactness', *Mol Biol*, vol. 42, no. 4, pp. 623–628, Aug. 2008, doi: 10.1134/S0026893308040195.

[168] O. V. Galzitskaya and S. O. Garbuzynskiy, 'Entropy capacity determines protein folding', *Proteins*, vol. 63, no. 1, pp. 144–154, Apr. 2006, doi: 10.1002/prot.20851.

[169] D. Seeliger and B. L. De Groot, 'Conformational Transitions upon Ligand Binding: Holo-Structure Prediction from Apo Conformations', *PLoS Comput Biol*, vol. 6, no. 1, p. e1000634, Jan. 2010, doi: 10.1371/journal.pcbi.1000634.

[170] R. Yoshida, ‘Effects of polymorphism in promoter region of human CYP2A6 gene (CYP2A6*9) on expression level of messenger ribonucleic acid and enzymatic activity in vivo and in vitro’, *Clinical Pharmacology & Therapeutics*, vol. 74, no. 1, pp. 69–76, Jul. 2003, doi: 10.1016/S0009-9236(03)00090-0.

[171] Y. Di, V. Chow, L.-P. Yang, and S.-F. Zhou, ‘Structure, Function, Regulation and Polymorphism of Human Cytochrome P450 2A6’, *CDM*, vol. 10, no. 7, pp. 754–780, Sep. 2009, doi: 10.2174/138920009789895507.

[172] S. Satarug, W. Tassaneeyakul, K. Na-Bangchang, J. Cashman, and M. Moore, ‘Genetic and Environmental Influences on Therapeutic and Toxicity Outcomes: Studies with CYP2A6’, *CCP*, vol. 1, no. 3, pp. 291–309, Sep. 2006, doi: 10.2174/157488406778249343.

[173] D. K. Brown *et al.*, ‘MD-TASK: a software suite for analyzing molecular dynamics trajectories’, *Bioinformatics*, vol. 33, no. 17, pp. 2768–2771, Sep. 2017, doi: 10.1093/bioinformatics/btx349.

SUPPLEMENTARY MATERIAL

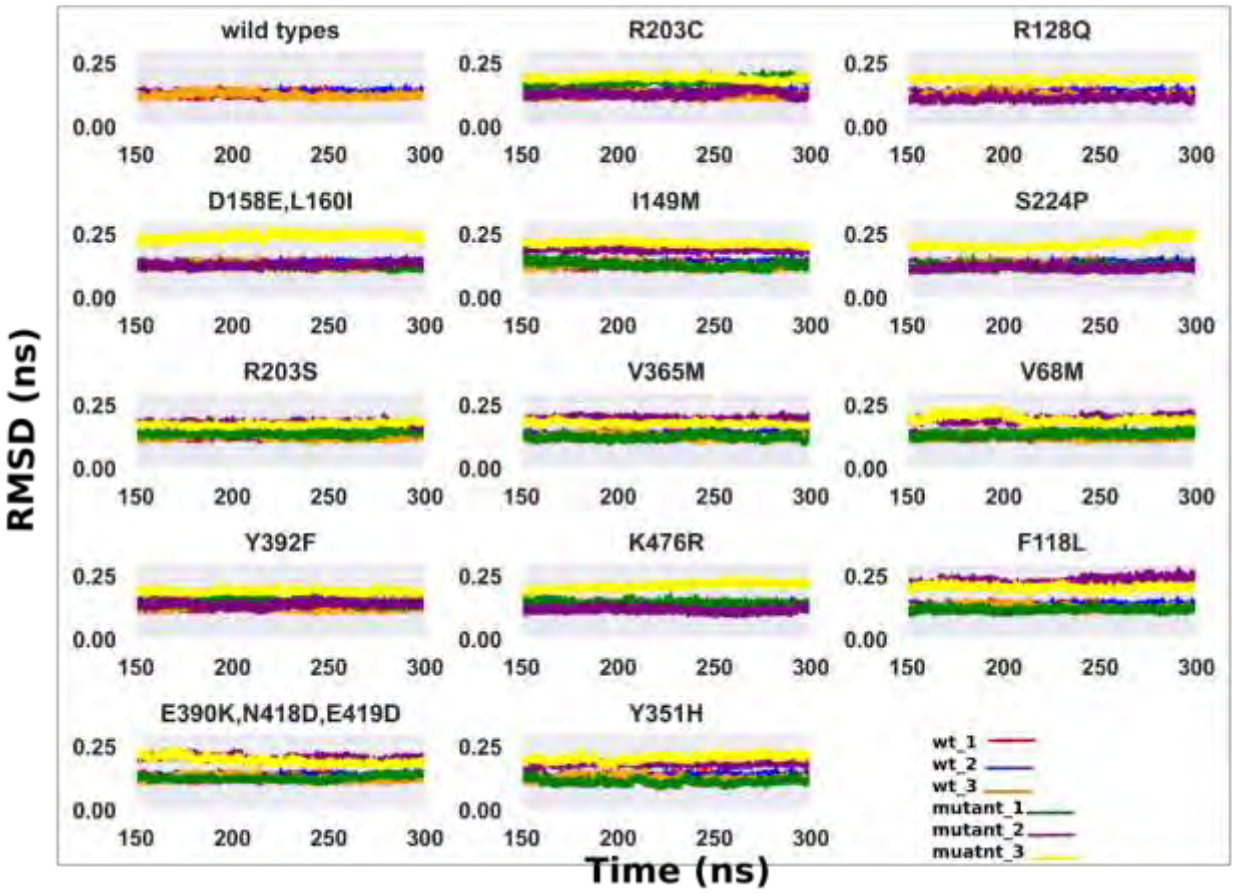


Figure S1. The RMSD line plots from 150ns to 300ns (Berendsen at production run results)

