Structural Bioinformatics Analysis of CYP2D6 Pharmacogenetic Variation Relevant to Sub-Saharan African Populations

Blessing Rotondwa Sitabule

Student Number: 1439665

Supervisors:

Dr Houcemeddine Othman

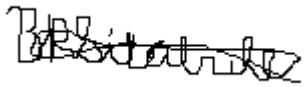
Prof Scott Hazelhurst

A dissertation submitted to the Faculty of Health Science, University of the Witwatersrand, Johannesburg, in fulfilment of the requirements for the degree of Master of Science in Medicine.

Johannesburg, April 2022

Declaration

I, Blessing Rotondwa Sitabule, declare that this dissertation is my own unaided work. It is being submitted as a degree of Master of Science in Medicine to the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination to any other University.



.....

(Signature of Candidate)

11th day of April 2022 in Johannesburg, South Africa.

Acknowledgements

To my supervisors, Houcemeddine and Scott, thank you very much for the support and guidance you provided me. Thank you for granting me an opportunity to work with you in this fascinating field of pharmacogenomics. The regular update meetings were helpful and provided a great platform for communication in the midst of the restrictions resulting from the COVID-19 pandemic.

To Jorge and David, thank you for the support and helping me to adapt to the new phase of my academic life. It was a great pleasure to be in a team with you guys. Thank you all my ADME team colleagues. Thanks to the NHLS, SBIMB and the university at large for granting me an opportunity to pursue this MSc project.

To my family, thank you for the financial, physical and emotional support as well as the motivations and prayers. Being part of the family is a great privilege. Thanks to my friends who have been supportive.

Thanks to my financial supporters:

To the Post Graduate Merit Award, thank you for funding my registration fees, tuition fees and providing me with quarterly stipends.

To the Wits Health Consortium, thank you for granting me with monthly stipends which were helpful with paying for conference fees, membership fees and purchasing data bundles for working online during the pandemic.

List of Presentations

17th H3Africa Consortium Meeting, 19-23 April 2021, Virtual Event

Oral presentation, Structural Bioinformatics Analysis of CYP2D6 pharmacogenetic variation relevant to Sub-Saharan African populations

ISCB Africa ASBCB 2021, the 8th bi-annual conference, was hosted by the International Society for Computational Biology (ISCB), 7-10 June 2021, Virtual Event

Oral presentation, Structural Bioinformatics Analysis of CYP2D6 pharmacogenetic variation relevant to Sub-Saharan African populations

Southern African Society for Human Genetics Young Researchers Online Symposium, 25-26 August 2021, Virtual Event

Oral presentation, Structural Bioinformatics Analysis of CYP2D6 pharmacogenetic variation relevant to Sub-Saharan African populations

MBRT Virtual Postgraduate Research Day, 9 December 2021, Virtual Event

Oral presentation, Structural Bioinformatics Analysis of CYP2D6 pharmacogenetic variation relevant to Sub-Saharan African populations

Abstract

Pharmacogenomics is a field of study that involves the association of genes involved in drug metabolism with drug response. The *CYP2D6* gene is important for drug metabolism. Different studies have identified a number of variations in the *CYP2D6* gene in African populations with potential functional impact. The aim was to gain insights of how missense variants found in sub-Saharan African populations may potentially impact the functionality of the CYP2D6 enzyme, with focus on a drug commonly used in Africa. Fifty missense variants identified in African populations were selected using the PharmVAR, and GnomAD databases. Missense variants were prioritised for molecular dynamics using the Structural Workflow for Annotating ADME gene Targets (*SWAAT*) and the H3Africa dataset to identify variants that are more common in Africa and have a potential significant impact on drug metabolism. Missense variants which were exceptions to the prioritisation criterion were also selected for molecular dynamics assessments. The complex CYP2D6/thioridazines structure was used to run the molecular dynamics simulations as the reference structure and for the selected variants. A total of 10 missense variants were selected for molecular dynamics assessment. From the 10, only Y355C, P34S and R365H were destabilising according to all the SWAAT features. The MD results showed how the 10 missense variants may affect the stability, flexibility, secondary structure and mobility of the enzyme. These findings may be used to expand our knowledge in pharmacogenomics which may be used to enhance precision medicine in Africa.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ADME	Absorption Distribution Metabolism Excretion
ADR	Adverse Drug Reaction
AFR	Africans
AMBER	Assisted Model Building with Energy Refinement
AMR	Admixed Americans
CHARMM	Chemistry at Harvard Macromolecular Mechanics
CYP	Cytochrome P450
DSSP	Dictionary of Secondary Structure Prediction
EAS	East Asians
EUR	Europeans
GROMOS	GROningen MOlecular Simulation
GSK	GlaxoSmithKline
H3Africa	Human Heredity and Health in Africa Initiative
ICSR	Individual Case Safety Report
IM	Intermediate Metaboliser
MD	Molecular Dynamics
ML	Machine Learning
MM	Molecular Mechanics
NM	Normal Metaboliser
OPLS	Optimised Potentials for Liquid Simulations
PCA	Principal Component Analysis
PGx	Pharmacogenomics
PKA	Protein Kinase A
PM	Poor Metaboliser
QM	Quantum Mechanics
RMSD	Root Mean Square Deviation
RMSF	Root Mean Square Fluctuation
SAS	South Asians
SB	Structural Bioinformatics
SNP	Single Nucleotide Polymorphism
SSA	Secondary Structure Analysis

SWAAT	Structural Workflow for Annotating ADME gene Targets
UM	Ultra-rapid Metaboliser
VMD	Visual Molecular Dynamics
VUS	Variant of Unknown Significance
WT	Wildtype

Chapter 1: Introduction

Precision medicine is an emerging approach that has been implemented to optimise the efficiency of medicine and reduce adverse drug reactions. Several factors are considered when implementing precision medicine, these include, genetic factors, environmental factors, age, lifestyle and sex. Genetic factors have shown to contribute to the implications of administered drugs on patients. The aim of relating an individual's genotype and drug response is referred to as pharmacogenomics. Pharmacogenomics studies have been performed to enhance the efficiency of precision medicine; however, African populations have been underrepresented. As a result, more studies are required for African populations to enhance the effectiveness of precision medicine in Africa. Among a number of genes involved in drug metabolism, the *CYP2D6* gene plays a critical role in the metabolism of a number of drugs such as anti-cancer drugs and analgesics which are commonly used in Africa. This study provides an insight into the implications of some missense variants on the activity of the CYP2D6 enzyme using *structural bioinformatics* approaches which may provide a better understanding on the implications of the variants on the phenotype of the individual. Moreover, the study identifies potentially damaging missense variants which may be potentially used as biomarkers for treatment guidelines in African populations, thus, enhancing precision medicine applications in Africa.

1.1 Precision Medicine

Beneficence and non-maleficence are two bioethics principles that should guide in health care to optimise therapeutic benefit and minimise toxicity from medical treatment (Beauchamp and Childress, 2019). In South Africa, an observational study reported that approximately 6.3% of admitted patients were hospitalised due to adverse drug reactions (ADRs), while an additional 6.3% of patients experienced significant ADRs within the hospital (Terblanche, 2018). Table 1.1 shows the number of individual case safety reports (ICSRs) in several African countries and the percentage of the total ICSRs in Africa from VigiBase by 2015 (Ampadu *et al*., 2016). In addition, 3.5% of admissions in hospitals resulted from ADRs and ~197,000 deaths occurred annually in Europe as a result of ADRs (Khalil and Huang, 2020). Furthermore, according to Duncan *et al.,* (2020), there are a number of drugs that have been reported to be potentially non-beneficial to some patients.

Precision medicine is a new approach which large amount of data generated daily in the healthcare system are used to provide the most effective treatment or preventive care, at the right time, to the patients that will benefit the most from it (Gameiro *et al*., 2018). This method incorporates a number of factors that are considered prior to the treatment process (Zia *et al*., 2013). These factors include the patient's age, gender, lifestyle and genome. As a result of these factors, individuals with the same condition may respond differently to the same drug as depicted in Figure 1.1 (Gill *et al*., 2021).

Table 1.1: Reported ICSRs in some African countries.

Country	**No. of ICSRs from VigiBase**	**Percentage of the total ICSRs in Africa according to VigiBase**
South Africa	28 609	27.64
Morocco	17 231	16.65
Nigeria	10 590	10.23
Egypt	8 474	8.19
Kenya	8 440	8.15
Tunisia	6 990	6.75
Democratic Republic of Congo	5 558	5.37
Ghana	2 900	2.80
Zimbabwe	2 155	2.08
Eritrea	1 982	1.91

(Ampadu *et al*., 2016)

Pharmacogenomics (PGx) aims to relate genetic variation with drug response amongst different individuals with studies that focus on ADME genes, which are a group of genes that encode proteins involved in the absorption, distribution, metabolism and excretion of drugs (SkaricJuric *et al*., 2018; Madian *et al*., 2012).

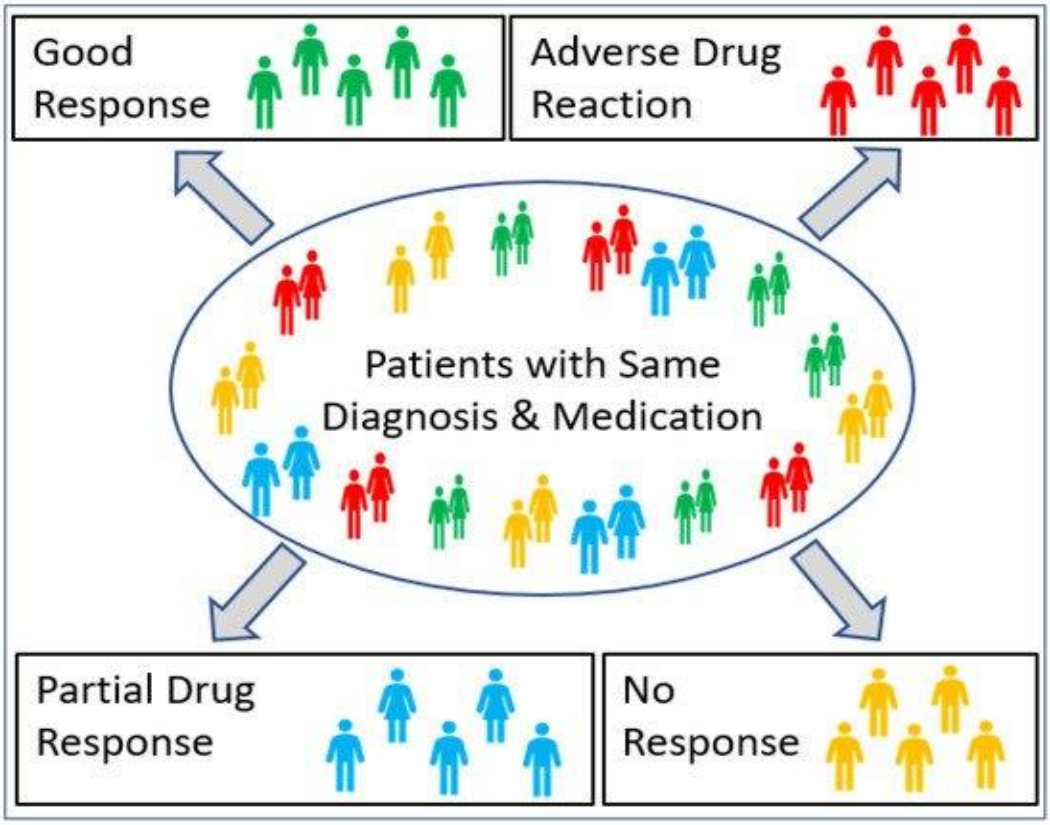


Figure 1.1: Diagram showing how a group of individuals with the same diagnosis and same medication may respond differently to the particular treatment. The possible outcomes of the treatment include a good response (shown on the top left subgroup), partial drug response (shown on the bottom left subgroup), adverse drug reaction (shown on the top right) and no response (shown on the bottom right) (Figure was taken from Gill *et al*., 2021) https://creativecommons.org/licenses/by/4.0/.

1.2 Cytochrome P450 Enzyme Roles and the CYP2D6 Genetic Variants

Among the proteins involved in drug metabolism, a group of enzymes called the Cytochrome P450s (CYPs) are synthesised mainly in the liver and several other tissues. They are divided into 18 families and 43 subfamilies according to the similarity of their sequences. Families 1, 2 and 3 play a role in the metabolism of drugs and xenobiotics (Preissner *et al*., 2013). These enzymes are membrane-attached proteins that play a role in numerous biotransformation processes within the endoplasmic reticulum and mitochondria. Figure 1.2 illustrates how a membrane-attached Cytochrome P450 enzyme (CYP3A4) may interact with the membrane. The membranes influence the behaviour of these proteins which may change the opening pattern of the access or egress channels as well as the substrate transport into or out of the enzyme's active site and the interaction with redox partners facilitated by electrostatic interactions (Srejber *et al*., 2018). Furthermore, most of these enzymes metabolise more than one drug (Figure 1.3b) and some drugs could be metabolised by more than one CYP. In addition, only CYPs are responsible for activating prodrugs which are drugs that are

converted into active drugs prior to having a pharmacological effect (Preissner *et al*., 2013). Cytochrome P450 2D6 (CYP2D6) has the highest number of single nucleotide polymorphisms (SNPs) and metabolises the second highest number of drugs Figure 1.3a.

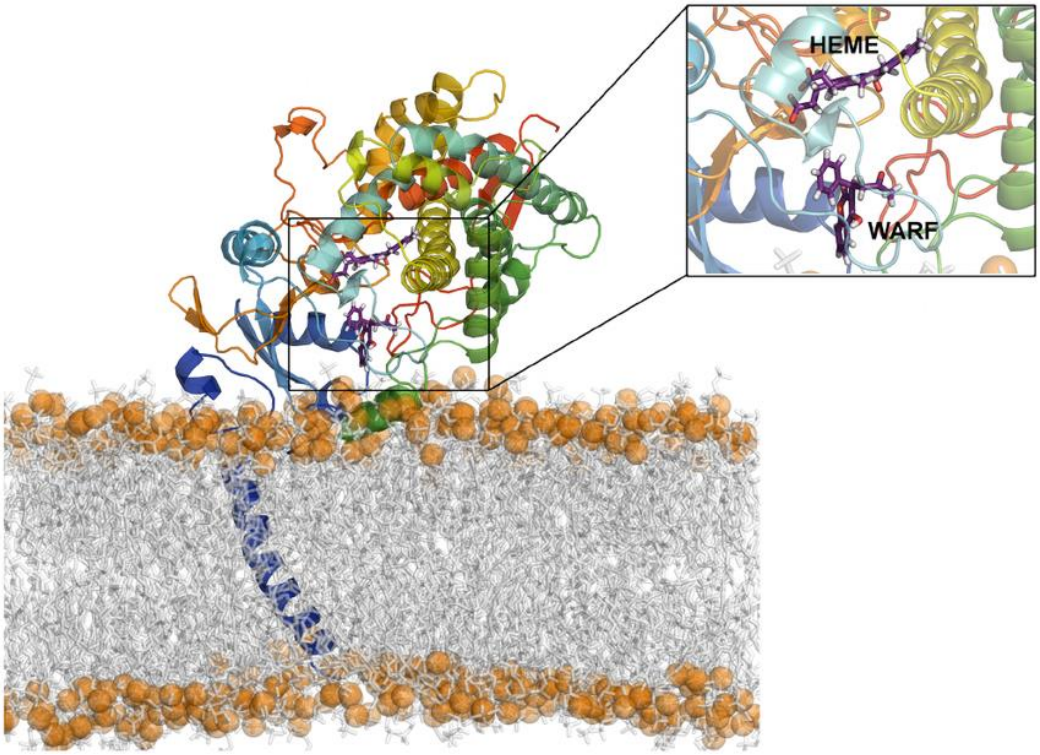


Figure 1.2: Diagram showing a membrane-attached cytochrome P450 enzyme (CYP3A4). Also shown is the enzyme's substrate, WARF (warfarin), and the heme group (Figure was taken from Lonsdale *et al*., 2014).

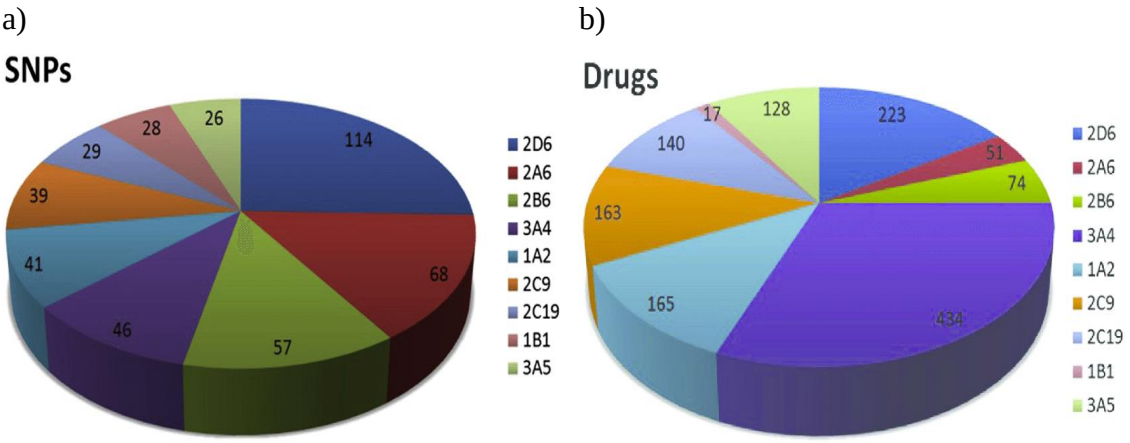


Figure 1.3: Diagram showing the number of known SNPs per CYP (shown on the left) and the number of drugs metabolised per CYP (shown on the right). The enzyme with the highest number of known SNPs is CYP2D6 (shown in a) and the enzyme with the highest number of substrates (drugs), shown in b, is CYP3A4 (Figure was taken from Preissner *et al*., 2013).

Most genetic studies have been conducted on non-African populations, therefore, leading to a dearth of data pertaining to African populations. In addition, African populations show a high genetic diversity (Tucci and Akey, 2019). For these reasons, this study focuses on African populations to enhance the efficiency of precision medicine in African populations by expanding the knowledge of pharmacogenomics in the context of Africa.

CYP2D6 is involved in drug metabolisation of several drugs, members of different families (Table 1.2) including, but not limited to, anti-cancer, anti-psychotic and analgesic drugs (Wang *et al*., 2015; Taylor *et al*., 2020). Furthermore, drugs such as metoprolol are mainly metabolised by CYP2D6 while drugs such as verapamil may be alternatively metabolised by another CYP450 enzyme known as CYP2C8 (Whirl-Carrillo *et al*., 2012).

Table 1.2: A list of some CYP2D6 substrates

Substrates	**Drug family**
Codeine	Analgesic
Desipramine	Anti-depressant
Fluoxetine	Anti-depressant
Imipramine	Anti-depressant
Metoprolol	β-blocker
Tamoxifen	Anti-cancer
Thioridazine	Anti-psychotic
Zuclopenthixol	Anti-psychotic

(Wang *et al*., 2015; Taylor *et al*., 2020).

Several studies have reported how genetic variants of the *CYP2D6* gene impact the functionality of the enzyme. The four possible phenotypes for drug response are ultra-rapid metabolisers (UM), normal metabolisers (NM), intermediate metabolisers (IM) and poor metabolisers (PM) (Jarvis *et al*., 2019). These phenotypes occur as a result of a combination of alleles which are usually classified according to the star-allele nomenclature which is a system that is commonly used for the designation of the ADME gene alleles. In many cases, the *1 denotes the wildtype (WT) allele while other star alleles such as *2 and *3 represent alleles that harbour at least one variant (Kalman *et al*., 2017).

As shown in Table 1.3, different types of variants have been reported in the CYP2D6 enzyme. These include frameshift, stop gain, splicing defects and missense variants. The *CYP2D6* star alleles are characterised by core variants which are also shown in Table 1.3. African populations have a high frequency of the *17 and *29 alleles and these are characterised by missense variants (Zhou *et al*., 2017). Furthermore, missense variants account for the highest portion of variants with unknown significance and variants that have been misclassified in terms of their implications on the phenotype (Dines *et al*., 2020). Moreover, the *CYP2D6* *74 and *73, which have been reported in sub-Saharan Africa, are defined by missense variants and are yet to be characterised (Gaedigk *et al*., 2018; Whirl-Carillo *et al*., 2012). In this study, we focused on missense variants to enhance our understanding on their potential impact on the functionality of the CYP2D6 enzyme and to expand on our knowledge that may provide insight into the significance of the variants.

Table 1.3: Allele frequencies of important CYP2D6 alleles in different populations and their implications on the functionality of the enzyme

Star allele	Variant type (Core variants)	Allele frequencies in indicated populations (%)					Functional consequences
		EUR	AFR	EAS	SAS	AMR	
*1	No variants	33.1	9.3	13.6	25.8	40.2	Normal
*1xN	*1 duplicated	1.0	3.3	1.0	0.5	0.5	Increased
*2	Missense (R296C, S486T)	34.3	26.7	14.0	36.2	32.7	Normal
*2xN	Missense (R296C, S486T) duplicated	1.3	6.0	1.0	1.0	0.5	Increased
*3	Frameshift	4.1	0.3	0.0	0.1	0.3	Inactive
*4	Splicing defect	15.5	11.9	0.4	11.6	15.7	Inactive
*5	CYP2D6 deleted	3.0	4.0	6.5	2.0	3.0	Inactive
*6	Frameshift	2.2	0.3	0.0	0.1	0.4	Inactive
*7	Missense (H324P)	0.0	<0.1	0.0	0.8	<0.1	Inactive
*8	Stop-gain (G169X)	0.0	<0.1	0.0	<0.1	0.0	Inactive
*9	Inframe deletion (K281del)	1.6	0.4	0.0	0.2	1.3	Decreased

*10	Missense (P34S, S486T)	0.2	3.2	58.7	6.5	0.0	Decreased
*11	Splicing defect	0.0	<0.1	0.0	0.0	0.0	Inactive
*12	Missense (G42R)	0.0	<0.1	0.0	0.0	0.0	Inactive
*14	Missense (G169R)	0.0	0.0	1.6	<0.1	0.0	Inactive
*17	Missense (R296C, T107I)	<0.1	19.7	0.0	0.1	0.7	Decreased
*29	Missense (R296C, S486T, V136I, V338M)	0.0	9.2	<0.1	<0.1	0.4	Decreased
*33	Missense (A237S)	0.7	0.2	0.0	0.7	0.1	Normal
*41	Splicing defect	3.0	3.0	3.0	13.5	3.5	Decreased
*42	Frameshift	0.0	0.2	0.0	0.0	<0.1	Inactive
*43	Missense (R26H)	<0.1	2.0	<0.1	0.8	0.2	Uncertain implication
*53	Missense (F120I, A122S)	0.0	<0.1	<0.1	<0.1	0.5	Increased
*62	Missense (R441C)	<0.1	<0.1	<0.1	<0.1	<0.1	Inactive

EUR, Europeans; AFR, Africans; EAS, East Asians; SAS, South Asians; AMR, Admixed Americans. (Zhou *et al*., 2017).

1.3 CYP2D6 Structure and Critical Residues

Protein function depends on the protein's structure and stability which depend on its sequence (Li and Koehl 2014). As a result, missense variants may potentially affect the protein's functionality as a result of altering the amino acid sequence, this however, depends on the nature of the mutation such as the location of the mutation and the impact of the amino acid change on the chemical property of the protein (Li and Koehl 2014). This is due to the different properties of amino acids which results in some amino acids being critical to the protein's structure and function. Figure 1.4 shows the structure of the CYP2D6 enzyme including the FG loop and the BC loop. The FG and BC loops function as a 'lid' over the active site cavity (Srejber *et al*., 2018).

Amino acids have unique chemical properties which include polarity, charge and other characteristics of their side chains i.e chain length and whether an aromatic ring is present or

not (Vnučec *et al.,* 2016). These unique characteristics are crucial for the roles played by some of the amino acids in protein structure and function (Ittisoponpisan *et al*., 2019). Some of these amino acids, such as cysteine, serine, histidine and threonine, may be crucial for catalysis, binding affinity for the substrate or serve as an important component of the protein structure and protein function (Holliday *et al*., 2009). These amino acids are referred to as critical residues and missense mutations that occur in these residues usually result in a significant impact on the functionality of the protein (Thibert *et al*., 2005).

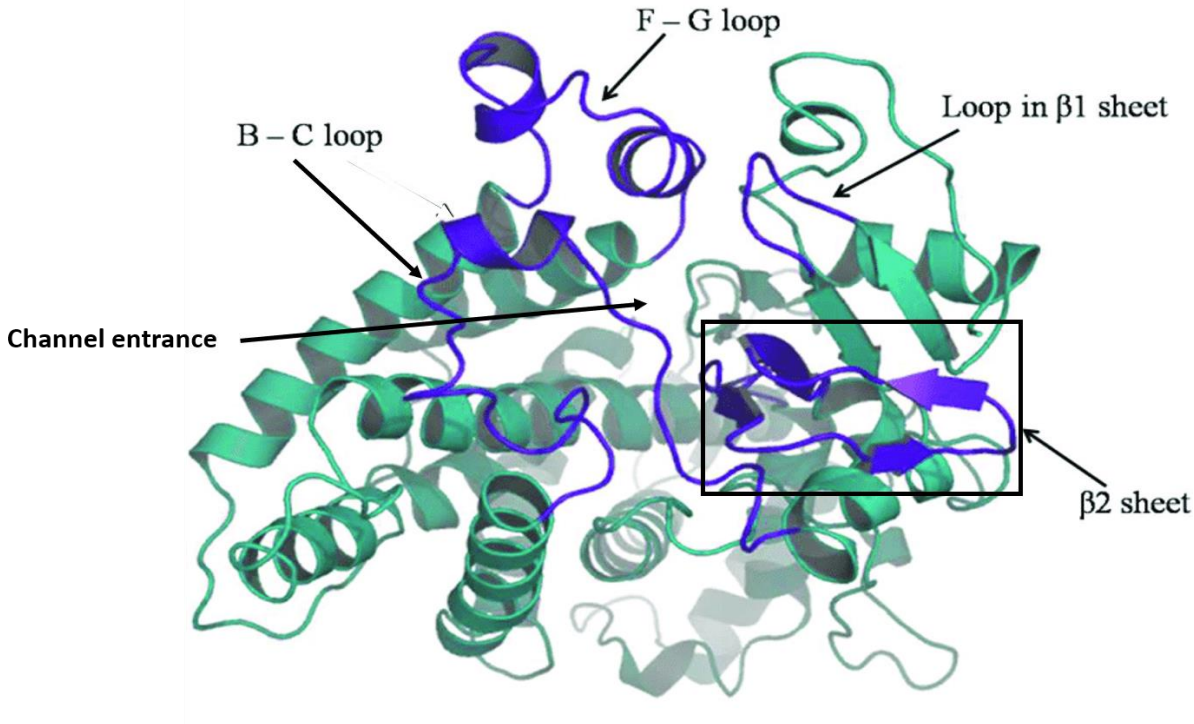


Figure 1.4: Diagram of the CYP2D6 structure showing the channel entrance indicated with an arrow. The FG and BC loops and the β-sheets are labelled. The PDB ID for the structure is 3QM4 (Figure adapted from Xin *et al*., 2020**)** https://creativecommons.org/licenses/by/4.0/.

As serine, threonine and tyrosine possess hydroxyl groups, these amino acids may serve as regulatory sites for proteins since they are phosphor-acceptors (Miao *et al*., 2018). As a result, the substitution of a phosphor-acceptor amino acid (which serves as a regulatory site) with an amino acid that lacks a hydroxyl group may potentially impact the functionality of the protein since the regulation of the protein would be affected by the amino acid change. It has been reported that the CYP450 enzymes undergo post-translational modifications (Lamb and Waterman, 2013). According to Oesch-Bartlomowicz and Oesch (2005), when a wild-type phosphor-acceptor Ser129 in a rabbit CYP2E1 protein kinase A (PKA) recognition motif Arg-Arg-Phe-Ser is absent, the regulation (via PKA-mediated phosphorylation) will take place on another site which enhances the catalytic activity which would be the opposite effect

as compared to when Ser129 is phosphorylated. Thus, missense mutations in CYP2D6 phosphorylation sites such as Ser135 may affect the enzyme's activity.

In addition to some amino acids being critical for protein function through forming a disulphide bond or serving as a regulatory site, there are other critical residues which may impact the protein function directly. These include amino acids that are located in active sites and play a key role in the catalysis of substrates by enzymes. Amino acids such as histidine may play a significant role in the catalysis of a substrate in enzymes such as carboxypeptidases and serine proteases (Bartlett *et al*., 2002). Furthermore, certain amino acids may facilitate the binding of the substrate to the enzyme. According to Marechal *et al*., (2008) Glu216 and Asp301 play a significant role in the binding of basic substrates in CYP2D6, as basic substrates require acidic residues. The basic nitrogen of bufuralol and dextromethorphan forms an ionic interaction with a negatively charged carboxylate group (Paine *et al*., 2003). The basic nitrogen of the substrates is known to lie close to the negative charge of the Glu216 carboxylate, but it is relatively distant from the one of Asp301 (Paine *et al*., 2003). Furthermore, Val370 and Phe483 are also known to be involved in substrate binding (Ingelman-Sundberg, 2005; Ito *et al*., 2008). Thus, missense mutations that occur in these residues may potentially impact the functionality of the enzyme, since the binding affinity of the substrate may be reduced (Huff *et al*., 2021). Figure 1.5 shows the CYP2D6 and its active site including the amino acids that interact with substrate.

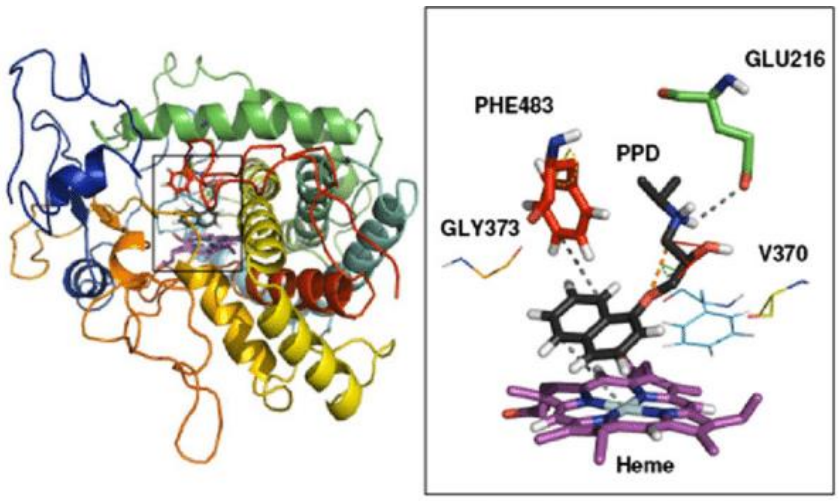


Figure 1.5: Diagram of the CYP2D6 and its active centre. The residues that form part of the active centre including, but not limited to, Phe483, Gly373, Val370 and Glu216 are shown. Also shown are the heme group and substrate propranolol (PPD) (Figure was taken from Nagy and Oostenbrink, 2012) https://creativecommons.org/licenses/by/4.0/.

Other catalytic roles, played by amino acids, include stabilisation, steric role, covalent catalysis, proton shuttling, hydrogen shuttling, electron shuttling and shifting the redox potential and nucleophilicity or electrophilicity of a reacting species (Holliday *et al.*, 2009). Missense mutations may also lead to damaging effects in other ways, such as, altering the hydrophobic protein core, altering the charge on the core residues of the protein, introducing a proline residue in the core region of the protein and by altering the van der Waals forces, consequently, affecting the structure of the protein significantly and ultimately its function (Ittisoponpisan *et al*., 2019). A number of methods are used to study the functional effects of variants such as *in vitro* and *in silico* studies.

1.4 In Silico Approach: Structural Bioinformatics

In silico studies involve experimentation that is conducted on computers to make simulations or models. This has advantages such as low cost, the ability of a quick execution and does not require biological samples (Andrade *et al*., 2016). *Structural bioinformatics* (*SB*), which is an *in silico* approach, is a bioinformatics category that involves the prediction of 3D structure of biological macromolecules such as proteins as well as its structure-function relationship (Choong *et al*., 2013). Multiscale molecular modelling is a *structural bioinformatics* approach which relies on the structure of the protein of interest to conduct the simulation and assessment. In general, experimental structures obtained from techniques such as x-ray crystallography and Nuclear Magnetic Resonance (NMR) are better than predictions obtained from theoretical building of the protein 3D coordinates. Furthermore, results and analysis from *in silico* approaches may be validated and strengthened by other experimental methods including but not limited to, mutagenesis assays, microcalorimetry, binding assays and knock-out assays (Frohlich and Salae-Behzadi, 2014).

Multiscale methods approaches are used in biological studies to determine how events, such as single nucleotide mutations, affect a phenotype by using structurally-based physicochemical models for the integration of temporal and spatial scales of biology. This improves the mechanistic understanding of the processes and their implications (Brown and Bishop, 2018; Boras *et al*., 2015). In the context of this study, the term "*structural bioinformatics"* is used interchangeably with "Multiscale methods" to describe the set of methods that uses biological macromolecular structures to run a computational analysis on mutations and intermolecular interactions. Thus, multiscale methods can be used to predict

the potential effect of a mutation which can be further assessed using approaches such as molecular docking and molecular dynamics (Boras *et al*., 2015).

A computational technique called molecular dynamics (MD) can be used to simulate the dynamic behaviour of the molecules as a function of time, with all entities (molecules and biological molecules) in a simulation box being regarded as flexible (Salmaso and Moror, 2018). MD involves the simulation of a classical system of *N* particles which is based on Molecular Mechanics (MM) physics (Santos *et al*., 2019). Albeit, MD simulations may also be performed on the basis of Quantum Mechanics (QM) which incorporates the direct electronic effects in molecules which are ignored by MM (Santos *et al*., 2019). The basis of many simulations is to begin with the initial positions and velocities for every particle then followed by the repeated application of an algorithm which updates the particles velocity and position from the initial time to the final time (Vollmayr-Lee, 2019). The dynamics is governed by Newton's second law of motion (Vollmayr-Lee, 2019). MD simulates the motions of particles that result from the application of molecular mechanics by solving Newton's equation of motion over a specific simulation time and by calculating the potential energy using one of the available force fields such as CHARMM, AMBER, OPLS and GROMOS (Lopes *et al*., 2015). Molecular dynamics yields a set of conformational ensembles of the protein called trajectories (Boras *et al*., 2015). Trajectories are sequential snapshots of a simulated molecular system which is a representation of atomic coordinates at a particular point in time (Likhachev *et al*., 2016). The trajectories are generated by following a number of steps as shown in Figure 1.6.

MD has several applications such as assessing conformational stability and flexibility, dynamic processes of macromolecules in a period of time, the impact of allelic variants on the protein structure, the manner in which binding of a substrate occurs on a protein and other physical properties of a protein (Hollingsworth and Dror, 2018; de Waal *et al*., 2014). MD can be used to understand the effect of variants by comparing the simulation of the native as well as the mutant protein (Kamaraj and Purohit, 2013). This approach reveals how non-synonymous variants affect the structure of a protein and consequently the phenotype of the organism (Abduljaleel, 2019). Moreover, MD has been used in PGx studies by evaluating the impact of variants on the enzyme's flexibility, stability and motion which have implications on the enzyme's functionality (Marquez *et al*., 2019).

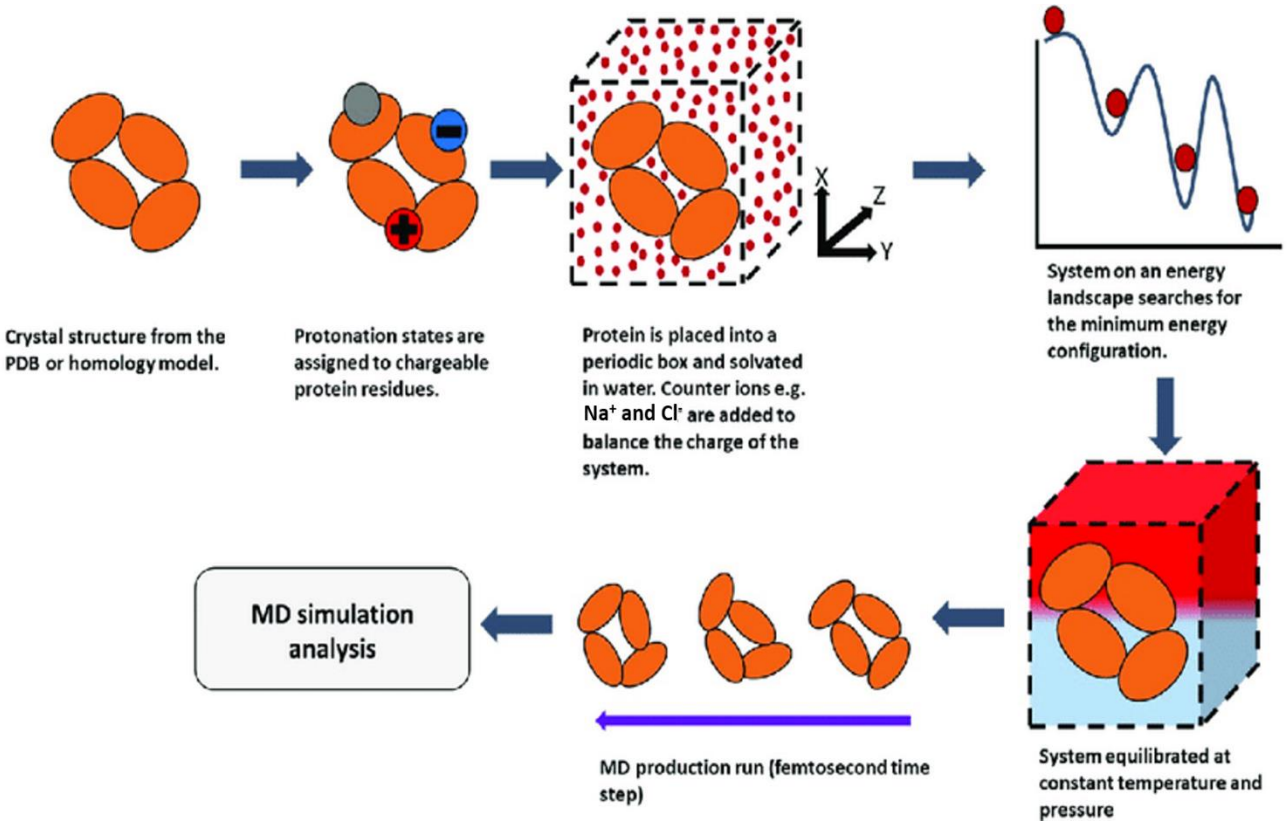


Figure 1.6: A schematic diagram of the MD steps. The first step involves the selection of a crystal structure for protein of interest from PDB (preferably a structure with a high resolution) or modelling a protein using a homologous structure through an approach called homology modelling. Protonation states are then assigned on amino acids with ionisable groups based on the pH level of the protein's environment. The protein is solvated with water and then the system is neutralised using either Na^+ or Cl^-. The system is minimised to obtain a minimum energy configuration, then this is followed by the equilibration of the system at a constant temperature. The MD simulation is then performed for a particular period and analysis is performed after generating the trajectories (Figure was taken from Pandya *et al*., 2018).

Chapter 2: Study Rationale and Aim and Objectives

2.1 Study Rationale

Precision medicine is an emerging approach that aims to give the right drug to the right person at the right time which reduces the occurrence of adverse drug events. To achieve this, pharmacogenomics is one of the methods that has been established. This relates an individual's genetic makeup to drug response and serves to reduce costs of trial and error for identifying the best available drug for the patient and reduce burdening hospitals with patients that are admitted due to ADRs which could be avoided.

African populations have been shown to have a high genetic diversity; however, they have also been underrepresented in genetic studies. For this reason, more studies are required in African populations to enhance the effectiveness of precision medicine among populations in Africa. Missense variants account for a high proportion of VUSs and misclassified variants. Furthermore, missense variants also form part of a number of pathogenic variants which have been implicated in CYP2D6, an enzyme that is involved in metabolising a variety of drugs. As a result, acquiring a better understanding of how missense variants impact the functionality of the enzyme is essential for knowing the implications at the phenotype level.

Tamoxifen and thioridazine are examples of various drugs metabolised by CYP2D6 which are commonly used in Africa. Some CYP2D6 alleles such as *17, which has missense variants as the core variants, have been suggested to result in a decreased activity of the CYP2D6 enzyme and occur mainly in African populations. Thus, we sought to gain a better understanding of how missense variants result in a significant impact at molecular the level and predict the possible implications of uncharacterised star alleles defined by missense variants relevant in sub-Saharan Africa.