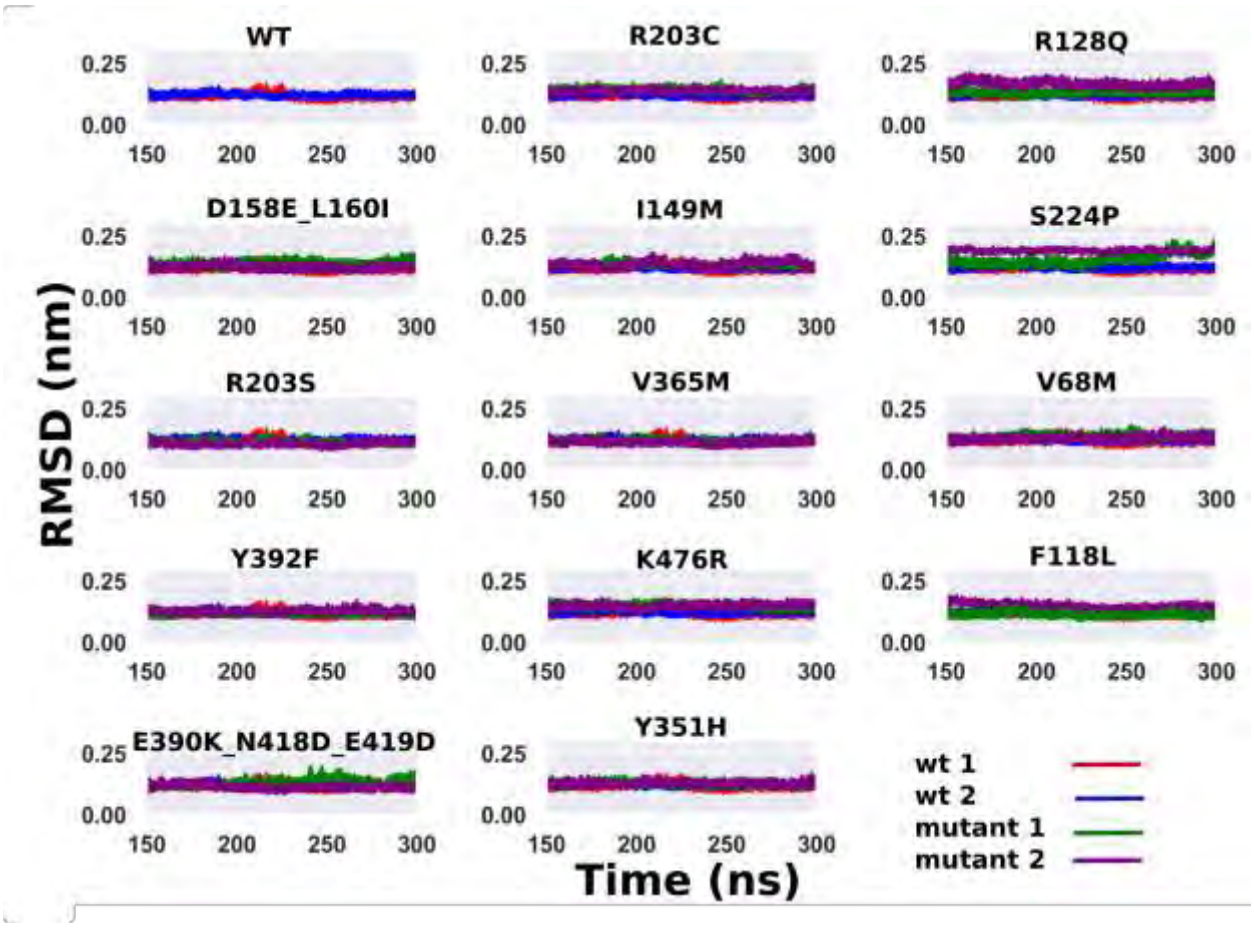


Figure S2. The RMSD line plots from 150ns to 300ns (Parinello-Rahman at production run results)