To achieve this, we used multiscale modelling, PharmVar and molecular dynamics to select and assess the impact of missense variants (found in sub-Saharan African populations) that have a potential significant effect. This provides a better understanding on how missense variants may potentially result in a significant effect on the functionality of the enzyme and provides a guide on identifying variants that may potentially have significant implications on drug response. In addition, this may also provide insight into the implications of future

discovered novel star alleles that consist of the implicated missense variants. Thus, this will contribute to the growth of knowledge of pharmacogenomics in Africa and enhancing the efficiency of precision medicine in Africa.

2.2 Aim and Objectives

2.2.1 Aim

Amino acids have unique characteristics and certain amino acids play an important role in the function of CYP2D6. With the paucity of pharmacogenomics studies in Africa, this study aims to gain insights into how missense variants found in sub-Saharan African populations may possibly impact the functionality of the CYP2D6 enzyme which may influence drug response, with focus to a drug commonly used in Africa.

2.2.2 Hypothesis

Given that amino acids have unique chemical properties and some amino acids have critical roles in the function of proteins, we therefore, hypothesise that *CYP2D6* missense variants influence the structural properties and dynamics of the CYP2D6 enzyme.

2.2.3 Objectives

- To catalogue African CYP2D6 variants using PharmVAR, GnomAD and the H3Africa/GSK ADME project dataset
- To discriminate variants of high impact of CYP2D6 from a large list of candidates using molecular modelling-based filtering approaches
- To select a drug/CYP2D6 complex from a set of available co-crystals in the Protein Data Bank based primarily on the quality of the structure and the relevance of the drug to African populations
- To provide insight into the molecular mechanism of variant effect on the protein function using multi-scale simulation methods which allows us to predict the possible implications of some uncharacterised star alleles relevant and identify potential causal variants for star alleles associated with a decreased enzyme activity in sub-Saharan African populations

Chapter 3: Methodology

3.1 Data Collection

The Pharmacogene Variation (PharmVar) serves as a hub for pharmacogene naming convention and its efforts are synchronised with other databases which include Pharmacogenomic KnowledgeBase (PharmGKB) and the Clinical Pharmacogenetic Implementation Consortium. In addition, PharmVar is known for incorporating a great amount of data on *CYPs* (Gaedigk *et al*., 2018). PharmVar was used for compiling a list of missense variants. No ethical clearance was required for the data acquired from public databases. Missense variants that were categorised as core variants for the *CYP2D6* star alleles were selected for further prioritisation from PharmVar using the GRCh37 reference which is the primary reference that is used by most human genetic tools (Li *et al*., 2021). Furthermore, non-zero frequency variants were retained for further analysis based on the prevalence in African populations calculated from GnomAD (Karczewski *et al*., 2020).

3.2 Variant Selection

Variant prioritisation consists of applying different filters over an initial list of variants in order to retain those with the most likely deleterious effect on CYP2D6 protein. At this stage, an efficient tool called *SWAAT* (*Structural Workflow for Annotating ADME gene Target*) was used to obtain insight into the effect of ADME variants at the structural level (Othman *et al*., 2020). *SWAAT* screens the missense variants in a VCF file, performs their structural modelling and outputs a rich report of different properties that includes, but not limited to, the variation of the folding energy upon amino acid changing from the wildtype to the variants (ΔΔG) as well as the entropy value (ΔΔS) through the integration of FoldX and ENCoM within the workflow. The integration of FoldX and ENCoM makes SWAAT a robust prediction tool, thus a preferable tool for identifying variants with potential destabilising effects. The tool also integrates a random forest model to predict the impact of variants by using a set of 15 structure-related and sequence-related features. In addition, there is an auxiliary workflow that helps in the preparation of the dataset required for SWAAT analysis. A VCF was generated to conduct *SWAAT* analysis for determining the missense variants that have a potential impact on the activity of the enzyme based on the ΔΔG energy, machine learning (ML) prediction and red flag features. These features indicate the implications of variants on the stability of a protein. A missense variant is retained if it fulfills the following

criteria: a ΔΔG > 1.0 kcal/mol, indicating a likely destabilising effect, a classification as a variant with significant impact according to the machine learning predictive model, and at least one red flag. Red flags consist of eleven molecular events originally defined by Missense3D that were found to be disease-associated features. Thereafter, the retention of missense variants in uncharacterised star alleles in PharmVar that were within the H3Africa ADME/GSK data set was performed to assess missense variants in uncharacterised star alleles and that are more common in the context of Africa. The H3Africa ADME/GSK data set includes 458 sequenced samples from various African countries (da Rocha *et al*., 2021). Ethical clearance was obtained from the Human Research Ethics Committee (Medical) of the University of the Witwatersrand (M200711) for accessing data generated by the H3Africa projects, the clearance certificate is shown in Appendix A.

Given that the prioritisation criterion was stringent, some variants were selected as exceptions. These included missense variants from star alleles commonly found in sub-Saharan Africa and defined by missense variants according to Zhou *et al*., (2017) were selected. A selection of highly prevalent (frequency > 0.1) star alleles found in sub-Saharan African populations associated with a decreased enzyme activity and characterised by missense variants was performed (Wright *et al*., 2010). These are the *29 and *17 which are highly prevalent in African populations (Masimirembwa *et* al., 1996, Wennerholm *et al*., 2001). The *29 and *17 have a frequency of 0.12 and 0.19 in sub-Saharan African populations, respectively (Whirl-Carillo *et al*., 2012). From the selected star alleles, missense variants that were solely harboured, as a core variant, in star alleles with a normal function were excluded in order to only assess variants that have a potential damaging effect according to the clinical data from PharmVar. This approach was conducted to identify potential causal missense variants from the selected star alleles. Furthermore, missense variants from star alleles that are uncharacterised and were firstly reported in African populations from recent studies (*73, *74 and *84) were selected (Wright *et al*., 2010, Gaedigk *et al*., 2017).

3.3 Selection of Reference Variants

Two CYP2D6 variants were selected as references for the computational analysis. These variants have been characterised by *in vitro* studies. P34S was selected, given that the variant results in the loss of function in CYP2D6, and has been reported to significantly decrease the enzyme's *in vivo* activities (Kim *et al*., 2013, Sakuyama *et al*., 2008). Furthermore, P34S

occurs in the highly conserved proline rich region of microsomal P450s, which may be responsible for acting as a hinge between the heme binding region of the CYP and the hydrophobic membrane that anchors the enzyme (Sakuyama *et al*., 2008). S486T was selected as a negative control as this variant does not affect the functional or structural basis of the mutant enzyme when catalysing bufuralol according to Kim *et al*., 2013. In addition, the S468T is the only core variant for *39 which is known to have a normal function and the variant results in the introduction of an amino acid that has similar properties to that of the wildtype amino acid. Thus, the amino acid change has not resulted in drastic implications (Gaedigk *et al*., 2018, Vnučec *et al*., 2016).

3.4 PDB Structure Preparation

The most complete human structure of all available CYP2D6 structures from the PDB was selected to generate the reference structure from the canonical CYP2D6 protein sequence (P10635) (Berman *et al*., 2000, The UniProt consortium, 2019). The PDB structure included two thioridazine drug molecules. Prior multiscale modelling, the PDB structure was prepared by removing molecules that were not required for the simulation. These included, the phosphate group, zinc and glycerol molecules since these were used for the purification and crystallisation of the enzyme. Furthermore, a single chain was retained as one biological entity. The heme group was retained since it is essential for the biological activity of the enzyme. The thioridazine drug molecules were bound on the active site and the entrance channel antechamber of the enzyme, respectively. Both drug molecules were retained to explore the potential impact of missense variants on the interactions of the active site and the antechamber entrance channel which play critical roles in the function of the enzyme (Dong *et al*., 2019). Following the local alignment between the canonical sequence and PDB structure using *EMBOSS Water* which uses the Smith-Waterman algorithm, an algorithm that is based on dynamic programming and provides optimal alignments (Madeira *et al*., 2019, Blazewicz *et al*., 2011). Modeller 9.24 (Blundell and Sali, 1993), which is an accurate and a commonly used modelling tool, was used to generate CYP2D6 structure conformed to the canonical Uniprot sequence. Two residues were replaced, namely, Lys32 into Arg32 and Leu33 into Tyr33.

3.5 Molecular Dynamics

It is important to gain insight into the dynamical and molecular effects of mutations to understand the mechanisms to which variants may impact the function of proteins. Molecular dynamics (MD) was the approach that was used to assess the dynamical impact of missense variants on the enzyme. This process includes the generation of topology files and coordinate files, energy minimisation, heating, equilibration and the production simulation steps. MD was performed using AMBER which provides programmes that setup, conduct and analyse MD simulations. The AMBER tool consists of classical molecular mechanics force fields that are primarily used for the simulation of biomolecules. In addition, the tool also has parameter sets that describe the most popular components of condensed matter and biomolecular simulations, which consists of parameters for naturally occurring solvents, amino acids, ions, lipids and carbohydrates (Salomon-Ferrer *et al*., 2012).

3.5.1 System Preparation for the MD Simulation

In preparation of the system, topology files and coordinate files of the wildtype and mutant CYP2D6 enzymes bound to the heme group and the drug molecule were required as input files for the further minimisation step. Mutagenesis was performed on the PDB crystal structure prior to generating the mutant topology and coordinate files. Mutagenesis involved the removal of atoms that were uncommon atoms, between the wildtype and mutant amino acids, on the PDB structure. Thereafter, the three-letter amino acid code of the wildtype was replaced with that of the mutant, on the PDB structure, and the mutagenesis process was completed with the topology generation which is described later. To further prepare the PDB structures, protonation states of histidine, aspartic acid, glutamic acid and lysine side chains were assigned at pH 7.0 (physiological pH of hepatocytes where the CYP2D6 enzyme functions) according to the pKa values determined through PROPKA (Dolinsky *et al*., 2007). Furthermore, the models were run through *pdb4amber*, for final preparation, to make the PDB file compatible with Amber (Case *et al*., 2018). The penta-coordinate ferric heme state was selected for the simulation process, since this state occurs before the binding and catalysis of the substrate through the heme group. The state was selected to mimic the biochemical state that occurs in a biological system, given that, the drug molecules are not directly interacting with the heme group in the selected PDB structure (Jandova *et al*., 2019). The force field parameters for the heme state were obtained from Shahrokh *et al*., (2012). For

the thioridazine molecules, parameter files were generated through the antechamber program. This was followed by generating topology files and coordinate files for the wildtype and mutant enzymes using *tleap* which included the generation of a bond between the proximal cysteine residue (residue 412) and the heme group (Case *et al*., 2018). The Amber 18 ff14SB force field, which is for proteins, was used for the topology generation. For the heme group and drug molecules, the General Amber Force Field (GAFF) was assigned and AM1-BCC was used to calculate the atomic point charges. The topology generation included the introduction of water molecules to solvate the system using TIP3PBOX 15.0. No counterions were required to neutralise the system.

3.5.2 Molecular Dynamics Minimisation

Prior to performing an *in vacuo* MD simulation, 500 step combined steepest descent and conjugate gradient *in vacuo* energy minimisation was performed on the CYP2D6 enzyme topology using the AMBER 18 prior to the solvation of the system (Case *et al*., 2018). This was conducted to minimise the energy of the system and remove any steric clashes between the atoms in the system which may potentially result in an unstable simulation in the later stages. Two stages of a 20 000 step combined steepest descent and conjugate gradient for *explicit* energy minimisation were then performed on the solvated system. The first stage involved 2 000 steps of the steepest descent and 18 000 steps of the conjugate gradient. Then the second stage included 4 000 steps of the steepest descent and then 16 000 steps of the conjugate gradient. This process allows for the minimisation of energy and incorporates the minimisation of both the solvent (water molecules) and solutes. The first stage of minimisation was conducted with restraints on the enzyme and ligands with a force constant of 500 kcal mol^{-1} $Å^{-2}$. The second stage of energy minimisation was performed on the entire system with no restraints. Xmgrace was used to visualise the minimisation energy plots and VMD 1.9.3 was used to visualise the generated topology following the minimisation (Humphry *et al*., 1996). Although energy minimisation was performed for all the enzymes, only the wildtype explicit energy minimisation plots are shown in Appendix B as the same principle applies to all the variants as well.

3.5.3 Heating and Equilibration

Following the energy minimisation process, the system was heated from an initial temperature of 50.0 K to a final temperature of 300.0 K using the Langevin dynamics to control the temperature with a collision frequency of 5.0 ps^{-1}. The heating was performed on the enzyme and ligand with a weak restraint of 10 kcal mol^{-1} $Å^{-2}$. The heating stage was performed for 20 ps with the SHAKE algorithm used for restraining bond lengths that involve hydrogen atoms and the random seed value (ig), which generates the velocity distribution, was kept constant at a value of 96465. This was followed by equilibration which was employed to relax the system at 300.0 K. The equilibration was performed at constant pressure with an average pressure of 1 atm which was maintained by isotropic position scaling with a relaxation time of 2 ps. This was conducted for a range of restraints with an initial force constant of 10 kcal mol^{-1} $Å^{-2}$ to a final force constant of 0 kcal mol^{-1} $Å^{-2}$ lowered with a decrement of 1 kcal mol^{-1} $Å^{-2}$ for each 80 ps which was then followed by the production simulation. The wildtype heating and equilibration plots are shown in Appendix C.

3.5.4 Molecular Dynamics Production Simulation

A molecular dynamics production simulation was employed to assess the potential impacts of missense variants at the molecular level. This method generates trajectories for the wildtype and mutant enzymes which can be analysed to understand how variants may affect the dynamics of the enzyme. Production simulation was performed for 500 ns on the wildtype and mutant enzymes using AMBER 18. Moreover, the simulation was conducted at a constant temperature of 300 K using the weak-coupling which ensures that the kinetic energy is appropriate for the selected temperature and the trajectory snapshots were obtained every 10 ps. Given that the time limit for running simulations using the CHPC GPU is 12 hours, the simulation for each enzyme (wildtype and mutants) was divided into eight stages. Another advantage is that this approach is cost effective, given that; the simulation continues from the previous stage instead of starting over completely; whenever technical disruptions occur during the run.

3.6 Data Analysis

AMBER includes several trajectory analysis tools such as *CPPTRAJ* and *PTRAJ*. *CPPTRAJ* is a tool that has several advantages over *PTRAJ*, such as, generating files that can be supported by Gnuplot and Xmgrace, which are visualisation tools, and can strip topology files to reduce disk space usage and accelerate the analysis process (Salomon-Fererer *et al*., 2012, Case *et al*., 2018). Prior to the analysis of the molecular dynamic trajectories, water molecules were stripped from the final topology files using *CPPTRAJ* (Roe and Cheatham, 2013). Data analysis involved several assessments including the Root Mean Square Deviation (RMSD), Root Mean Square Fluctuation (RMSF), Principal Component Analysis (PCA), Porcupine Plot analysis and Secondary Structure Analysis (SSA). RMSD, RMSF and PCA were performed on the Jupyter environment to generate and visualise the plots. The RMSD and RMSF trajectory was read using *MDTRAJ*, which is an open and lightweight Python library that is employed to manipulate and analyse MD trajectories (McGibbon *et al*., 2015). *PYTRAJ*, which is another python library that is utilised for assessing MD trajectories, was used for the PCA calculation (Nguyen *et al*., 2018). *CPPTRAJ* was used to generate the SSA plots, which are saved as gnuplot files, and Porcupine plots that are visualised using VMD. Visualisation of the SSA and the Porcupine plots were conducted using the Gnuplot 5.2.7-qt and VMD 1.9.3, respectively.

3.6.1 Root Mean Square Deviation

The RMSD measures the average displacement of atoms relative to a reference structure during a simulation (Knapp *et al*., 2011). This approach assesses the time-dependent movements of a structure and evaluates the evolution of a trajectory that shows the stability of the structure (Martinez, 2015, Knapp *et al*., 2011). We employed this method to assess the potential impact of the selected variants on the stability of the CYP2D6 enzyme. This provides insight on how the missense variants may influence the stability of the enzyme.

3.6.2 Root Mean Square Fluctuation

The RMSF measures the displacement of particular atoms with respect to the reference structure with an average over the number of atoms (Knapp *et al*., 2011). This method was used to evaluate the local fluctuations of the residues on the CYP2D6 enzyme as a result of

the missense variants. This reveals the potential effects of the mutations on the flexibility of the enzyme which has potential implications on the catalytic efficiency of the enzyme.

3.6.3 Principal Component Analysis

Principal Component Analysis (PCA) is a multivariate statistical method employed to systematically decrease the number of dimensions required to describe protein dynamics by means of a decomposition process which filters observed motions from the greatest to the least spatial scales (David and Jacobs, 2014). The implementation of PCA on a protein trajectory is referred as "Essential Dynamics", since essential motions are extracted from sampled conformations (David and Jacobs, 2014). As a result, this method yields dominant modes in the movements of the atoms from the MD trajectory (Haider *et al*., 2008). These significant movements correspond to the collective movements of groups of molecules in normal modes analysis (Teeter *et al*., 1990). Furthermore, this method generates eigenvectors and eigenvalues. The eigenvector represents the motion whereas the eigenvalue corresponds to the eigenvector and represents the energetic contribution of the particular component on the motion (Haider *et al*., 2008). The PCA method has several advantages, one advantage is that information from any atom may be used to acquire the PCA modes of that particular subspace (David and Jacobs, 2014). Another advantage is that statistics from numerous trajectories can be pooled which allows flexibility for the manner in which data from a number of trajectories may be combined (David and Jacobs, 2014). Although, one limitation of essential dynamics is that whenever variables are not intrinsically linearly related, nonlinear relationships that are present are not described properly (David and Jacobs, 2014). In this study, the PCA approach was employed to evaluate the implications of missense variants on the motion of the enzyme which may influence the efficiency of the enzyme.

3.6.4 Porcupine Plots

Assessing the stability and flexibility of a protein does not provide extensive information pertaining the nature of movements from the CYP2D6 enzyme residues which have implications on enzyme activity. To assess this effect resulting from missense variants, Porcupine plots were generated from the eigenvectors that were produced using PCA. Porcupine plots show a graphical representation of the summarised enzyme movements (Wang *et al*., 2011). This method reveals the movements of the residues on the enzyme by producing spikes, from the residues with significant motion, that represent the direction and

the magnitude of the enzyme residues (Chen *et al*., 2016). The advantage of this method is that it shows which residues on the protein exhibited significant movements, including their magnitude and direction (Wang *et al*., 2011). Thus, this approach was used to assess the implications of the missense variants on the movements of the residues on the enzyme. This provides better understanding on how the enzyme may potentially interact with its substrate as a result of altered movements.

3.6.5 Secondary Structure Analysis

Understanding the impact of mutations on the secondary structure provides insight into how mutations may potentially affect the functionality of the protein, given that the protein's structure is critical for its function (Ma *et al*., 2018). The SSA was implemented to assess the impact of the missense variants on the secondary structure of the enzyme. The dictionary of secondary structure prediction (DSSP) algorithm was implemented, through *CPPTRAJ*, for generating secondary structure analysis plots. The DSSP algorithm is based on the use of hydrogen bonding patterns on the backbone of the protein for classification and is robust in distinguishing α-helices from β-helices (Nagy and Oostenbrink, 2013). In addition, the DSSP algorithm provides insight in bends and turns on the structure of the protein (Case *et al*., 2018). The scripts used for the study can be accessed on github on the following link https://github.com/BR-Sitabule/CYP2D6_SB.

Chapter 4: Results

4.1 Overview

We sought to evaluate the potential impact of the selected missense variants on the enzyme's which may potentially influence the enzyme activity. Missense variants were prioritised using the *SWAAT* tool and additional variants were selected as exceptions to the criteria that were implemented. These were analysed using MD which included the implementation of several assessments to assess the dynamic effects of the prioritised missense variants on the CYP2D6 enzyme. This provided insight on the potential implications of the variants on the functionality of the enzyme. These MD assessments include the RMSD, RMSF, PCA, Porcupine plot and Secondary Structure analyses and were performed on the backbone atoms (Cα, C, N) of the enzyme. The outcomes for all the steps which were carried out are discussed in greater detail in the subsequent sections.

4.2 PDB Structure Selection and Cleaning

Structures of ligand (drug) bound to CYP2D6 available on the Protein Data Bank (PDB) were evaluated to identify the most suitable ligand with a structure (Berman *et al*., 2000). The ligands that contained structures included, prinomastat (anti-cancer drug), thioridazine (anti-psychotic drug), ajmalicine (treat high blood pressure), quinidine (treat psychomotor-seizures), quinine (anti-malarial drug) and BACE1 inhibitors (treat Alzheimer's disease). DrugBank was used to determine the details regarding the medicinal use of the drugs (Wishart *et al*., 2006). Thioridazine was selected as the drug for this study as it is known to be used as an anti-psychotic drug worldwide. In addition, thioridazine has been found to be a potential anti-cancer, anti-inflammatory and anti-microbial drug (Baig *et al*., 2018; Thanacoody, 2007; Chang *et al*., 2018). Furthermore, drugs such as prinomastat, quinidine and quinine, which are inhibitors of CYP2D6, were excluded given that they are not substrates of CYP2D6 (Wang *et al*., 2015). The PDB ID of the selected structure is 3TBG. The PDB structure is a human CYP2D6 enzyme that is bound to two thioridazine molecules and consists of four subunits, namely, chain A/B/C/D (Wang *et al*., 2015).

The PDB structure was prepared for MD by removing atoms that were not required for the modelling step, these included, the phosphate group, zinc, water and glycerol molecules since these were used for the purification and crystallisation of the enzyme. A single chain (chain

A) was retained as one biological entity was adequate for the evaluation of the potential impact of missense variants on the enzyme activity. The heme group was retained since it is essential for the biological activity of the enzyme. The two thioridazine molecules were bound on the active site and the entrance channel antechamber of the enzyme, respectively. Both thioridazine molecules were retained to explore the potential impact of missense variants on the interactions of the active site and the antechamber entrance channel, which play critical roles in the function of the enzyme, with the drugs (Dong *et al*., 2019).

4.3 Variant Selection

A total of 50 missense variants, using the GRCh37 build, from African populations were selected from PharmVar. The details for these 50 variants are shown in Appendix D. A total of eight variants were retained using the *SWAAT* tool and these are discussed in more detail in section 3.4. This was followed by the retention of missense within the H3Africa dataset resulting in a final of two variants, namely, Y355C and R365H. Due to the stringent criteria, exceptions to the criteria were selected. This included variants harboured in *17 and *29 which are highly prevalent *alleles in sub-Saharan Africa as well as *73, *74 and *84 which have only been reported in Africa to date. These variants are, namely, T107I (*17), V338M (*29), V136I (*29), L91M (*74), V104M (*74) and P267H (*84). A summary of variant selection is shown in Figure 4.1. A final of eight variants and two reference variants, P34S and S486T, were selected for MD assessment. Table 4.1 shows the details of the missense variants selected for the MD analysis which includes the reference variants and Figure 4.2 shows the PDB structure of CYP2D6 annotated with the eight selected variants as well as the two reference variants. The PDB structure was visualised using PyMOL 2.4.1 (Schrödinger, 2010).

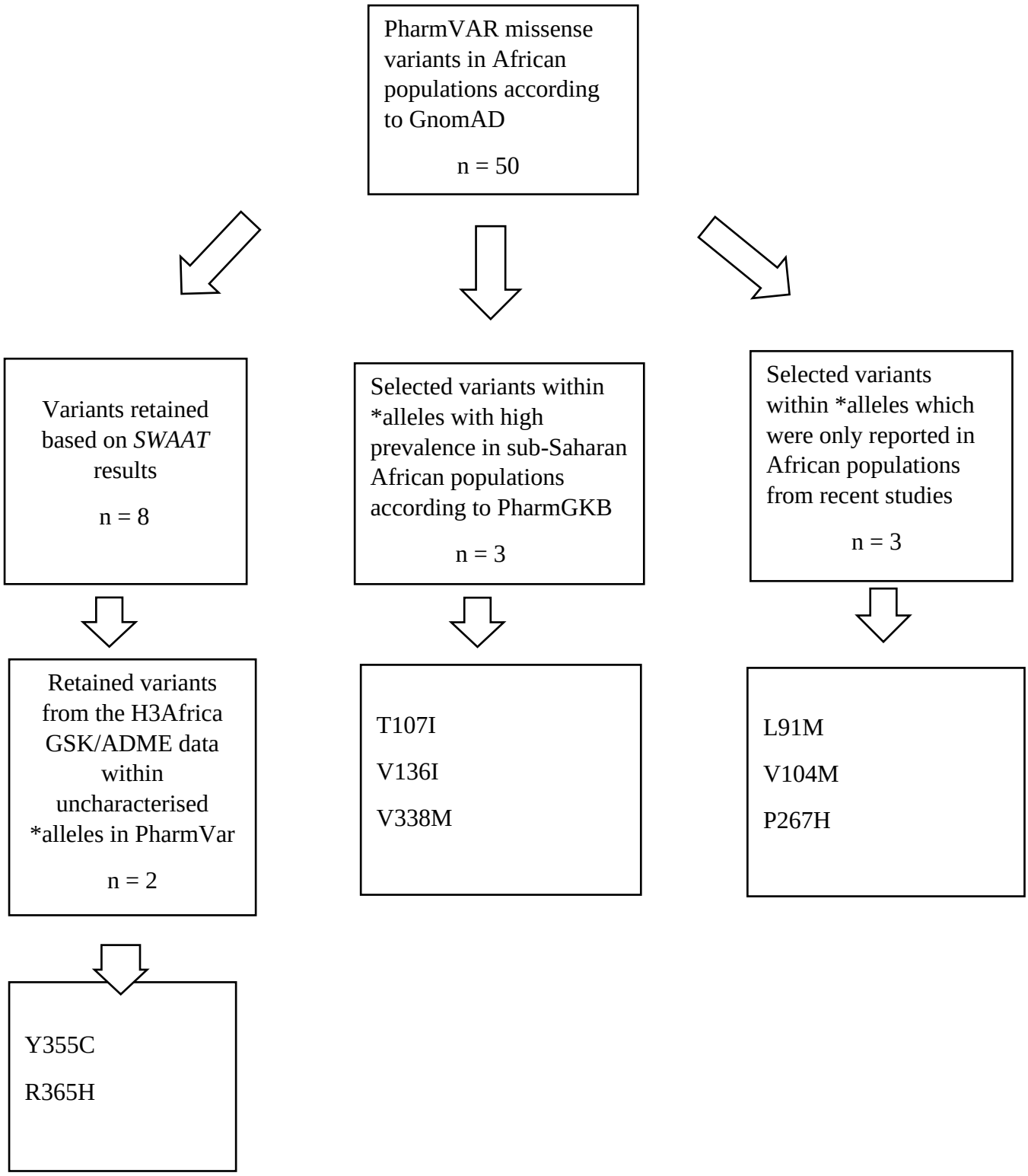


Figure 4.1: Schematic diagram showing the selection criteria for variants selected for the MD assessment. The "n" denotes the number of variants retained in each step.

Table 4.1: Table showing the details of the variants selected for the MD simulations and their frequencies in African populations according to GnomAD.

Variant ID	Nucleotide change (GRCh37)	Amino acid change	Frequency in African populations according to GnomAD
rs1065852	g.42526694G>A	P34S	0.12
rs28371703	g.42525821G>T	L91M	0.03
rs267608308	g.42525782C>T	V104M	-
rs28371706	g.42525772G>A	T107I	0.19
rs61736512	g.42525134C>T	V136I	0.10
rs148769737	g.42524219G>T	P267H	<0.01
rs59421388	g.42523610C>T	V338M	0.09
rs202102799	g.42523558T>C	Y355C	<0.01
rs1058172	g.42523528C>T	R365H	0.03
rs1135840	g.42522613G>G (C>G)*	S486T	0.35

*GRCh37 denotes the variant as the reference, thus the mutation (at genomic level) is shown using GRCh38 in brackets.

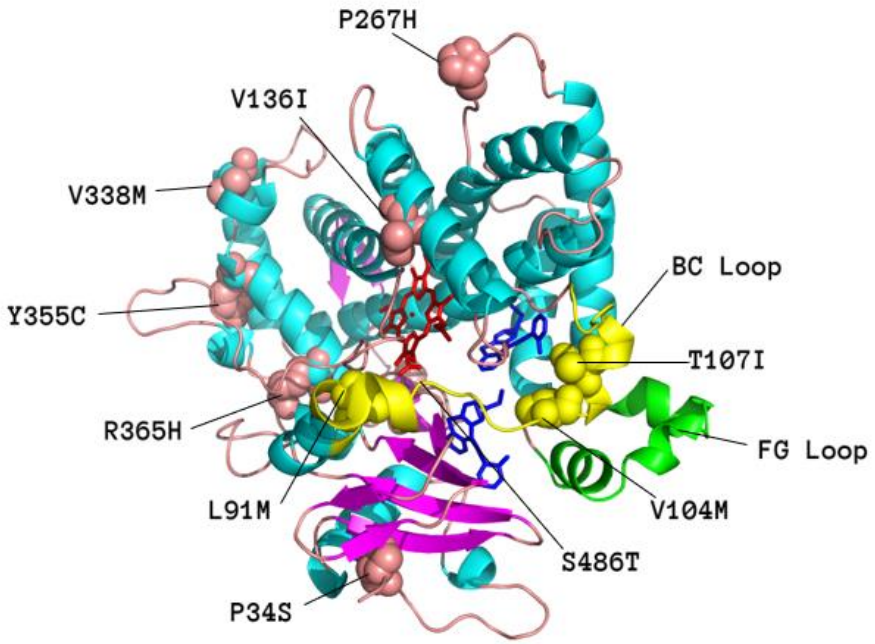


Figure 4.2: Diagram showing the CYP2D6 enzyme structure and the missense variants to be assessed using MD, with P34S and S486T serving as controls. The heme group is shown in red, the two thioridazine molecules are shown in blue and the variants of interests are depicted as light pink spheres. The helices are shown in cyan, the sheets are depicted as magenta and the loops are portrayed in light pink. The FG loop is shown in green and the BC loop is shown in yellow on the Figure (Figure generated using PyMOL 2.4.1).

4.4 SWAAT Analysis

SWAAT analysis was performed on 50 missense variants to prioritise variants that have potential significant implications. The *SWAAT* results for all the assessed variants are shown in Appendix E. From the 50 missense variants, V7M, V11M, R26H and R28C were not processed, given that the *SWAAT* tool algorithm incorporates the available protein structures on PDB and the CYP2D6 enzyme crystal structures lack the first 31 residues. Figure 4.3 shows a histogram of the frequencies and ΔΔG distribution for the 46 variants. A total of 15 variants had a ΔΔG > 1.0 kcal/mol which indicates that they are potentially destabilising. From these variants, eight missense variants were retained by using two other features as described in the methods section and Table 4.2 showing the eight missense variants and their *SWAAT* results.

As shown in Table 4.2, the R365H mutation had the highest ΔΔG score, which was 8.525 kcal/mol, suggesting that the mutation may have the most destabilising effect compared to the other seven variants. All eight variants had an ML prediction score of 2 which indicates that they have potential significant implications on the enzyme activity. From the eight variants, only M279K and P469A were not flagged as variants that occur in hotspot patch regions. The variants that are flagged as "hotspot patch" are potentially deleterious as they occur in a hotspot patch region. The P34S, L231P and P469A mutations were flagged as "buried exposed switch" which indicates that these variants may be damaging as a result of destabilising the structure through an amino acid change that involves proline that possesses a bulky side chain. The L31P mutation was also flagged as a "buried proline introduced" which also implies that the variant may destabilise the protein through the introduction of the proline amino acid. The M269K was flagged as "salt bridge formation" which results from the charged lysine amino acid. This mutation affects the nature which the residue may interact with the solvent as a result of the introduced charge. The P34S, G42R and G42E were the only variants that have been characterised with *in vitro, in vivo* or biochemical assays according to Kim *et al.,* (2013) and Tsuzuki *et al*., (2003).

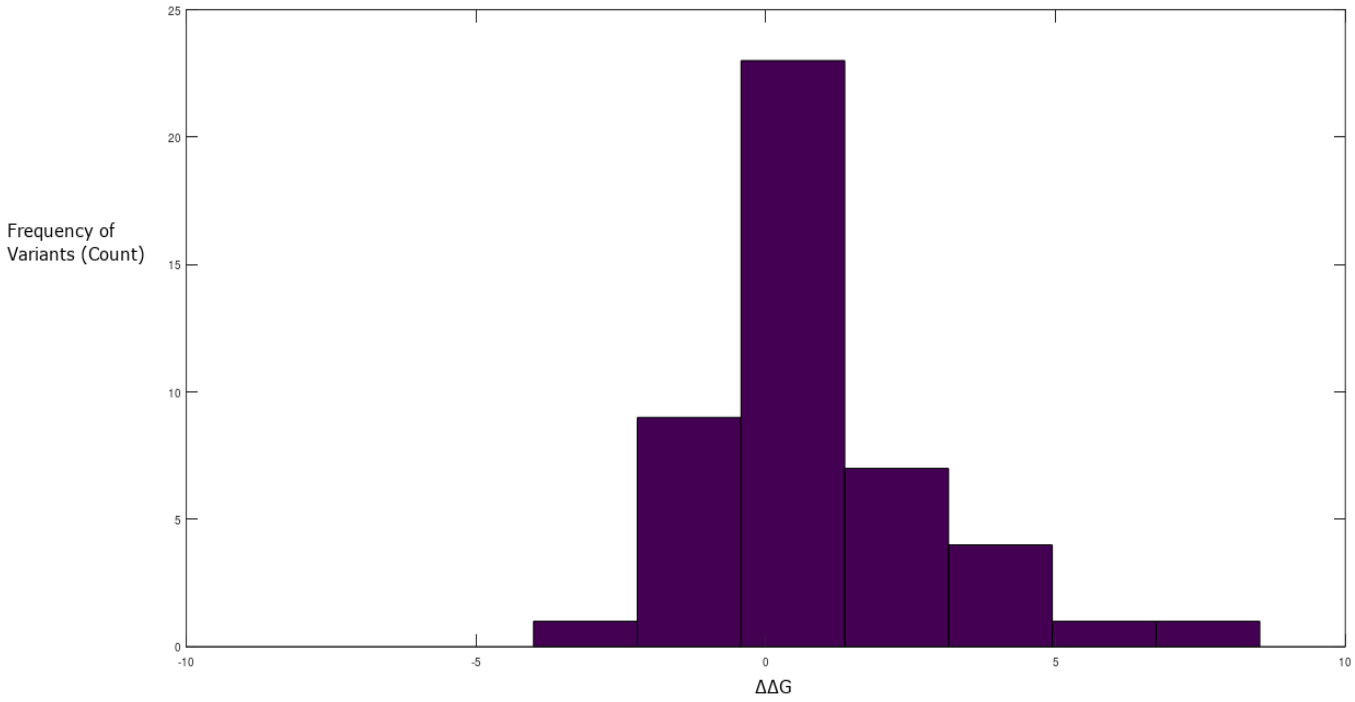


Figure 4.3: Diagram showing the frequencies and distribution of the ΔΔG for the 46 African missense variants selected from PharmVar.

Table 4.2: Variants that have a potential significant effect according to the *SWAAT* features used for prioritisation and their consequences according to *in vitro, in vivo* or biochemical studies.

Variant	ΔΔG value (kcal/mol)	Machine Learning Prediction score	Red flag description(s)	Reported impact on enzyme functionality according to literature
Y355C	2.290	2	hotspot patch	Unknown implications
P34S	4.575	2	buried exposed switch; hotspot patch	Decreased function (Kim *et al*., 2013)
R365H	8.525	2	hotspot patch	Unknown implications
G42R	2.243	2	hotspot patch	Decreased function (Tsuzuki *et al*., 2003)
G42E	2.768	2	hotspot patch	Decreased function (Tsuzuki *et al*., 2003)
L231P	6.389	2	buried proline introduced; buried exposed	Unknown implications

			switch; hotspot patch	
M279K	1.878	2	Salt bridge formation	Unknown implications
P469A	2.020	2	buried exposed switch	Unknown implications

4.5 RMSD

RMSD results were visualised using *MDTRAJ* following the generation of MD trajectories through *AMBER*. RMSD is an MD assessment that was used to assess the stability of the enzymes by analysing the time evolution of the backbone deviation of a variant compared to the reference structure, corresponding to the crystal structure of CYP2D6 used to run the molecular dynamics simulation (wildtype structure) as shown in Figure 4.4.

Figure 4.4 shows that some of the variants had similar trends. The Y355C and R365H variants yielded similar trends which were also similar to that of the wildtype trajectory. The P34S, V338M, V104M and V136I mutations showed similar patterns to each other, as these showed some fluctuations on the trajectory and were less steady compared to the wildtype trajectory. The L91M, P267H and S486T yielded trajectories that were fairly stable although the L91M and P267H trajectories had some fluctuations while S486T was steadier.

The results for L91M show that the wildtype converges around 200 ns and the mutant enzyme RMSD trajectory is initially stable between the 0 and 300 ns. At this point the trajectory fluctuates and regains a steady value. Minute deviations were observed between the mutant and the wildtype RMSD trajectories which suggests that the mutation may potentially destabilise the enzyme; however, this may not be significant.

For P34S, the mutant trajectory maintains a steady value between 0 and 120 ns. However, moving to the later stages of the simulation, there are deviations that occur between the mutant and wildtype enzyme and convergence occurred later for the mutant enzyme. This suggests that the P34S may potentially destabilise the enzyme which may have implications on the functionality of the enzyme.

The results for the S486T mutant showed that the mutant enzyme had a steady trajectory through the entire simulation. The trajectory had lower RMSD values in comparison to the wildtype. This suggests that the mutant has greater stability compared to the wildtype enzyme.

For the Y355C enzyme, the trajectory of the mutant coincided with the trajectory of the wildtype through the entire simulation. This suggests that the mutation may not have implications on the stability of the enzyme.

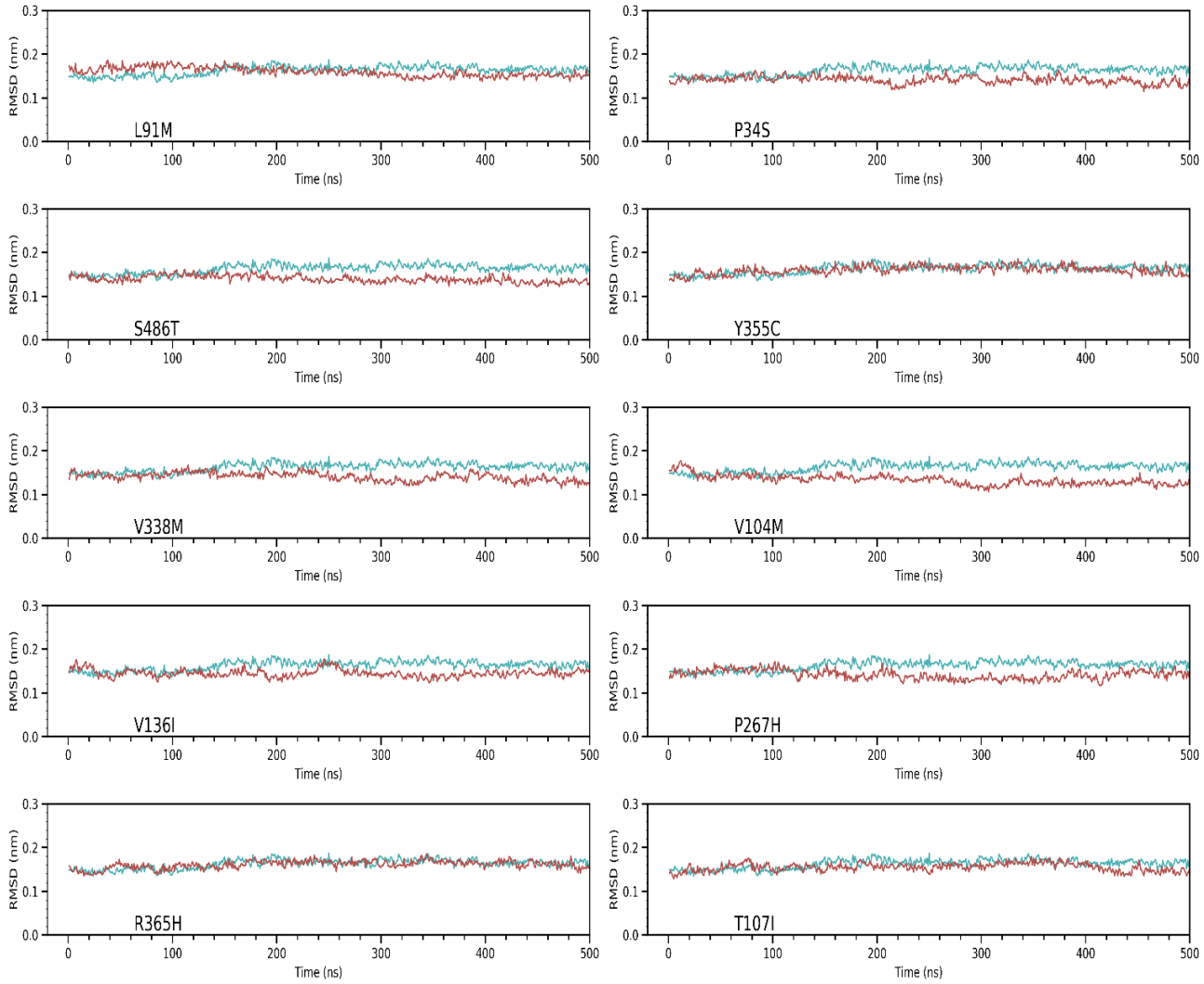


Figure 4.4: A diagram illustrating the comparison of RMSD trajectories of the mutant and wildtype enzymes in a 500 ns simulation. The wildtype is shown in cyan and the mutant is shown in red. The RMSD (nm) is shown on the y-axis and the time (ns) is shown on the x-axis.