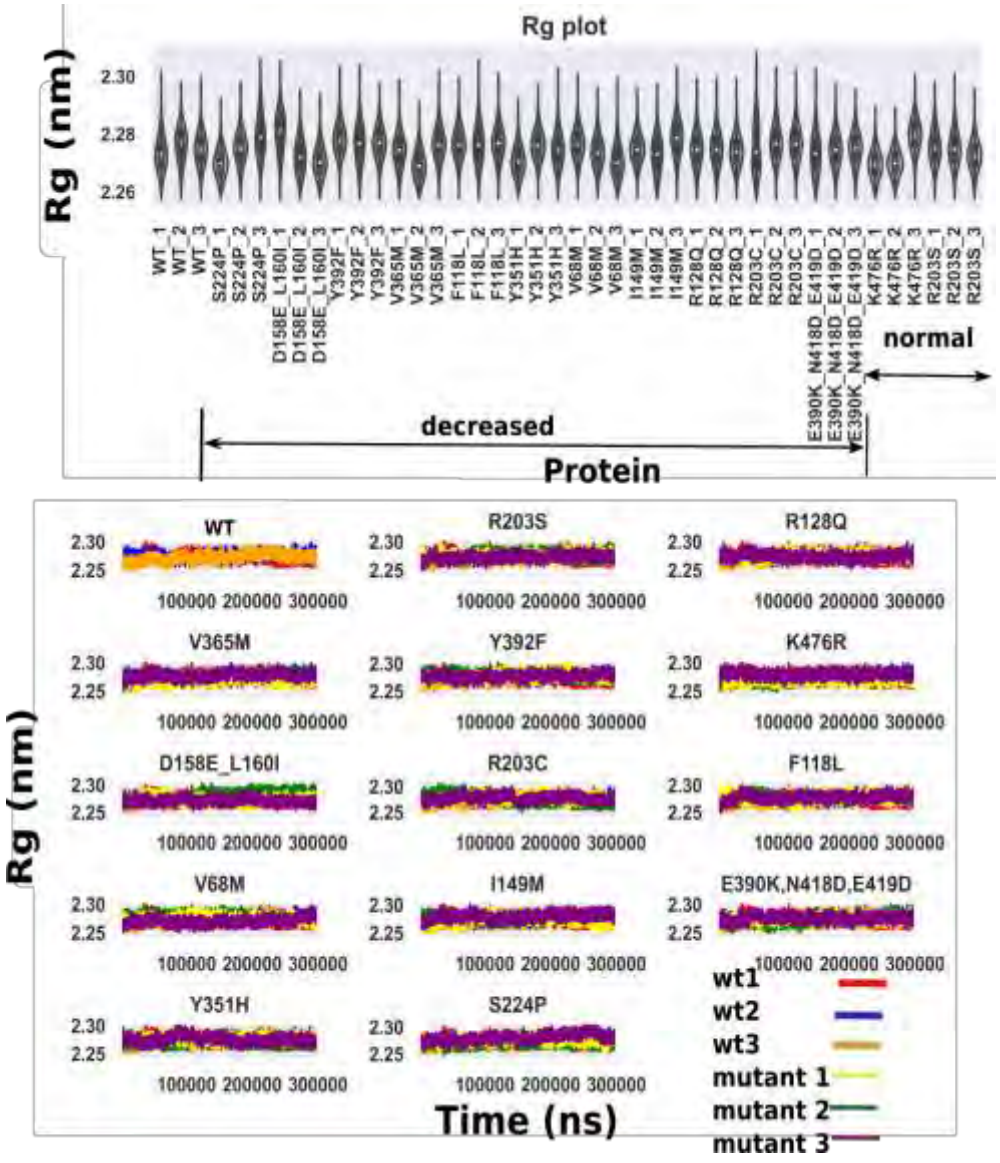


Figure S3. Rg plots to measure the compactness of residues (Berendsen results)