For the V338M mutation, the mutant enzyme's trajectory was initially steady; however, some fluctuations were observed from ~140 ns and a more stable trajectory was observed later in

the simulation compared to the wildtype enzyme. This suggests that the mutation may potentially have an impact on the stability of the enzyme.

For the V104M trajectory. The trajectory of the mutant converges at ~340 ns, which is later than the wildtype enzyme, indicates that the mutation may potentially affect the stability of the enzyme and ultimately the functionality of the enzyme.

The results for the V136I mutation showed that the mutant only stabilises at 300 ns. This implies that the mutation may potentially result in altering the enzyme's stability. However, the trajectory is steady from 0-120 ns and 300-500 ns, suggesting that the variant did not impact the enzyme's stability drastically.

The P267H trajectory stabilised at around 200 ns together with the wildtype enzyme. This suggests that the mutation may potentially have no impact on the stability of the enzyme and ultimately no significant impact on the functionality of the enzyme.

For the R365H enzyme results, the trajectory of the mutant enzyme coincided with the enzyme through the 500 ns simulation. Thus, this may suggest that there is no impact on the stability of the enzyme by the mutation.

For the T107I mutant and the wildtype enzyme, the mutant trajectory coincided with the wildtype from 0-400 ns; however, it deviates from the wildtype from ~400 ns and only stabilises in ~460 ns. This may imply that the mutation may have an effect on the stability of the enzyme.

4.6 RMSF

For further assessments, RMSF analysis was applied to observe the local fluctuations that occurred on the residues of the enzyme. The local fluctuations reflect the flexibility of the enzyme. Figure 4.5 shows the RMSF of the wildtype enzyme and the mutant enzymes. Variation of the RMSF between the wildtype enzyme and the L91M mutant was observed in some residues. The wildtype has greater fluctuations in the first N-terminal residues on the RMSF plot. Furthermore, the wildtype has higher fluctuations around residue 200 and in approximately residue 260. These indicate that the wildtype has a greater flexibility on those

regions which has implications on how the enzyme interacts with the substrate, which is thioridazine in this case.

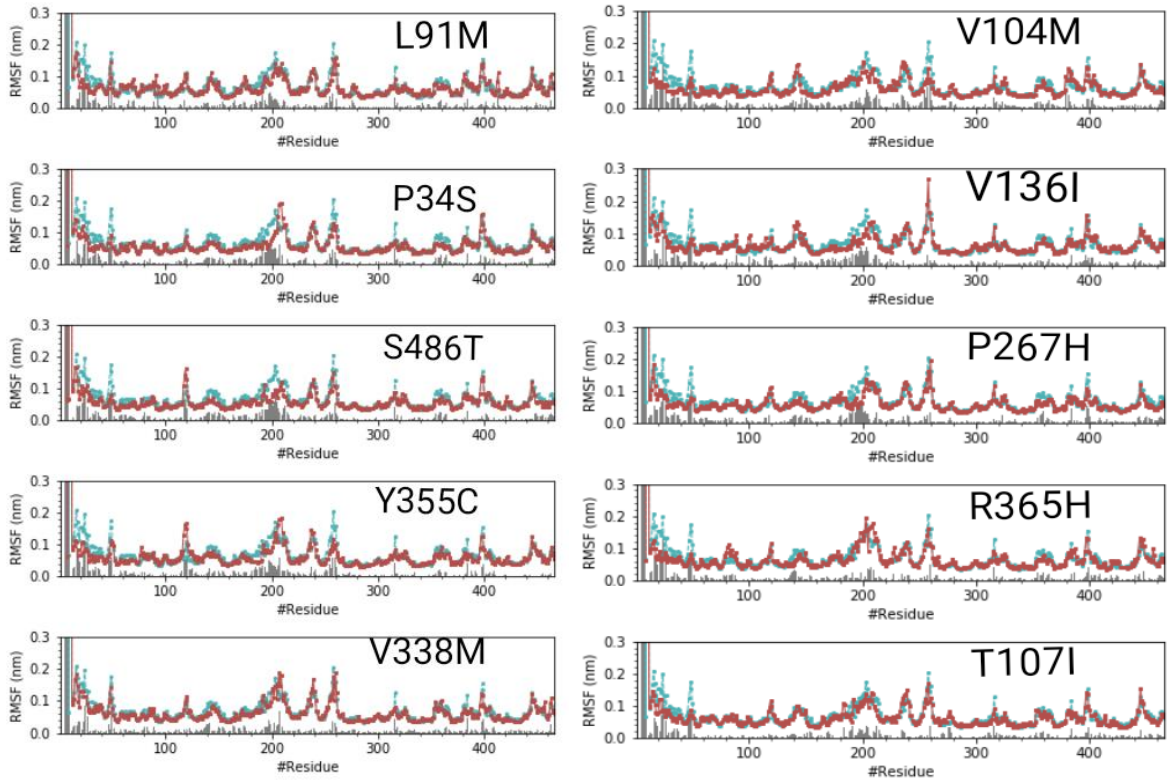


Figure 4.5: A diagram depicting the comparison of the RMSF of the mutant enzymes and wildtype enzyme in a 500 ns simulation. The wildtype is shown in cyan and the mutants are shown in red. The residue positions of the enzyme are shown on the x-axis and the Root Mean Square Fluctuation (nm) is shown on the y-axis.

The results from Figure 4.5 show that the variants yielded similar patterns for the first 100 residues and the last 100 residues. Various implications were observed in the central regions of the protein for the different variants. These implications are described in more detail for each variant in the subsequent paragraphs.

For the P34S mutation, there are higher fluctuations on the wildtype from the N-terminal residues, near residue 140, 240 and 360. However, there were lower fluctuations compared to the mutant around residue 210. These differences indicate that there are differences in the flexibility from the wildtype enzyme and the mutant enzyme.

The results for S486T show that there were higher fluctuations observed from the wildtype from the N-terminal residues, residue ~200 and at ~260. This shows that the wildtype has a higher flexibility compared to the mutant enzyme in the affected residues.

For the Y355C mutation analysis, higher fluctuations were observed in the wildtype in the first N-terminal residues and the residue around 260. However, on residue ~120 and ~210 the mutant has higher fluctuations. This shows the variation in flexibility between the two.

The results for the V338M mutation show that the wildtype did not have prominent differences in fluctuations compared to the mutant enzyme. This indicates that the mutation may not have much implications on the flexibility of the enzyme.

For the V104 mutation, the plot shows that there are differences in the N-terminal residues, residue ~200 and ~260. The wildtype had higher fluctuations in those residues. This suggests that the mutation may potentially result in a decreased flexibility.

The V136 mutation RMSF analysis shows that from the N-terminal residues and residue ~200 the wildtype enzyme had higher fluctuations compared to the mutant. However, in residue ~260, the fluctuations are higher for the mutant which shows some increase in flexibility due to the presence of the mutation.

For the P267H mutation, prominent differences in fluctuations are observed in the N-terminal residues, residue ~200 and ~400. The wildtype enzyme has higher fluctuations in all those sites, showing that the mutant has a reduced flexibility when compared to the wildtype enzyme.

The results of the R365H mutation showed that prominent differences were only observed in the first 30 residues on the plot with the wildtype enzyme having the highest fluctuations. These indicate that the wildtype is more flexible than the mutant in that region.

The results for the T107I mutation revealed that differences were only observed on several residues from the N-terminal end of the enzyme. The wildtype has more fluctuations compared to the mutant in this region, which indicates that the wildtype has a higher flexibility than the mutant in those residues.

4.7 Principal Component Analysis (PCA)

To assess the conformational variance in protein folding and motion between the wildtype enzyme and mutant enzymes, PCA (Principal Component Analysis) was performed. The CYP2D6 trajectories were projected on the subspaces and the calculations were based on the backbone atoms (Cα, C, N) of the enzyme. This approach compares the most significant variations between the wildtype enzyme and the mutant enzyme. The results pertaining the effect of the missense variants on the conformation of the enzyme are shown in Figure 4.6 (PC1 vs PC2) and Figure 4.7 (PC3 vs PC4) for the mutant enzymes compared to the wildtype enzyme.

The PC1 vs PC2 results showed that all the variants had implications on the movement of the enzyme. The S486T, L91M and V104M resulted in similar movement patterns on the enzyme. The other seven variants yielded unique movement patterns. For the PC3 vs PC4 plots, the S486T, T107I, R365H, V119L and V136I showed similar movement patterns and were similar to the wildtype movements. V104M and V136I showed some similar implications for the enzyme's movement. P267H, Y355C and L91M had unique implications on the movements of the enzyme.

The PC1 vs PC2 plot for L91M shows that there are differences in motion of the wildtype and mutant enzyme; however, the differences are not as prominent with the PC3 vs PC4 plot. These differences indicate that the mutation has potential implications on the movement of the enzyme which may potentially affect how the enzyme interacts with the drug and its functionality.

For the P34S results, differences were observed from the PC1 vs PC2 plot, however, the differences were moderate. Minor differences were also observed with the PC3 vs PC4 plot. This suggests that the mutation does not cause drastic changes in the motion of the enzyme.

The S486T PCA analysis results for the PC1 vs PC2 plot showed some differences in motion between the mutant and the wildtype enzyme. The PC3 vs PC4 shows minor differences between the mutant and wildtype enzyme. This suggests that the mutation may alter the enzyme's motion.

For the V104M PCA analysis, results showed that there were prominent differences in the motion of the wildtype and mutant enzyme. The PC3 vs PC4 had drastic differences between the mutant enzyme and the wildtype enzyme. This implies that the V104M may potentially have a prominent effect on the enzyme's motion.

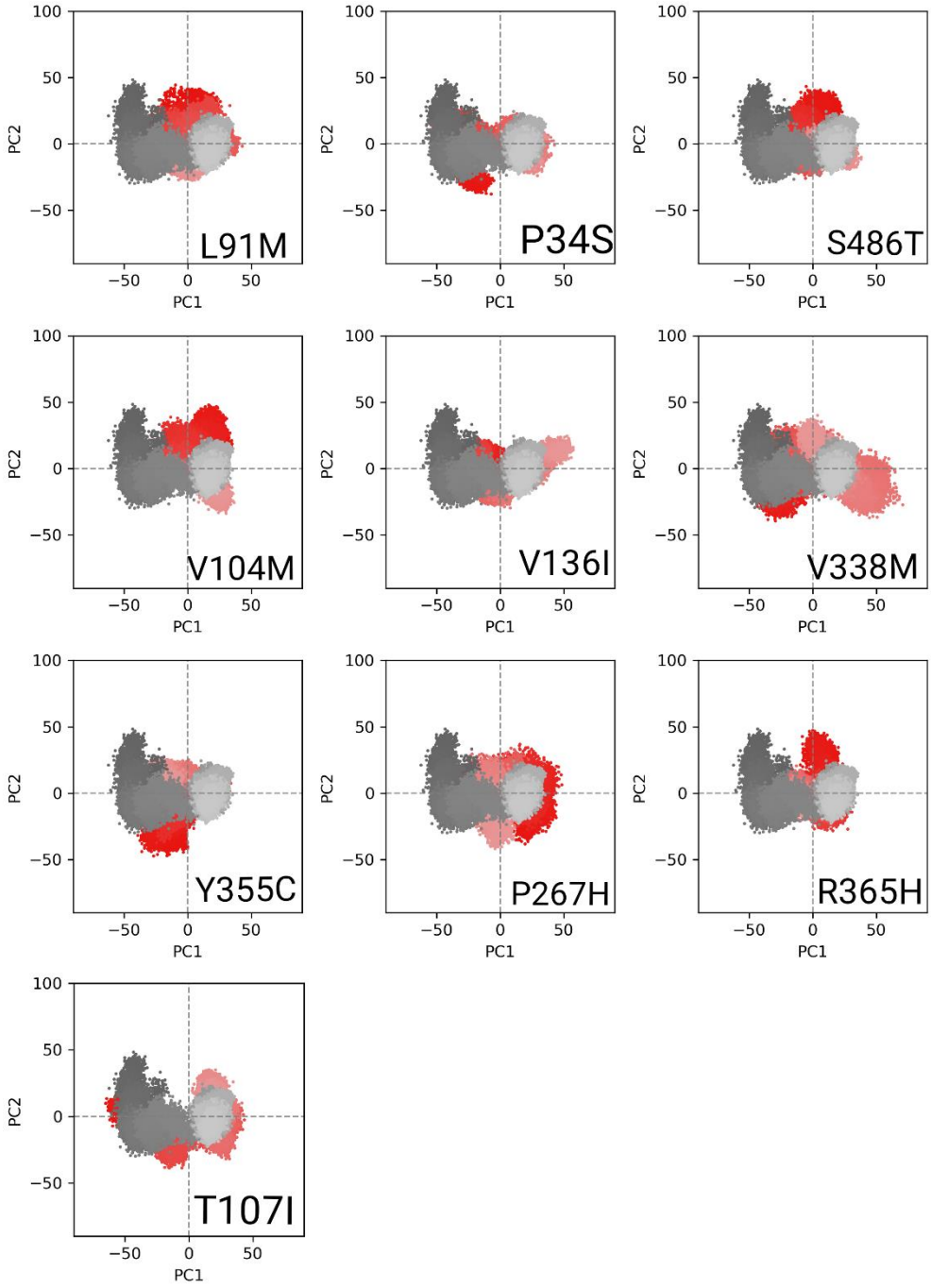


Figure 4.6: The diagram showing the PC1 vs PC2 principal component analysis of the wildtype and mutant enzyme's motion. The wildtype is shown in grey and white plots and the mutant is shown in red and salmon.

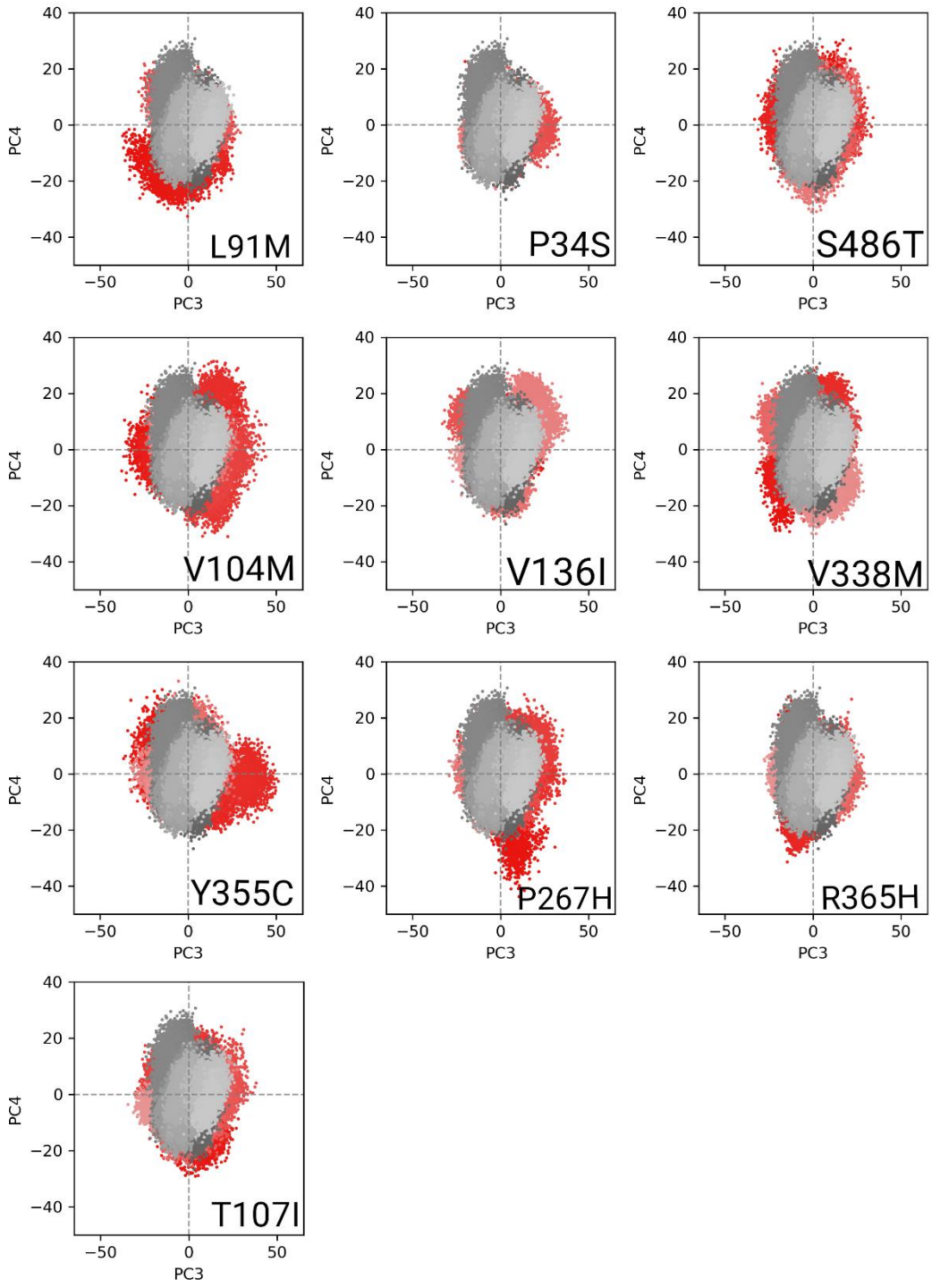


Figure 4.7: The diagram showing the PC3 vs PC4 principal component analysis of the wildtype and mutant enzyme's motion. The wildtype is shown in grey and white plots and the mutant is shown in red and salmon.

The PC1 vs PC2 plots for the V136I mutation shows a difference in motion between the two enzymes. However, the PC3 vs PC4 has minor differences. These results show that the mutation results in changes on the motion of the enzyme when compared to the wildtype enzyme.

For the V338M mutation PCA analysis, the PC1 vs PC2 outcome showed prominent differences in the motion between the wildtype and the mutant enzyme. In addition, the PC3 vs PC4 also had moderate differences in the motion of the two enzymes. This suggests that the mutation has implications on affecting the direction and magnitude of the enzyme's motion.

The Y355C PCA results showed that the mutant had great differences in motion (for the PC1 vs PC2) compared to the wildtype enzyme. In addition, the differences were prominent for the PC3 vs PC4 analysis as well. This suggests that the mutation may have a great effect on the motion of the enzyme.

The P267H mutation PCA assessment revealed that the PC1 vs PC2 plot showed great differences in the motion of the mutant compared to the wildtype. This is also observed with the PC3 vs PC4 analysis. These results suggest that the mutation may potentially have a significant effect on the enzyme's functionality by affecting the enzyme's mobile characteristics.

The R365H PCA outcome showed that the PC1 vs PC2 shows the differences in the motion of the wildtype enzyme and the mutant enzyme. The differences on the PC3 vs PC4 are not prominent. However, these results still show the potential implications on the enzyme's mobility which results from the introduction of the mutation.

The T107I PCA results showed that the PC1 vs PC2 plot shows variation in the motion between the wildtype and mutant enzyme. The PC3 vs PC4 plot showed moderate differences in the motion of the mutant enzyme and the wildtype enzyme. These findings show that the mutation has an impact on altering the motion of the enzyme.

4.8 Porcupine Plots

Porcupine plots were generated to evaluate the magnitude and directions of the movements of the residues during the simulation. These plots portray the enzyme structure with spikes that represent the movements of the residues. The length of the spike represents the magnitude of the motion and the direction of the motion is represented by the point of the spike. The mutants were compared to the wildtype which are shown on Figure 4.8. All the variants altered the movement patterns of the enzymes. This included altering the movements in the FG loop, the BC loop and/or the central regions. However, the new movements were unique among the variants in terms of the magnitude and the direction of the movements for the residues.

For the L91M Porcupine plot analysis, great movements were observed in both enzymes on the FG loop (which is a flexible loop). However, the direction of the motions was different. Furthermore, some of the residues located in the BC loop had greater motions. As a result, these alterations may potentially result in inadequate interactions between the enzyme and the substrate.

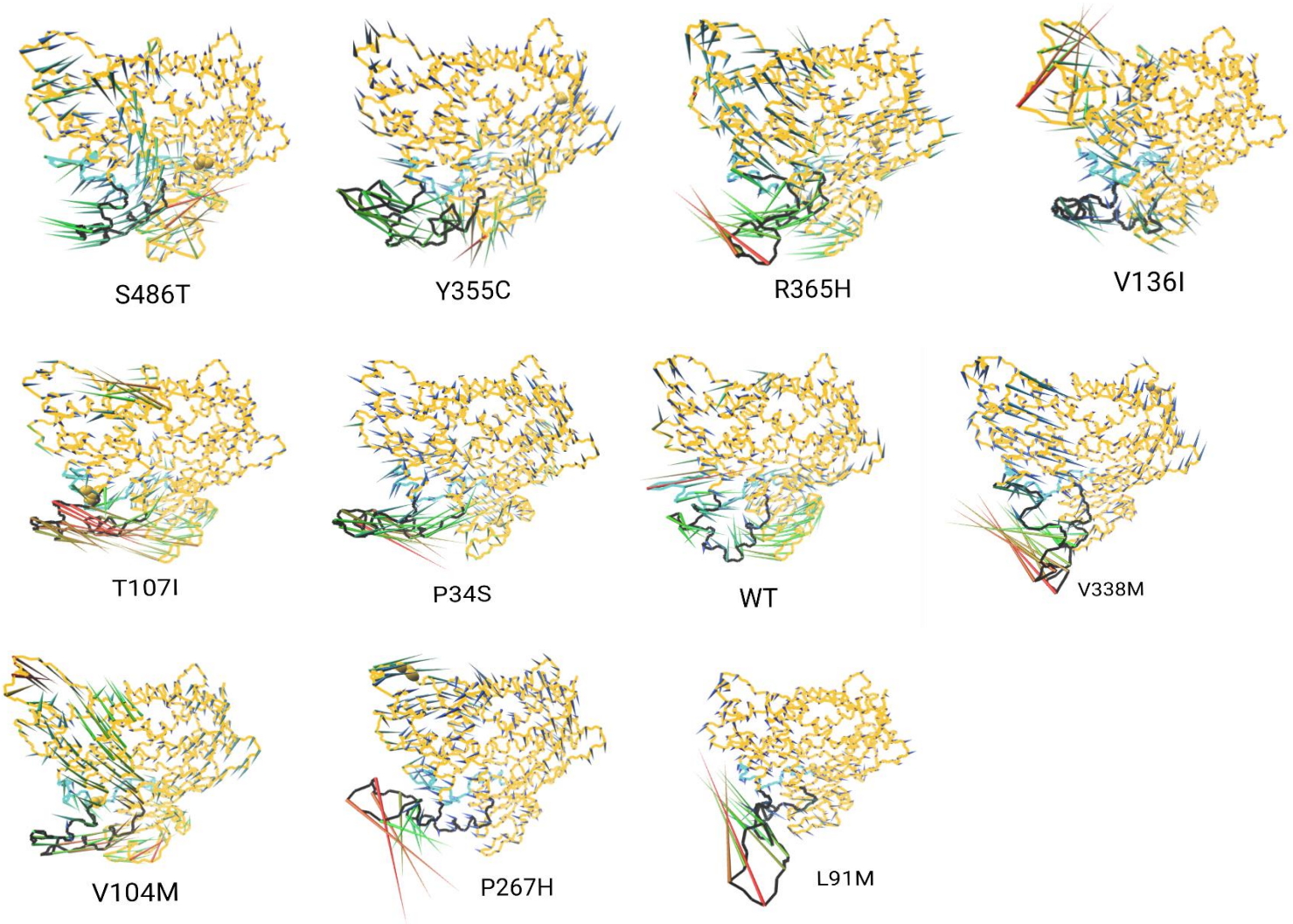


Figure 4.8: A diagram showing the Porcupine plots of the wildtype enzyme and mutant enzymes. The variants are illustrated as spheres on the enzyme structures and the spikes depict the magnitude and direction of the motions from the residues. The FG loop and BC loop are shown in black and cyan, respectively.

For the Y355C mutation, the motion from the residues on the FG loop were similar for the mutant and wildtype enzyme. The BC loop residues also showed similar patterns. This suggests that this mutation may not have an impact on the enzyme's motions significantly.

For the V136I Porcupine plot results, both the wildtype and mutant enzyme had similar motions on the BC loop and FG loop; however, the mutant has greater movements with unique directions on the central site of the enzyme. Furthermore, larger movements are also observed on the neighbouring region. This suggests that the mutation may result in influencing the functionality of the enzyme.

The V338M Porcupine plot showed that the direction of the movements of the FG loop residues differs between the two enzymes. Furthermore, the mutant residues located in the central region have greater motion and move in different directions compared to the wildtype enzyme. Thus, this mutation may potentially result in inefficient interactions between the enzyme and the drug.

The P34S mutation Porcupine plot results show that the direction of the residues located in the FG loop differ between the two enzymes. Furthermore, the P34S BC loop residues move in an opposite direction of the wildtype BC loop residues. This may have implications on the functionality of the enzyme.

The S486T Porcupine plot assessment revealed that the magnitude of the motions on the BC and FG loop were not different between the two enzymes. However, there were some differences in the direction of movements for some of the residues. These may potentially affect the functionality of the enzyme.

The V104M Porcupine plot assessment revealed that the residues on the wildtype FG loop moved in a different direction and magnitude compared to the mutant enzyme. In addition, the centrally located residues of the mutant have greater movements and move in a different direction compared to the wildtype. As these differences involve the catalytic site, the mutation may disrupt the functionality of the enzyme significantly.

The T107I Porcupine plot results showed large differences in the nature of the motion from the FG loop of the mutant and wildtype enzyme. Furthermore, there are greater motions in the opposite region of the mutant enzyme compared to the wildtype enzyme. This suggests that the mutation may potentially cause drastic changes on the functionality of the enzyme.

The P267H mutation results for the Porcupine plot showed that the FG loop residues from the mutant enzyme had movements in the opposite direction compared to the wildtype enzyme. Differences were also observed on the central residues; thus, this mutation may potentially result in significant implications on the enzyme's functionality.

With the R365H mutation Porcupine plot assessment, the BC loop residues on the mutant enzyme moved in a different pattern compared to the wildtype and the residues which are located centrally had a unique pattern in comparison to the wildtype enzyme. These differences show the potential impact of the mutation on the enzyme.

4.9 Secondary Structure Analysis (SSA)

To determine the impact of missense variants on the secondary structure of the enzyme, secondary structure analysis was implemented. The dictionary of secondary structure prediction (DSSP) algorithm was used for this analysis which is embedded in *CPPTRAJ*. The secondary structure elements which were assessed included structural bending, turns, pi (3-14) helix, alpha helix, 3-10 helix, anti-parallel beta sheets and parallel beta sheets. The assessment was carried out over a 500 ns simulation to observe the difference in conformational evolution between the wildtype and mutant enzymes. The results are shown in Figure 4.9. The diagram shows that the variants had some effects on the secondary structure of the enzyme. However, the impacts varied with the different variants.

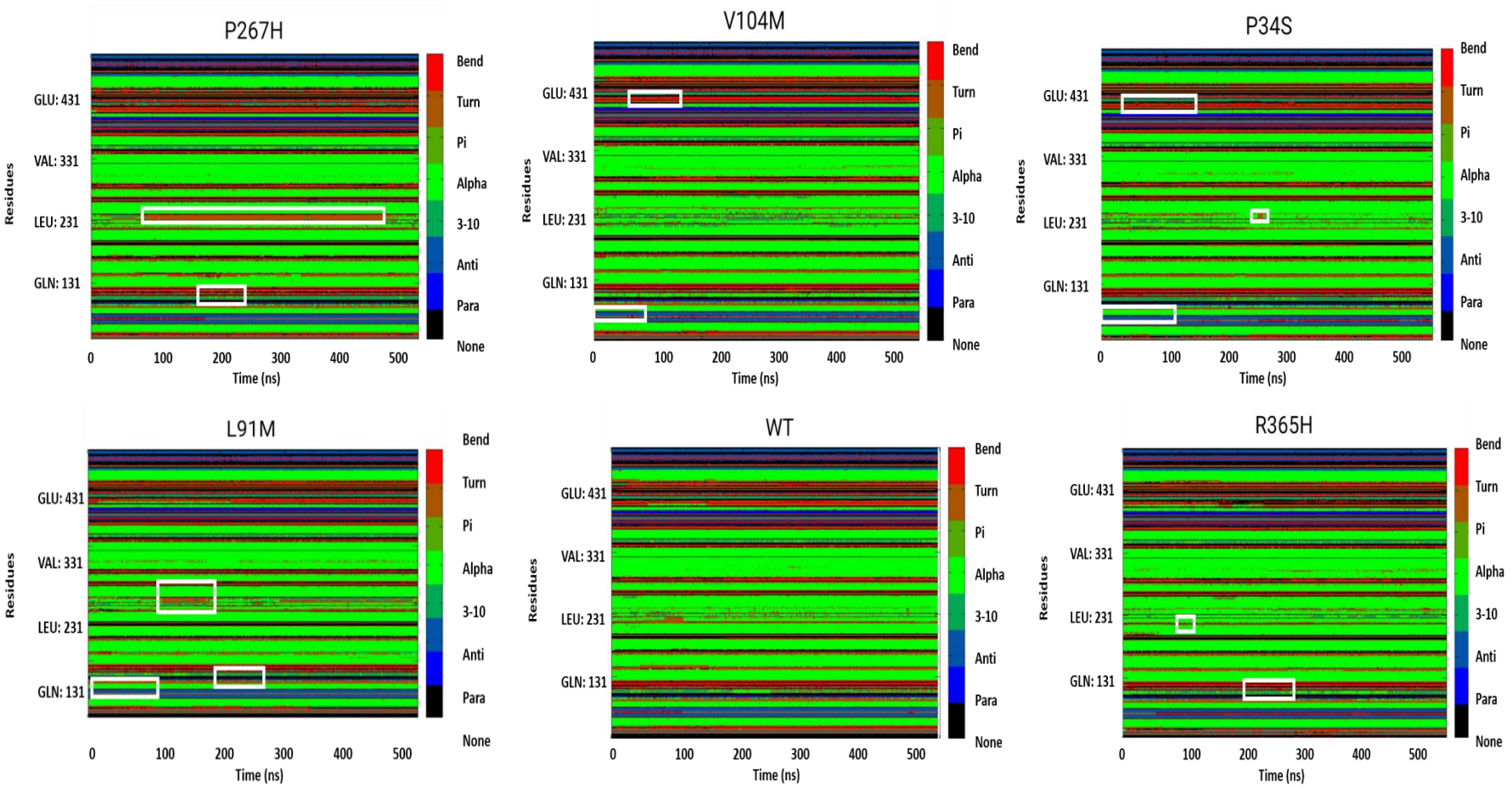
The SSA analysis for the V338M mutant show that the mutation results in changes which include the introduction of a turn on the secondary structure around residue 231 within the simulation period of 200 ns and 400 ns. The 3-10 helix around the 100^{th} residue which occurs throughout the simulation is not present on the mutant enzyme. In addition, the turn that is present on the wildtype throughout the simulation around residue 400 is absent on the mutant enzyme. These alterations may have potential implications on the functionality of the enzyme.

The SSA results for the P267H mutation showed that, on the mutant enzyme, a turn is observed around the 240th residue between the 100 and 400 ns period which is not present on the wildtype enzyme. An alpha helix is observed on the wildtype enzyme between the 200 ns and 300 ns period near the 100th residue was not observed on the mutant enzyme.

For the R365H SSA analysis, notable differences included a turn that only occurs on the 230th position of the wildtype at 100 ns whereas the mutant enzyme had an alpha helix at that position. In addition, the alpha helix observed in the wildtype around residue 100 (200 ns to 300 ns) became more prominent at a later stage of the simulation on the mutant enzyme.

The S486T SSA results included the introduction of a bend around residue 80 from approximately 500 ns. An alpha helix was also observed between ~10 ns and 350 ns from approximately residue 100; however, this element was only evident between 200 and 300 ns on the wildtype enzyme. Around the position which the amino acid change took place, no conformational changes were noticeable.

For the V104M SSA results, the bend observed between 1-100 ns on approximately residue 80 of the wildtype was not observed on the mutant enzyme. This is replaced by a turn on the mutant enzyme. The wildtype enzyme has an alpha helix around position 400 between 100 ns and 200 ns period; however, the mutant consists of a bend in this region.


P267H
V104M
P34S
L91M
WT
R365H
Residues
GLU: 431
VAL: 331
LEU: 231
GLN: 131
0
100
200
300
400
500
Time (ns)
Bend
Turn
Pi
Alpha
3-10
Anti
Para
None

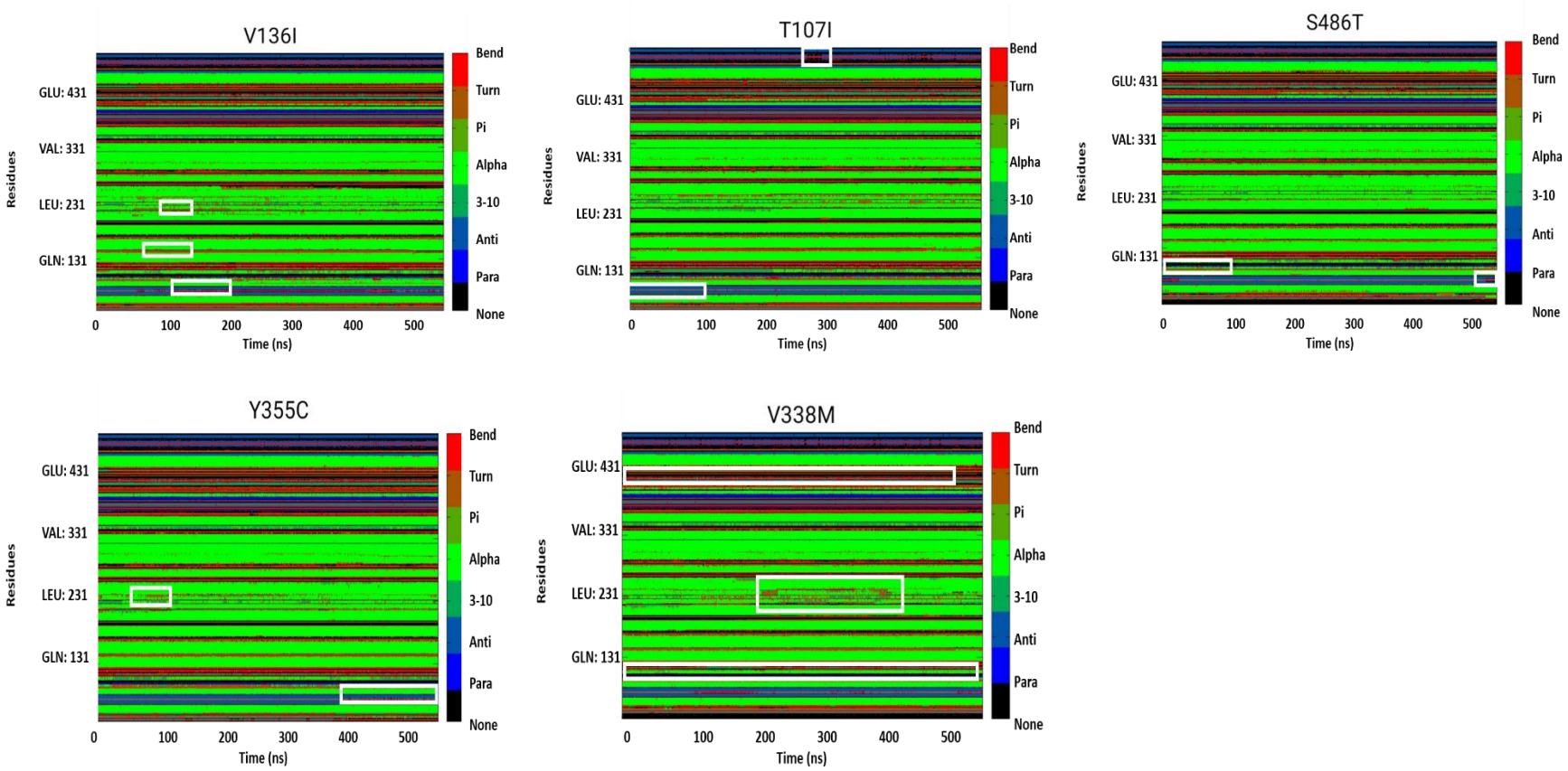


Figure 4.9: Diagram showing the secondary structure analysis of the wildtype enzyme (WT) and the mutant enzymes. The different colours on the plots represent the secondary structure elements. These include the Bend (shown in red), Turn (shown in brown), Pi (3-14) helix (shown in forest-green), Alpha helix (shown in green), 3-10 helix (shown in dark green), anti-parallel Beta sheet (shown in dark blue) and parallel Beta sheet (shown in blue). Black represents "none" of the secondary structure elements. The x-axis represents the simulation time and the y-axis represents the residue positions. The white rectangles on the mutant enzyme plots indicate the regions that had different secondary structure elements compared to the wildtype (WT) enzyme.

The P34S SSA assessments show that the mutant enzyme had a turn that was introduced on the centre of the plot (around 231st residue and ~250 ns). The bend which is observed between 1 and 100 ns on approximately residue 80 of the wildtype enzyme only occurs at a later stage of the simulation on the mutant. Furthermore, the alpha helix element observed on approximately residue 400 (between 100 ns and 200 ns) on the wildtype is replaced by a bend on the mutant enzyme.

The L91M SSA results revealed that a turn was introduced at around residue 231 (between 100 ns and 200 ns) on the mutant enzyme which was not observed on the wildtype structure. The alpha helix that occurs around position 100 in the period between 200 ns and 300 ns on the wildtype enzyme is not present on the mutant structure. A bend appears more prominently between 1 ns and 100 ns around position 80 on the wildtype compared to the mutant enzyme.

For the V136I mutation SSA results, the turn observed around the 231st residue of the wildtype between 100 ns and 110 ns is not evident on the mutant. Furthermore, a bend is introduced on the mutant between 100 ns and 400 ns around residue 80. A turn observed on the wildtype enzyme between 50 ns and 150 ns near residue 131 was not present on the mutant enzyme.

The conformational evolution results of Y355C show a notable alteration on the mutant enzyme is the absence of a turn around 100 ns on approximately residue 231, which is observed on the wildtype enzyme. A bend only occurs on the later stages of the simulation in approximately residue 80.

For the T107I SSA results, a notable bend on the wildtype enzyme around residue 80 (between 1 ns and 100 ns) was replaced by a turn on the mutant enzyme. Furthermore, the anti-parallel beta sheet that occurs on the wildtype around residue 480 in the 300 ns point in time is lost on the mutant enzyme.

Chapter 5: Discussion

The CYP2D6 enzyme is involved in the metabolism of ~25% of commonly used drugs (Monte *et al*., 2018). As a result, variation in the gene coding for the enzyme has implications on how individuals may respond to several families of drugs including anti-psychotics, anti-cancer drugs and analgesics (Ingelman-Sundberg, 2005). Most genetic studies have focused on non-African studies, thus there is a paucity in data for African populations. We, therefore, aimed to assess the potential impact of missense variants on the functionality of the CYP2D6 enzyme relevant to sub-Saharan populations using *structural bioinformatics* methods. The evaluation of the potential impact of the selected variants was performed using *SWAAT* and MD. *SWAAT* was used to identify missense variants with potential significant effects using three features, namely, the ΔΔG score, ML prediction score and red flag. For the ten variants that were selected for MD (Table 4.2), the variants' potential implications on the enzyme's stability, flexibility, movement and secondary structure were assessed using various MD assessments. These included the RMSD, RMSF, PCA, Porcupine plots and Secondary Structure analyses. Table 5.1 shows a summary of the results for the eight variants that were retained based on the SWAAT results and the ten variants that were selected for the MD assessment.