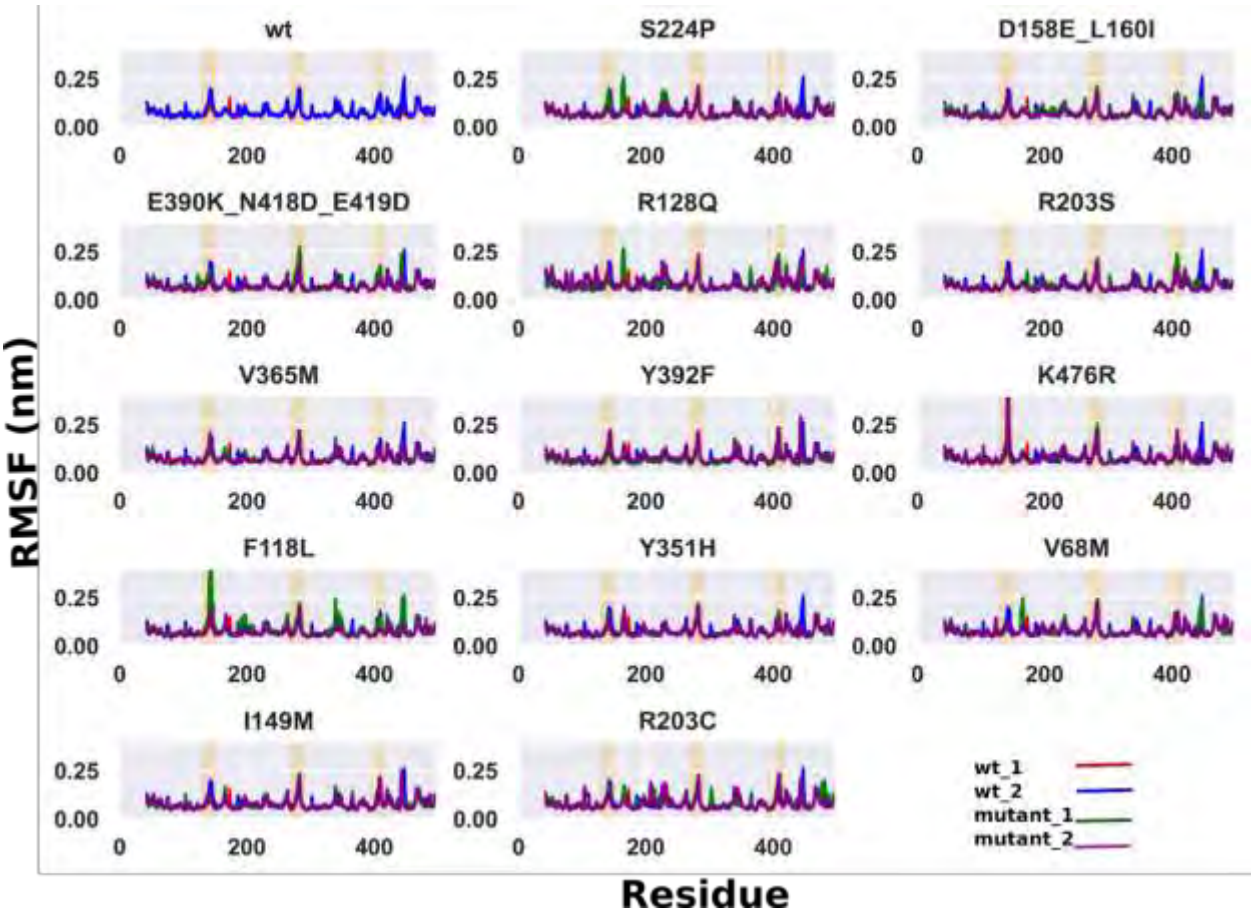


Figure S4. Fluctuations on residues (Parinello results). The most fluctuating regions are highlighted in orange

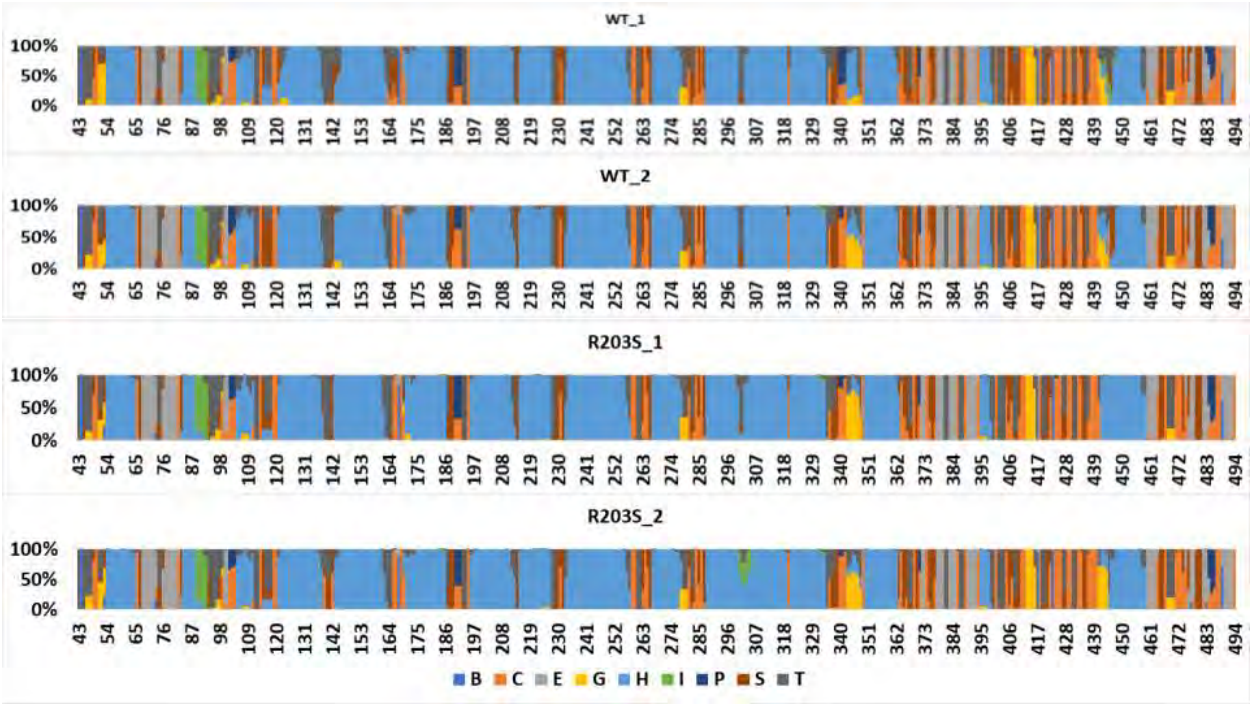


Figure S5. DSSP results for R203S. Secondary structure differences variant compared to the reference structure (WT_1, WT_2)