Table 5.1: Summary of SWAAT results for variants retained at SWAAT stage and MD results

SWAAT Results			
Variant	**Implication based on ΔΔG Value**	**ML Prediction Outcome**	**Red Flag Description(s)**
Y355C	Destabilising Effect	Significant Impact	Hotspot patch
P34S	Destabilising Effect	Significant Impact	Buried exposed switch; Hotspot patch
R365H	Destabilising Effect	Significant Impact	Hotspot patch
G42R	Destabilising Effect	Significant Impact	Hotspot patch
G42E	Destabilising Effect	Significant Impact	Hotspot patch
L231P	Destabilising Effect	Significant Impact	Buried proline; Buried exposed switch; Hotspot patch;

M279K	Destabilising Effect		Significant Impact	Salt bridge formation	
P469A	Destabilising Effect		Significant Impact	Buried exposed switch	
MD Results					
Variant	**RMSD**	**RMSF**	**PCA**	**Porcupine plot**	**SSA**
L91M	Converged later than wildtype	Reduced flexibility in FG loop and residue ~260	Altered conformation	Altered movements in FG loop and BC loop	Altered secondary structure elements at residues ~231, ~100 and ~80
P34S	Converged later than wildtype	Reduced flexibility in D helix, KL and GH loop and enhanced flexibility in FG loop	Similar conformation as wildtype	Altered movements in FG loop and BC loop	Altered secondary structure elements at residues ~80, ~231 and ~400
S486T	More stable trajectory compared to wildtype	Reduced flexibility in FG loop and GH loop	Similar Conformation	Similar movements to wildtype with minor differences	Altered secondary structure elements at residues ~80 and BC loop
Y355C	Trajectory coincided with wildtype	Increased flexibility in D helix and FG loop	Altered conformation	No prominent difference compared to wildtype	Altered secondary structure elements at residue ~231 and ~80
V338M	Converged later than the	No prominent impact on the	Altered conformation	Altered movements in	Altered secondary

	wildtype	flexibility		FG loop	structure elements at residue ~100, ~231 and 400
V104M	Converged later than the wildtype	Reduced flexibility in GH loop and FG loop	Altered conformation	Altered movements in FG loop	Altered secondary structure elements at residue ~80 and ~400
V136I	Converged later than the wildtype	Reduced flexibility in GH loop and FG loop	Altered conformation	Altered movements in central residues	Altered secondary structure elements at residues ~231, ~80 and ~131
P267H	Converged earlier than the wildtype	Reduced flexibility in FG loop	Altered conformation	Altered movements in FG loop	Altered secondary structure elements at residues ~100 and ~240
R365H	Converged earlier than the wildtype	No prominent impact on flexibility of enzyme	Altered conformation	Altered movements in BC loop and central residues	Altered secondary structure elements at residues ~100 and ~230
T107I	Converged later than the wildtype	No prominent impact on flexibility of enzyme	Altered conformation	Altered movements in FG loop	Altered secondary structure elements at

					residues ~80 and ~480

5.1 SWAAT Analysis

Fifty *CYP2D6* core missense variants which occur in African populations, according to GnomAD, were selected from PharmVar and processed through *SWAAT*. However, four were not processed successfully, namely, V7M, V11M, R26H and R28C. This is due to the *SWAAT* algorithm that incorporates the crystal structures of proteins from the PDB and the CYP2D6 crystal structures are incomplete, with the first 31 residues missing on the structure. From the 46 variants that were processed successfully, a total of 15 missense variants had a $\Delta\Delta G > 1.00$ as shown on the frequency distribution diagram (Figure 4.3). Only eight variants from the 15 were retained through the *SWAAT* tool which included variants with a machine learning prediction score of 2 and at least one red flag. These variants are shown in Table 4.2 with the results of the machine prediction score and red flag description.

From the eight variants retained based on the *SWAAT* results, all the variants showed a $\Delta\Delta G > 1.0$ and a machine learning score of 2. This indicates that the variants were potentially destabilising and had a potential significant impact on the enzyme's functionality. The R365H mutation had the highest $\Delta\Delta G$ score which suggests that the mutation may be the most destabilising variant based on the score. The P34S, L231P and P469A were mutations that involved the proline residue. These three variants were flagged as "buried exposed switch" which would result in a destabilising impact. The destabilising effect may be as a result of the altering of the backbone hydrogen-bonding pattern due to the removal or introduction of the proline residue (Ittisoponpisan *et al*., 2019). In addition, a buried exposed switch is damaging (Ittisoponpisan *et al*., 2019). Furthermore, according to Kim *et al*., (2013), the P34S mutation is known to reduce enzyme activity. The L231P variant was also flagged as a "buried proline introduced", as a result, this proline may destabilise the backbone structure of the enzyme (Ittisoponpisan *et al*., 2019).

The P469A and M279K were the only two variants from the eight that were not flagged as "hotspot patch" which indicates that the other six variants may be damaging as they occur in hotspot patch regions. From the six missense variants that occur in hotspot patch regions,

G42R and G42E are known to result in a decreased enzyme function according to Tsuzuki *et al*., (2003) which agrees with the *SWAAT* predictions. These effects may be due to the mutations resulting in an acidic amino acid (glutamic acid) or basic amino acid (arginine) which have different charges and larger side chains compared to glycine (Tsuzuki *et al*., 2003; Vnučec *et al*., 2016). G42 forms part of a hydrophobic region (LPLPGL) with bulky amino acids, as a result, the introduction of arginine or glutamic acid, which have longer side chains and are hydrophilic amino acids, may potentially destabilise hydrophobic interactions and ultimately the enzyme (Schrödinger, 2010; Vnučec *et al*., 2016).

The M279K variant was the only mutation to result in a salt bridge formation from the *SWAAT* results. The introduced lysine residue may form a salt bridge with an aspartate (D292) or glutamate residue (E293), and this type of interaction is stronger than common hydrogen bond interactions. However, the strength of a salt bridge may vary with the type of solvent and pH level (Xie *et al*., 2015). According to Sindelar *et al*., (1998), salt bridge formation may destabilise a protein, as a result, the M279K is potentially damaging given the introduction of a strong interaction in a site that normally has weaker interactions.

From the eight variants retained based on *SWAAT*, only P34S, R365H and Y355C were found in the H3Africa dataset, which suggests that these were more common in African populations as compared to the other five. Moreover, R365H and Y355C are core variants of uncharacterised star alleles, *108 (H352R and Y355C), *127 (R365H and Y355C) and *139 (R365H), according to PharmVar (Gaedigk *et al*., 2018). As a result of the potentially deleterious characteristics of the variants, these star alleles may be associated with a decreased enzyme activity. To obtain more insights, further assessments of these two variants were performed with MD. In addition, the P34S, which has been reported as a damaging variant in other studies, was used as a reference for the MD assessments. The assessments are discussed in the next subsections.

5.2 Molecular Dynamics

5.2.1 L91M Mutation

To assess the deviation between the wildtype and mutant enzymes, the Root Mean Square Deviation analysis was performed. From Figure 4.2, the results show the trajectories for the

wildtype and the mutant variants during the 500 ns simulation. The L91M mutation is a core variant for *74 which is an uncharacterised allele that is commonly found in sub-Saharan African populations (Gaedigk *et al*., 2018, Whirl-Carillo *et al*., 2012). The results show that the L91M mutation yields a trajectory that equilibrates later than the wildtype enzyme. This suggests that the conversion of leucine into a methionine in position 91 may potentially result in a destabilising effect. The destabilising effect may be as a result of the unique properties of methionine that is introduced from the mutation. Methionine consists of a sulfur atom and can form bonds with aromatic rings (Lim *et al*., 2019). This suggests that the introduction of methionine may result in new interactions such as the formation of a bond with a residue, potentially F433, which possesses an aromatic ring thus affecting the stability of the protein. Incorporating hydrogen bond analysis and tools such as Dynamut may provide better insight into how atomic interactions are affected by the mutation (Rodrigues *et al*., 2018). The observed RMSD results agree with Don and Smiesko (2018), who assessed missense variants in CYP2D6 *4, which consists of L91M as one of the core variants, and found that the trajectory converged later compared to the wildtype. However, the assessed *4 allele also harbours P34S, which is known to have a damaging effect. As a result, the delayed convergence may be solely due to the P34S mutation. Although, this does not rule out the possibility of the L91M impact being masked by the effect of P34S which would mean that L91M is potentially destabilising together with P34S within *4.

The RMSF analysis was performed to evaluate the flexibility of the residues on the mutant and wildtype enzymes, since flexibility is known to have a significant role in substrate specificity and access (de Waal *et al*., 2014). The RMSF results are shown in Figure 4.3. Some of the most critical loops include the BC loop, which includes residues 59-83 on the RMSF plots (residue position numbering is based on modellers numbering system i.e. 1-466), and the FG loop, which includes residues 186-210 on the RMSF plots (Marquez *et al*., 2018). The BC loop is involved in substrate binding and substrate recognition and the FG loop functions as a hatch (de Waal *et al*., 2014, Marquez *et al*., 2019). The molecular dynamics system was solvated with water and there was no membrane in the system, thus high fluctuations were observed in the residues around the N-terminal proline rich region of the enzyme which is normally anchored by a membrane (Don and Smiesko, 2018). The RMSF results for L91M show that higher fluctuations were observed in the FG loop (around residue 200) and ~260 for the wildtype compared to the mutant. These findings suggest that the FG loop and the residues around residue 260 (I helix region) are more flexible for the wildtype

enzyme compared to the mutant enzyme. These findings reveal that the mutation has an impact on distal residues (i.e around residue 260 with a distance of ~40 Å from the mutation) and not only residues with close proximity (i.e around residue 200 with a distance of ~30 Å from the mutation) at the mutation site. This amino acid change involves a change from leucine which is non-polar and aliphatic while methionine is amphipathic and contains a sulphur (Vnučec *et al*., 2016). Thus, it is expected for this amino acid change to affect the enzyme's flexibility due to the unique amino acid properties, given that methionine can form new interactions, with other residues, that influence the flexibility (Lim *et al*., 2019).

PCA was performed to assess the motion of the wildtype enzyme compared to the mutant enzymes. Furthermore, the movements of the residues were assessed using Porcupine plots which were generated from the PCA eigenvectors. The Porcupine plots show the distance and the direction of the movements of the residues. The PCA and Porcupine plot results are shown in Figures 4.4-4.6. For L91M, the PCA results showed some variation in motion between the mutant enzyme and the wildtype enzyme which suggests that the mutation may have affected the enzyme's nature of movement and further assessment was performed with Porcupine plot analysis to gain more insight and predict the potential implications. The Porcupine plot (Figure 4.6) showed the greatest movement in the FG loop and the BC loop of the mutant. The direction of the movements of the residues in the FG loop were different compared to the wildtype enzyme. The results suggest that the mutation results in altered motions in the FG loop. As a result, since the FG loop functions as a hatch, the altered direction in motion may result in poor accessibility for the substrate to the active site which would diminish the enzyme's activity (Fukuyoshi *et al*., 2016). Given that L91M is understudied, more assessments may be performed to explore the implications of such a mutation.

To elucidate the impact of missense variants on the secondary structure of the enzyme, secondary structure analysis was performed using the DSSP algorithm. The results are demonstrated on Figure 4.7 which shows the comparison between the wildtype and the mutant enzymes. From the secondary structure plot (residue position numbering is based on the PDB crystal structure i.e. residues 31-497) the BC loop includes residues 90 to 114 and the FG loop includes 186 to 210 (Marquez *et al*., 2018). The SSA results for L91M show that the mutation altered the secondary structure of the enzyme at residue ~231, ~100 and ~80

during the simulation. Since this affects the BC loop, these results suggest that the mutation may potentially have a damaging effect on the enzyme. The overall analysis of the L91M mutation suggests that the mutation may potentially result in a damaging effect, thus the *74 allele may be associated decreased enzyme activity.

5.2.2 P34S Mutation

RMSD was used to evaluate the impact of P34S on the stability of the enzyme. The P34S mutation is one of the core variants of the *10 allele, which results in a decreased enzyme activity according to PharmVar (Gaedigk *et al*., 2018). This variant was selected as a positive control for this study. The RMSD trajectory for this mutation converged later than the wildtype and the *SWAAT* results in Table 4.2 indicate that P34S results in a destabilising effect. Furthermore, our results showed that the trajectory for the mutant was lower than that of the wildtype which agrees with Xin *et al*., (2020). A potentially damaging effect was expected as this variant is known to lower the activity of the CYP2D6 enzyme. According to Silvino *et al*., (2016) the P34S mutation results in the disruption of the local hydrophobic interaction network and has an impact on the rigidity of the backbone which explains the observed potential destabilising effect from RMSD and *SWAAT*. Furthermore, the side chain carbon atoms from the P34 residue were found to perform a main chain to side chain polar interaction together with a neighbouring beta strand (Silvino *et al*., 2016).

The impact of P34S on the flexibility of the enzyme was assessed using RMSF. Fluctuations were observed around residues 140 (D helix), 240 (GH loop), 360 (KL loop) and 210 (FG loop), with the wildtype having higher flexibility in all the regions excluding the latter (FG loop). This agrees with Xin *et al*., (2020) who observed that the P34S mutation resulted in altered flexibility in the FG loop from the RMSF results; however, they observed a higher flexibility for the wildtype in the FG loop which is the opposite of our case. Although, the unique changes in our case may still have implications on how the enzyme interacts with the substrate which would ultimately affect the enzyme's activity. The observed results may a be a result of P34 playing a role in imparting rigidity to the backbone, through hydrophobic interactions with V68 and performing a main-chain to side-chain polar interaction with a beta strand (Silvino *et al*., 2016). Therefore, the P34S mutation resulting in higher degree of freedom within the backbone (Silvino *et al*., 2016). Furthermore, our results agree with Don and Smiesko, 2018 who reported that *4 which harbours the P34S had higher fluctuations

compared to the wildtype in the FG loop, however, this was for the apo enzyme, whereas in this study the enzyme was assessed with thioridazine. In addition, this mutation was a positive control, thus potential damaging effects were expected as observed from these findings given that P34 plays a role in maintaining the rigidity of the backbone through local hydrophobic interactions.

The PCA results for the P34S mutation showed moderate and minor differences compared to the wildtype enzyme. However, the Porcupine plot showed the prominent differences which included the altered directions of the FG loop and BC loop residues. Given that the BC loop is responsible for substrate recognition and substrate binding, the altered direction would potentially reduce optimal interaction between the enzyme and the substrate (de Waal *et al*., 2014). Furthermore, the altered direction of the motion in the FG loop residues may also impede optimal interaction between the enzyme and the substrate. This implies that the mutation results in a decreased enzyme activity which agrees with Hongkaew *et al*., (2021) who reports that *10 which harbours P34S has a decreased activity. Furthermore, these results were expected given that the variant was a positive control.

To assess the impact of P34S on the secondary structure of the enzyme, SSA was conducted. The results show that the mutation results in a change in residue ~80, ~231 and ~400. These changes may result from the removal of a proline which may change the backbone hydrogen-bonding (involving the nitrogen atom) pattern thus disrupting the secondary structure (Ittisoponpisan *et al*., 2019; Deepak and Sankararamakrishnan, 2016). These may result in a significant impact on the enzyme's functionality. Furthermore, the P34 residue forms part of the proline rich region which is highly conserved in microsomal P450s and may function as a hinge between the hydrophobic membrane and the heme-binding region of the enzyme (Sakuyama *et al*., 2008). Thus, the P34S mutation may result in a damaging effect and this agrees with Kim *et al*., (2013), who stated that the P34 is located in a conserved region suggesting that P34S is a destabilising variant. The overall results suggest that the mutation is damaging, this was expected since this mutation was previously reported as damaging.

Interestingly, even though the P34S mutation portrays damaging characteristics, the mutation has a high frequency (0.12) in African populations (Table 4.1) (Karczewski *et al*., 2020). This may be due to a number of possible reasons, which include, hitchhiking or reduction in selective constraints such as reduced exposure to xenobiotics that are only metabolised by the

CYP2D6 enzyme (van Hooft *et al*., 2014). Furthermore, some drugs or other xenobiotics are alternatively metabolised by other CYPs (Lynch and Price, 2007). As a result, the deleterious mutation may not cause severe effects on the organism even though the CYP2D6 enzyme may have a reduced activity. Another possible explanation may be that the P34S may exhibit a substrate-specific effect (Marcath *et al*., 2019). Thus, only certain substrates are metabolised by CYP2D6 inadequately. This would reduce the odds of individuals, with the mutation, possibly encountering adverse drug reactions to a point that the P34S is not eliminated through negative selection.

5.2.3 S486T Mutation

The S486T mutation is a neutral control for this study. The impact of S486T on the enzyme's stability was elucidated using the RMSD. This mutation defines the *39 allele and is one of the core variants of other star alleles. The *39 allele is known to have a normal function according to PharmVar (Gaedigk *et al*., 2018). Thus, we selected this variant as a neutral control. The RMSD results showed that the trajectory was stable throughout the 500 ns simulation suggesting that the mutation stabilises the enzyme and has no negative impact. This is supported by Kim *et al*., (2013) who state that the S486T has no impact on the structural basis or functionality of the CYP2D6 enzyme. In addition, this is supported by the high frequency of 0.35 for S486T (Karczewski *et al*., 2020).

The RMSF results showed that the wildtype had higher fluctuations in residues ~200 (FG loop) and ~260 (GH loop). This contradicts Don and Smiesko, 2018 who observed higher fluctuations in the loops for all variants including *2 which has a normal function and has the S486T mutant. However, this was for the APO enzyme and not the ligand bound enzyme. Furthermore, this study assessed the thioridazine drug whereas Don and Smiesko (2018) assessed other drugs which may have implications on the changes that were observed resulting from the unique drug type. This study used the AMBER 18 ff14SB force field which differs from the OPLS force field implemented by Don and Smiesko, (2018). This may have also contributed to the differences observed in the results. Moreover, *2 also includes another variant (R296C) which may have contributed to the observed changes. Our findings imply that the mutation results in altered flexibility of the enzyme in the mentioned regions. However, given the nature of the mutation, which involves two amino acids that share some chemical properties (both hydroxylic), and mutation being known to have not significant

impact on the enzyme activity, these changes may not affect the functionality of the enzyme significantly.

The PCA results for S486T show that the motion of the wildtype and the mutant enzyme differed, however the differences were not that prominent. This suggests that the mutation does not have much implications on the enzyme's functionality. In addition, the Porcupine plots showed that the magnitude of the motions were similar for the residues in the wildtype and the mutant enzyme. However, there were differences in the direction of the motions in some of the residues, although these were not as prominent as in other variants thus the mutation would not have a drastic impact on the enzyme activity. This agrees with Kim *et al*., (2013) who states that only a minor decrease in catalytic efficiency occurred, with dextromethorphan as the substrate, as a result of S486T.

The SSA results reveal that the S486T which was the neutral control resulted in changes in residue ~80 and residue ~100 (BC loop). In contrast, serine and threonine have similar chemical properties (both possess hydroxyl groups) thus not much structural differences were expected between the wildtype and the mutant enzyme (Vnučec *et al*., 2016). These results indicate that the mutation may affect the functionality by altering the secondary structure. This agrees with Kim *et al*., (2013), who reported that the S486T resulted in a small decline (~24%) in catalytic efficiency. However, this may not be a significant effect since the mutation is known to have no significantly altered enzyme activity. The MD results were contrasting, thus the effect of S486T is inconclusive from the study; however, this variant is known to have no drastic impact on the functionality of an enzyme. Thus, the changes observed between the wildtype enzyme and mutant enzyme may not impact the enzyme's functionality significantly. Although, the effect of the variant differs with substrates as reported by Kim *et al*., (2013) who showed different implications of S486T on the catalytic efficiency when assessing bufuralol and dextromethorphan. Moreover, CYPs have shown to have variant alleles that exhibit functional consequences that vary with different substrates (Marceth *et al*., 2019). As a result, we cannot rule out the possibility of S486T having a significant impact, resulting from the altered dynamic effects, on the catalytic efficiency for thioridazine specifically.