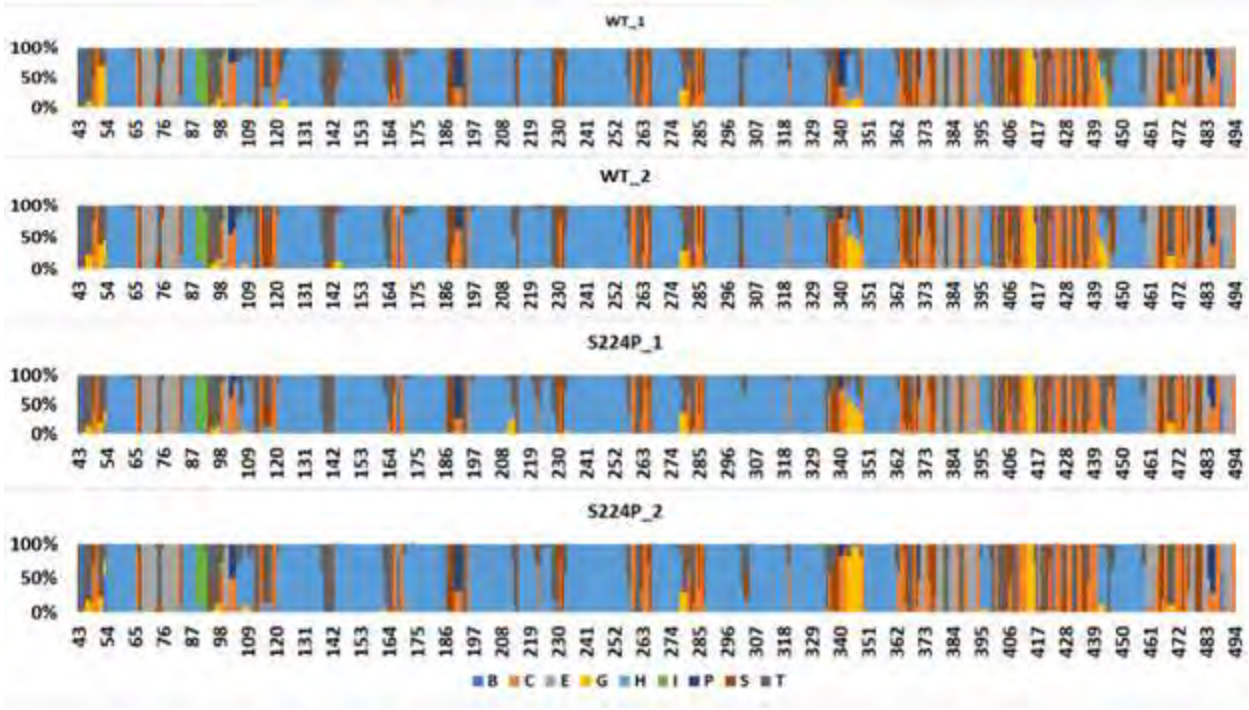


Figure S6. S224P DSSP results

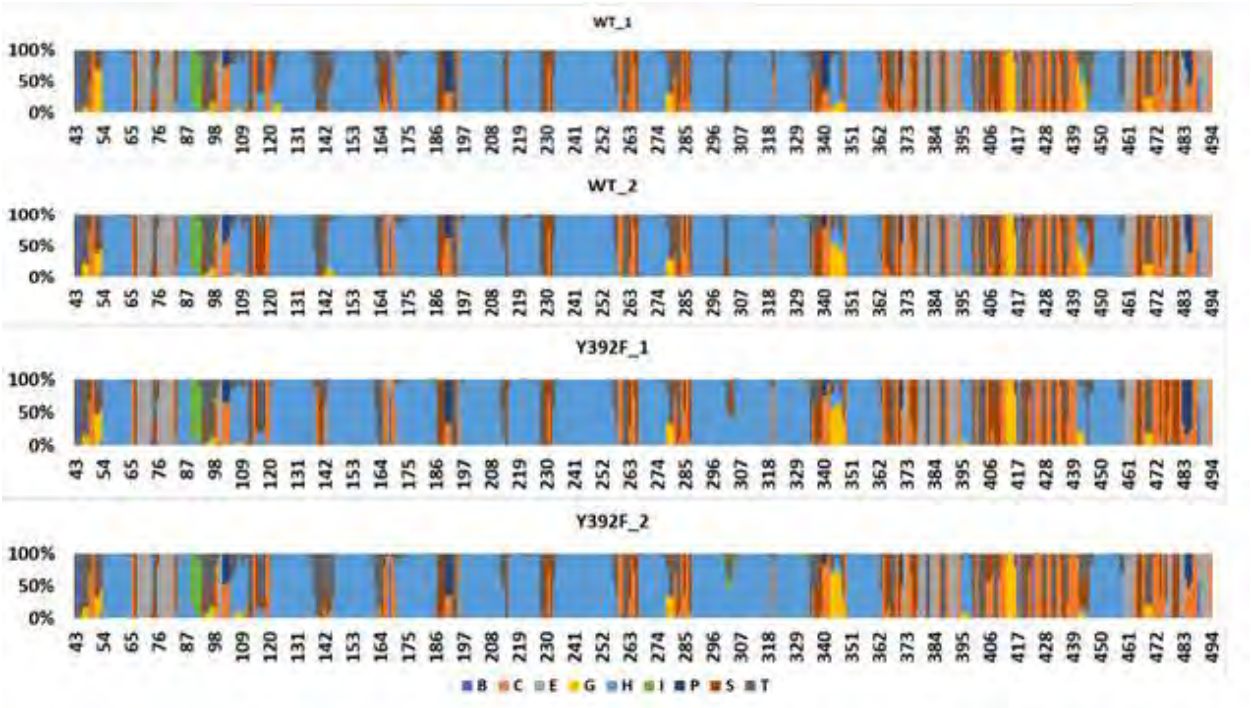


Figure S7. Y392F DSSP results

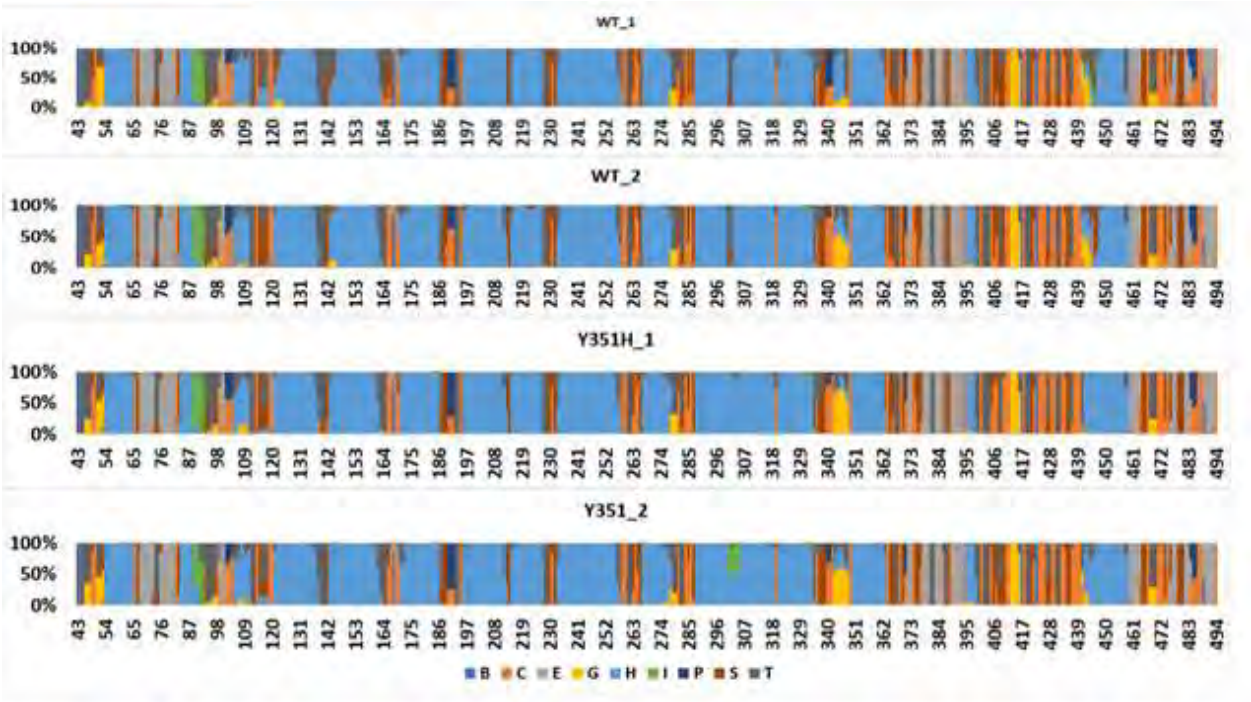


Figure S8. Y351H DSSP results

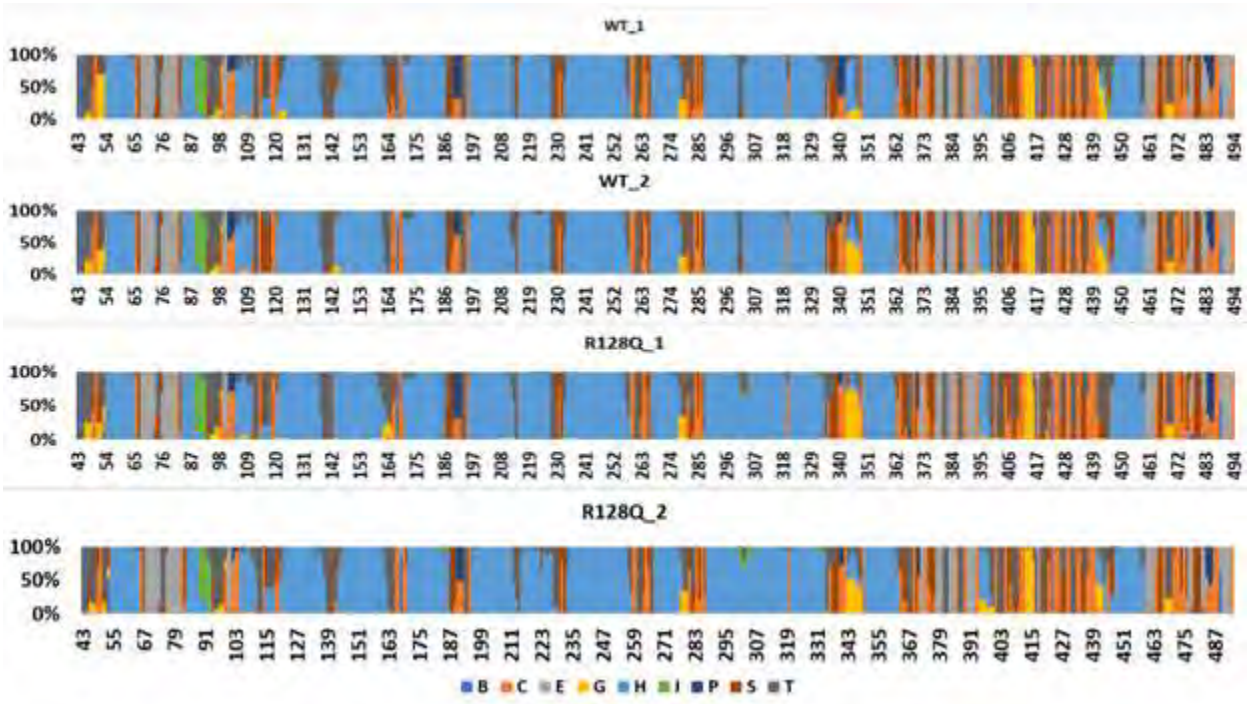


Figure S9. R128Q DSSP results

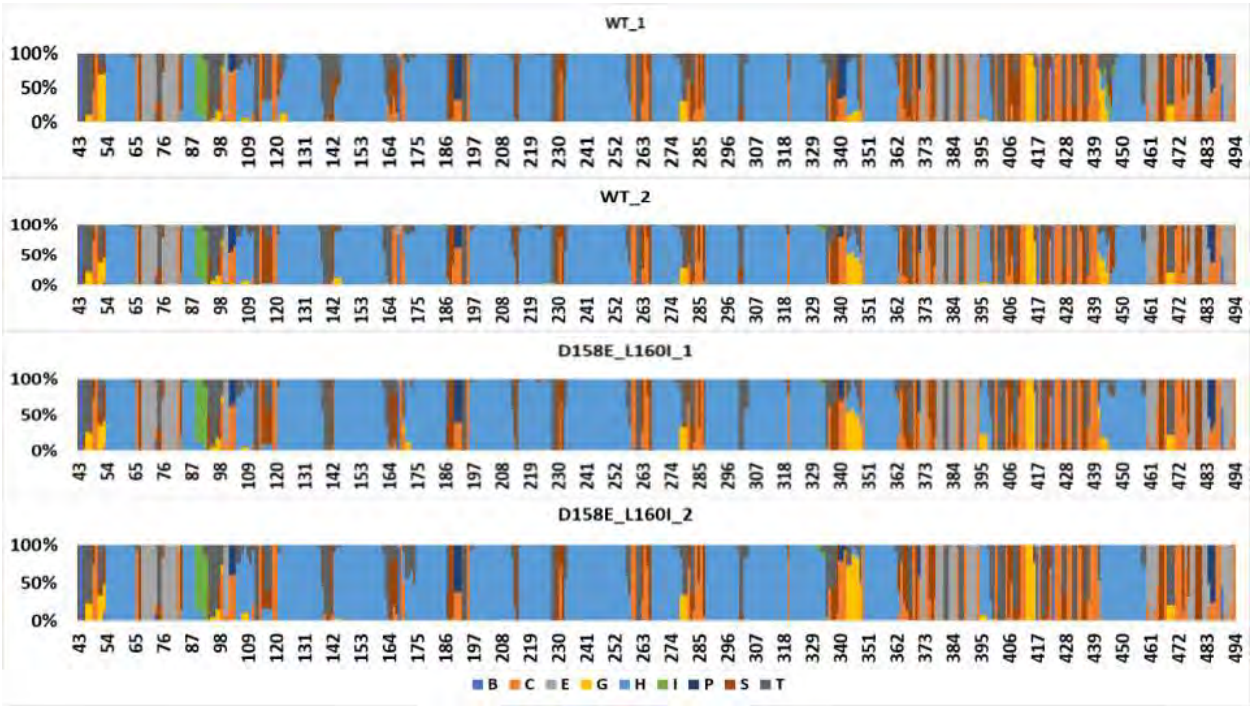


Figure S10. D158E_L160I DSSP results

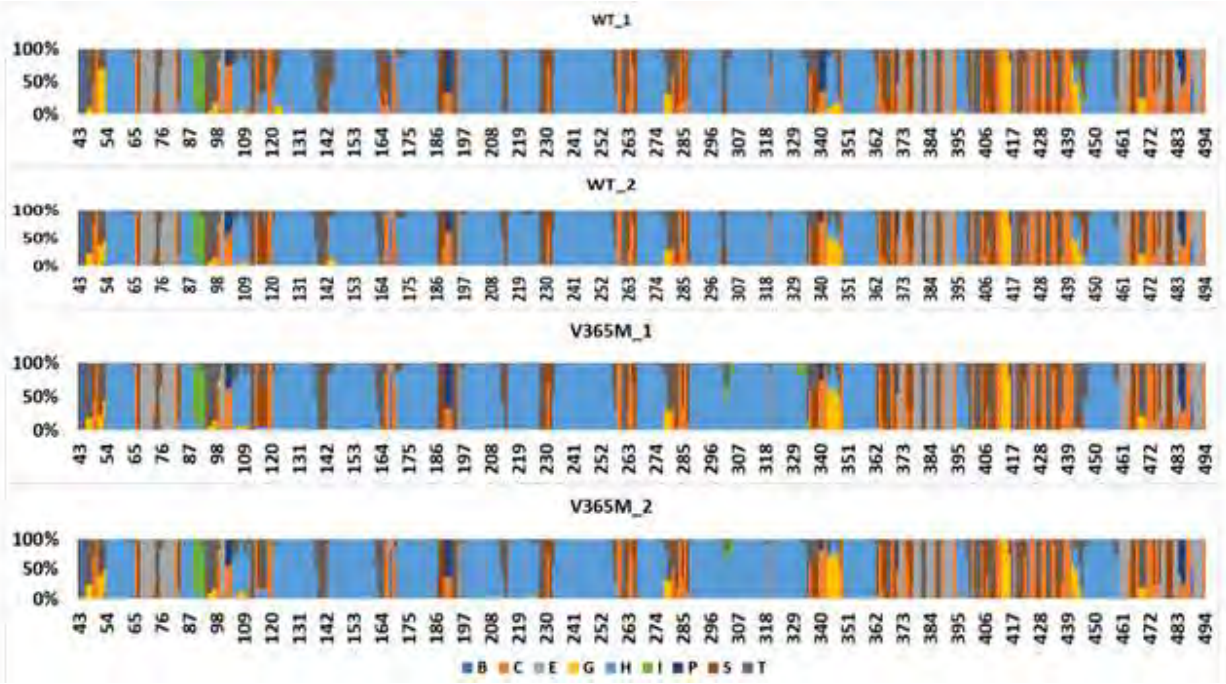


Figure S11. V365M DSSP results

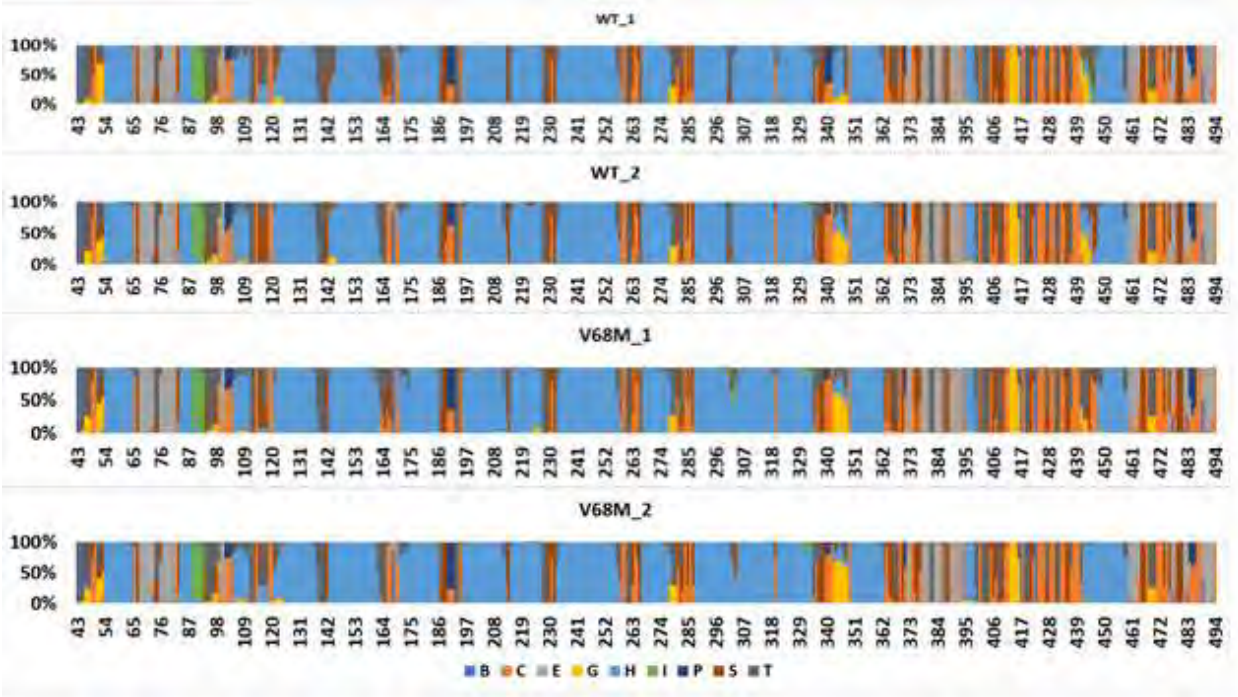


Figure S12. V68M DSSP results

TABLE S1.

Cluster percentages for each system

Protein	**Cluster 1 percentage**	**Cluster 2 percentage**	**Cluster 3 percentage**
WT_1	0.946	0.053	0.000
WT_2	0.913	0.059	0.028
D158E_L160I_1	0.997	0.003	0.000
D158E_L160I_2	0.836	0.123	0.041
E390K_N418D_E419D_1	0.977	0.012	0.011
E390K_N418D_E419D_2	0.993	0.045	0.021
F118L_1	0.979	0.020	0.001
F118L_2	0.929	0.052	0.019
I149M_1	0.946	0.052	0.002
I149M_2	0.972	0.028	0.000
K476R_2	0.990	0.005	0.005
K476R_2	0.992	0.008	0.000
R128Q_1	0.977	0.020	0.004
R128Q_2	0.590	0.409	0.001
R203C_1	0.995	0.005	0.000
R203C_2	0.999	0.001	0.000
R203S_1	0.972	0.016	0.012
R203S_2	0.998	0.002	0.000
S224P_1	0.982	0.018	0.000
S224P_2	0.975	0.025	0.000
V68M_1	0.965	0.035	0.000
V68M_2	0.976	0.023	0.001

V365M_1	0.989	0.007	0.004
V365M_2	0.940	0.060	0.000
Y351F_1	0.980	0.015	0.004
Y351F_2	1.000	0.000	0.000
Y392F_1	0.975	0.024	0.001
Y392F_2	0.890	0.086	0.024

TABLE S2.

RMSD values after aligning the WT and representative mutant clusters

System	**RMSD value system vs WT_1**	**RMSD value system vs WT_2**
WT_1	0	1.248
WT_2	1.248	0
D158E_L160I_1	1.294	1.229
D158E_L160I_2	1.213	1.292
Y351H_1	1.169	1.519
Y351H_2	1.112	1.327
E390K_N418D_E419_D_1	1.353	1.451
E390K_N418D_E419_D_2	1.090	1.307
S224P_1	1.341	1.613
S224P_2	1.215	1.512
R203C_1	1.228	1.295
R203C_2	1.267	1.472
R203S_1	1.164	1.351
R203S_2	1.187	1.249
V68M_1	1.137	1.259
V68M_2	0.968	1.387
R128Q_1	1.133	1.381
R128Q_2 cluster 1	1.436	1.617

R128Q_2 cluster 2	1.575	1.470
V365M_1	1.218	1.309
V365M_2	1.022	1.372
Y392F_1	1.251	1.269
Y392F_2	1.229	1.272
I149M_1	1.245	1.296
I149M_2	1.034	1.381
F118L_1	1.588	1.801
F118L_2	1.094	1.005
K476R_1	1.266	1.551
K476R_2	1.218	1.265