5.2.4 Y355C Mutation

Figure 4.2 shows the RMSD trajectories of the wildtype enzyme and the Y355C enzyme. The Y355C variant is found in *108 which is uncharacterised (Gaedigk *et al*., 2018). The results show that the mutant and wildtype trajectories coincided during the entire 500 ns simulation. This suggests that the Y355C mutation does not have a destabilising effect on the enzyme. However, this contradicts the *SWAAT* analysis results, shown in Table 4.2, which predicted that the variant would have a destabilising impact. Furthermore, this was expected to have a destabilising effect given the nature of the amino acid change, i.e., tyrosine and cysteine have different chemical properties (Vnučec *et al*., 2016). Running a longer simulation may potentially reveal the difference in trajectories at a later stage of the simulation provided the mutation is destabilising.

To assess the impact of Y355C on the flexibility of the enzyme residues, RMSF was conducted and the results show that the mutation resulted in higher fluctuations in ~120 (D helix) and ~210 (FG loop) compared to the wildtype. These fluctuations may result from the different chemical properties between the two amino acids, as cysteine has a thiol group while tyrosine has a benzene ring and a hydroxyl group and, thus, this may have implications on how the residue interacts with the other residues (Vnučec *et al*., 2016). The loss of the hydroxyl group as a result of the mutation may potentially result in a loss of a hydrogen bond that can occur on the hydroxyl group which serves to stabilise proteins (Pace *et al*., 2001). Thus, these results suggest that the mutation may potentially affect the flexibility and ultimately functionality of the enzyme.

The results from PCA for the Y355C mutation showed that the mutation causes some changes in the motion of the CYP2D6 enzyme. This results from the Y355C mutation. However, the Porcupine plots demonstrated no prominent differences from the residues of the mutant and the wildtype enzyme. This suggested that the Y355C mutation in the enzyme may not have significant implications on the enzyme's functionality.

The evaluation of the impact of Y355C on the secondary structure was conducted using SSA. The mutation resulted in structural changes in residue ~231 and ~80. These changes may potentially affect the enzyme activity. However, it cannot be ruled out that the changes may not result in a significant effect. In addition, the SSA results were similar to the S486T, which also suggests that the variant may be benign. Analysis of the overall results for Y355C are

inconclusive. Other approaches such as *in vitro* approaches may be performed to further assess the variant to draw a clearer conclusion.

5.2.5 V338M Mutation

To evaluate the impact of V338M on the stability of the enzyme, RMSD was performed and the results are shown on Figure 4.2. The V338M missense variant is one of the core variants of *29 which is known to be associated with a decreased function according to PharmVar (Gaedigk *et al*., 2018). In addition, the star allele is commonly found in sub-Saharan African populations (Whirl-Carillo *et al*., 2012). The V338M yielded a less stable trajectory which converged later than the wildtype enzyme, implying that the variant results in a destabilising effect. Furthermore, the nature of the V338M trajectory is similar to the P34S, which is the positive control. This may suggest that the V338M variant may be the causal variant for the clinical effects associated with *29. However, another variant (V136I) may contribute to the clinical effect observed in *29, since it is also a core variant of *29 (Wang *et al*., 2014). This variant will be discussed later.

The influence of V338M on the flexibility of the enzyme was assessed and the RMSF results showed no prominent differences between the wildtype and the mutant. This suggests that the V338M mutation does not have much effect on the flexibility of the enzyme. These findings suggest that the mutation may not have a significant impact on the functionality of the enzyme. This however, contradicts the implications suggested by the RMSD results for this mutation. Furthermore, this variant is found in *29 which is known to have a decreased activity. Thus, these results suggest that V338M may be in linkage disequilibrium with the causal variant, potentially V136I, rather than V338M being the causal variant of the clinical effects associated with *29. However, these findings cannot rule out the possibility of V338M having a damaging effect for other drugs rather than thioridazine as different drugs have different chemical properties (Don and Smiesko 2018).

The PCA results for the V338M revealed that there were differences in the movements of the wildtype and the mutant enzyme. In addition, the Porcupine plots indicated that the direction of the movements in the FG loop of the mutant enzyme differed compared to the wildtype and the central residues had greater movements and moved in different directions compared to the wildtype enzyme. This suggests that the mutation may result in a reduced enzyme

activity which may be due to the altered enzyme's motion which affects the enzyme and substrate interaction. This agrees with Leitao *et al*., (2020) as this variant is also found in *29 and this allele is associated with a decreased function. This may suggest that both V338M and V136I may be contributing in the reduced enzyme activity associated with *29.

The Secondary Structure Analysis (SSA) results for V338M show that the mutation results in a change in secondary structure elements in some of the regions during the simulation which includes ~231st residue, the BC loop (residue 100) and the 400th residue. This suggests that the change from a valine to a methionine in position 338 has implications on the secondary structure of the enzyme. This may be due to the nature of the amino acid change, which involves two amino acids that have some unique chemical properties, i.e., valine lacks a sulphur atom while methionine contains one (Vnučec *et al*., 2016). As the secondary structure changes involve the BC loop, this indicates that the mutation may disrupt the catalytic activity of the enzyme, given that the BC loop is involved in substrate binding and recognition (de Waal *et al*., 2014). This further suggests that V338M is one of the causal variants of the phenotype associated with *29. The different analysis, excluding the RMSF, suggested that the mutation may potentially affect the functionality of the enzyme. Thus, the V338M mutant might be contributing to the phenotype associated with *29, although more studies may be performed for verification of the mutant's implications.

5.2.6 V104M Mutation

The potential impact of V104M on the stability of the enzyme is illustrated on Figure 4.2. The V104M variant is harboured in *73, which is an uncharacterised star allele commonly found in sub-Saharan African populations (Gaedigk *et al*., 2018, Whirl-Carillo *et al*., 2012). The RMSD results show the trajectory of V104M and the wildtype enzyme. The trajectory of V104M converged later which indicates that the mutation may potentially have a destabilising effect and similar trends were observed with the positive control. Given that the nature of the amino acid change is the same as the V338M variant, similar implications were expected and this was observed since both mutations yielded results that suggested a destabilising impact. Furthermore, this may potentially imply that the *73 allele may result in a decreased enzyme activity due to this mutation. The other two core variants of *73 include S486T and R296C which define *2 (has a normal function). Thus, the V104M may define the

outcome of the *74 implications given that the other two variants do not have a significant effect on the enzyme's functionality (Wang *et al*., 2014).

The V104M implications on the flexibility of the enzyme were evaluated using RMSF. The wildtype enzyme had higher fluctuations in residues ~200 (FG loop) and ~260 (GH loop) which suggests that the mutation reduced the flexibility in those regions. This may potentially have significant implications on the functionality of the enzyme. Interestingly, the effects of this mutation differ from V338M even though the nature of the mutation is the same. This shows that the site of the mutation is also a factor that contributes to the implications of the mutation. These results further support that the V104M may result in a decreased function in *73.

The PCA results for V104M showed that there were prominent differences between the mutant and the wildtype motions for both the PC1 vs PC2 and PC3 vs PC4. This suggests that the mutation alters the mobile characteristics of the enzyme. This may affect the efficiency of the catalytic processes. In addition, the Porcupine plots display that the movements of the central residues and the FG loops differed between the wildtype and the mutant enzyme. These results propose that the V104M mutation may potentially reduce the enzyme activity as a result of the altered movements which would diminish the interactions between the enzyme and the drug. Furthermore, this suggests that *73 is associated with a decreased enzyme activity.

The evaluation for the impact of V104M on the secondary structure of the enzyme was performed using the SSA. The results reveal that there were structural changes at residue ~80 and ~400. These effects may potentially impact the functionality of the enzyme. Furthermore, these results show that the position of the amino acid change has implications at the effects of the mutation, as these effects differ from V338M even though they involve the same amino acid substitution. This analysis together with the other results, suggests that *73 may have an altered enzyme activity resulting from the V104M mutation.

5.2.7 V136I Mutation

The effect of V136I on the enzyme's stability was assessed using RMSD. The V136I mutation is also a core variant of the *29 allele. The RMSD results show that the trajectory of

the mutation only equilibrates later than the wildtype, which was also observed in the P34S mutant, positive control. This implies that this mutant will potentially result in destabilising the enzyme. Thus, this mutant may contribute to the clinical effects associated with *29 which has a decreased enzyme function (Zhou *et al*., 2017).

The impact of V136I on the flexibility of the enzyme was assessed with RMSF and the results show that the wildtype enzyme has higher fluctuations in the FG loop and the GH loop compared to the mutant. This suggests that the V136I variant results in a decreased function which agrees with *29 having a decreased activity (Leitao *et al*., 2020). Moreover, the results suggest that the V136I may be a causal variant for the phenotype associated with *29.

The PCA results for V136I demonstrate that there were differences in the motion of the mutant and wildtype enzyme. This indicates that the mutation causes a change in motion on the enzyme. Moreover, the Porcupine plots revealed that the differences in movements were mainly observed in the central residues. As a result, this may have implications on how the drug may interact with the active site which ultimately affects the catalytic process. This agrees with Leitao *et al*., (2020) since *29 has a decreased function and the allele harbours V136I. Furthermore, this would imply that the mutation is one of the causal variants if not the only causal variant for the *29 phenotype.

The assessment of the impact of V136I on the enzyme's secondary structure was performed using the SSA. The results show that the mutation causes an alteration on the secondary structure in some of the regions. This includes residue ~231, ~80 and ~131 on the enzyme. These changes may potentially reduce the enzyme activity and this would suggest that the V136I mutation contributes in the decreased enzyme activity associated with *29 (Gaedigk *et al*., 2017). All the analyses suggest that the mutation may be damaging and, thus, V136I may be a causal variant of the phenotype associated with *29.

5.2.8 P267H Mutation

The potential impact of P267H on the stability of the enzyme was evaluated with RMSD and the results are shown on Figure 4.8. The P267H is found in *84 which was firstly reported in South Africa and is still uncharacterised (Gaedigk *et al*., 2018). The mutant RMSD trajectory stabilised early together with the wildtype enzyme, which suggests that the mutation does not

result in a destabilising effect. This, however, contradicts with Dodgen *et al*., (2013) who reported that the P267H would result in an altered function due to the altered charge that results from the amino acid change. Histidine consists of 3 pKa values which have implications on the protonation state of Histidine. The pKa values are, ~2.0, ~6.6 and ~9.0 (Grimsley *et al*., 2009). As a result, histidine is considered to have no charge at neutral pH and a positive charge in a pH near 6.0 (Kampmann *et al*., 2006). This contradicts the notion of the P267H mutation introducing a charge (altering the charge) given that both histidine and proline would be neutral at pH 7.0 (Dolinsky *et al*., 2007). Thus, other assays such as *in vitro* studies may be conducted to further assess the implications and provide clarification as well as a greater insight.

The RMSF results for P267H show that the wildtype enzyme had higher fluctuations compared to the mutant enzyme in ~200th residue (FG loop). This suggests that the mutation results in altered flexibility on the FG loop which functions as a hatch. As a result, this may have harmful consequences on the functionality of the enzyme which agrees with Dodgen *et al*., (2013). This suggests that *84 may have a decreased enzyme activity.

The results of the P267H mutation in PCA showed that the mutation resulted in a drastic difference in motion between the mutant enzyme and the wildtype enzyme. Furthermore, the movements from the FG loop residues and the central residues differed with the magnitude and direction between the wildtype and the mutant enzyme. This suggests that the mutation may result in a reduced enzyme activity resulting from these changes. This agrees with Gaedigk *et al*., (2017) who reports that the P267H may cause a decreased function, thus *84 may be associated with a decreased enzyme activity.

The impact of P267H in the secondary structure was evaluated using SSA. The results show that the mutation causes structural changes in the 240th residue and BC loop (100th residue) during the simulation. Proline is a hydrophobic amino acid, in contrast histidine is a hydrophilic and basic amino acid. Thus, these unique properties may account for the effects observed from the mutation (Vnučec *et al*., 2016). Furthermore, the changes may affect the functionality by causing poor interaction between the enzyme and the drug due to the altered BC loop structure resulting from the mutation. This agrees with Gaedigk *et al*., (2017), who report that the P267H may result in a decreased function. Thus, the overall results suggest

that *84 is associated with a decreased enzyme activity, although the results on the stability were inconclusive as the *SWAAT* and RMSD results were contradicting.

5.2.9 R365H Mutation

The R365H mutation is found in an uncharacterised allele (*139). Figure 4.9 shows the impact of this mutation on the stability of the enzyme. The results show that both the wildtype and R365H trajectories converged early. This suggested that the mutation does not have an impact on the stability of the enzyme. This may be due to both histidine and arginine being basic amino acids thus share some chemical properties. However, this contradicted the *SWAAT* results (Table 4.2) that indicated that the mutation results in a destabilising effect. In addition, arginine's side chain has a very high pKa value (~12.48) compared to histidine's side chain (~6.6), thus the histidine side chain is deprotonated at a neutral pH while arginine's side chain is still protonated (Dolinsky *et al*., 2007). As a result, the two amino acids will likely have different charges in the physiological pH of hepatocytes (pH ~7.0) which may be potentially be destabilising and this supports the *SWAAT* results (Berezhkovski *et al*., 2013).

The R365H results show no prominent differences other than from the first 30 residues in the N-terminal region which may be due to the absence of the membrane that stabilises this region. This suggests that the R365H does not have a significant impact on the flexibility of the enzyme and potentially the functionality of the enzyme which contradicts the results from *SWAAT* which suggest that there may be a potential negative impact on the enzyme's functionality.

The PCA results for R365H revealed that the mutation results in different motion which may potentially affect how the enzyme interacts with the enzyme. In addition, the movements of the residues in the BC loop and the central residues in the mutant differed from the wildtype enzyme. This suggests that the R365H mutation may result in a decreased enzyme function. This agrees with Ittisoponpisan *et al*., (2019) who reports that an amino acid substitution that alters a charge may be damaging (histidine's side chain is deprotonated in the physiological pH of hepatocytes whereas with arginine the side chain is still protonated).

The SSA results for R365H demonstrated that the mutation has an impact on the secondary structure of the protein. The R365H mutation caused structural changes in ~230th residue and the BC loop (residue 100) in the simulation. Even though both amino acids are basic, arginine has a higher pKa value which has implications on the charge differences between histidine and arginine at pH 7.0 (Pace *et al*., 2009). This may have implications on the secondary structure which has been observed in these results. The structural change that occurred in the BC loop suggests that the mutation results in a potential significant impact on the enzyme's functionality. The overall results were inconclusive, given that there were contrasting results observed in a number of analyses. Thus, more studies should be carried out to assess the variant.

5.2.10 T107I Mutation

The T107I mutation is one of the core variants of *17 which is associated with a decreased function (Gaedigk *et al*., 2018). The allele is also found in sub-Saharan Africa (Whirl-Carillo *et al*., 2012). The RMSD results for this variant showed that the mutant trajectory coincided with the wildtype enzyme for the first 400 ns, however between 400 and 460 ns there were fluctuations and stabilisation only took place in ~460 ns. This suggested that the mutation had a destabilising effect on the enzyme, given that convergence took place later. Moreover, the variant may be the causal variant for the decreased activity associated with *17. These results agree with Don and Smiesko (2018) which reported that the *17 RMSD took longer to converge compared to the wildtype enzyme.

The RMSF results for T107I show that the wildtype had higher fluctuations in the first residues from the N-terminal region which may be a result of the absence of a membrane and no prominent differences were observed from other regions. This suggests that there may be no significant impact on the flexibility and ultimately the functionality of the enzyme which may explain the high frequency of 0.19 for T107I (Karczewski *et al*., 2020). However, this contradicts Wright *et al*., (2010) who reports that the T107I contributes to the diminished enzyme activity. Furthermore, Don and Smiesko (2018) had higher RMSF fluctuations for the *17 allele compared to the wildtype. However, this allele includes other two variants which may also have accounted on the higher fluctuations observed in *17. This is supported by Cacabelos *et al*., (2012), who stated that T107I showed different implications on the enzyme when assessed with other variants as compared to when assessed as the only variant

in the enzyme. In addition, the T107I mutation has been reported to show substrate-specific effects which may be a plausible reason why the results from this study contrasted Don and Smiesko (2018) who assessed different substrates compared to the one (thioridazine) in this study (Cacabelos *et al*., 2012).

The PCA results of T107I showed that the mutation caused an altered motion between the mutant and the wildtype. The Porcupine plots portrayed that the mutation resulted in different movements of the FG loops and other regions of the enzyme which suggests that the mutation may potentially result in a detrimental effect. This agrees with Leitao *et al*., 2020 as it reports that *17 which harbours T107I has a decreased enzyme activity. Furthermore, *17 was found to have ~10-30% of the wildtype activity according to Zanger *et al*., (2020).

The effect of T107I on the secondary structure of CYP2D6 was evaluated using the SSA. The changes that resulted from this mutation occurred in residues ~80 and ~480. These may have significant implications on the functionality of the enzyme. In addition, the chemical properties of the two amino acids are unique, i.e., threonine has a thiol and is hydrophilic, while isoleucine is hydrophobic. This suggests that this change may be detrimental. Furthermore, this may suggest that the mutation is the causal variant of the decreased enzyme activity in *17 (Del Tredici *et al*., 2018). Except for the RMSF results, all the MD results suggested that the mutation may be potentially damaging and a causal variant of the enzyme activity associated with *17. Future MD studies may incorporate all three core variants to assess the overall effect, as some of the MD assessments showed that S486T, which is one of the *17 core variants, may potentially affect the functionality of the enzyme.

5.3 Strengths and Limitations

This study provides an insight of how missense variants may potentially impact the structure and functionality of an enzyme at molecular level. The findings of the study may provide a direction in identifying some biomarkers for treatment guidelines in Africa. Our approach included controls which were used as a guide to interpret the results for some of the assessed variants and increase the reliability of our findings. Furthermore, we managed to observe the impact of the variants at molecular level in a 500 ns simulation which is shorter than the 1 µs simulation employed by Don and Smiesko (2018). As a result, this approach may provide reliable results in a cost-effective manner.

However, there are some limitations from this study. Firstly, the *SWAAT* tool incorporates PDB structures for analysis, as a result, missing residues from the PDB structure are not processed. In this study, the V7M, V11M, R26H and R28C residues were not processed by *SWAAT* since the first 31 residues are missing on the CYP2D6 PDB structures. Secondly, even though studying the effects of the variants in isolation provides an idea of whether a variant may be a potentially a pathogenic variant or not, the epistatic effects are not observed and, thus, the findings do not show how the variants would influence the phenotype in the presence of other variants (haplotype groups). Thirdly, our molecular dynamics system did not include a membrane and the enzyme model was incomplete, as a result of the missing residues on the crystal structure. Thus, our system may not be the best representation of a biological system. However, our system is still valid as it includes some properties that mimic biological systems. Fourthly, the method used in this study does not assess the impact of missense variants (involving regulatory sites) on post-translational modification which would also have implications on drug response. Fifthly, only one drug was included in the simulation, as a result, the impacts observed in this study may only be relevant for this particular drug as different drugs have different chemical properties and, thus, they may interact differently with the same enzyme and have unique implications for the catalytic efficiency. Lastly, no functional assay was performed in this study to verify the prediction.

Chapter 6: Conclusion

In conclusion, only eight missense variants from the fifty missense variants, obtained from PharmVar, were retained based on the SWAAT results. Only two missense variants from the eight were selected for MD analysis. MD assessments were performed on ten selected missense variants using a CYP2D6/thioridazine complex model (3TBG) that was acquired from PDB. The study provided an insight of the potential impact of variants in uncharacterised alleles, such as the L91M which has potential damaging effects and is harboured in an uncharacterised allele (*74). The study also provided an idea of potential causal variants in some of the alleles that are known to have a decreasing enzyme activity effect such as *29 and *17. This provides potential biomarkers that may be used for treatment guidelines of drugs involving the CYP2D6 enzyme. The findings also show that mutations involving amino acids with similar properties may also potentially cause significant effects such as with R365H (basic amino acids). The findings support the hypothesis, as it was observed that the missense variants affected the dynamics of the enzyme. These findings serve as a stepping stone for optimising the efficiency of precision medicine in Africa.

6.1 Future Work

To better understand how variants may impact the enzyme's functionality at molecular level, more studies may be performed which include other molecular dynamics assessments on the studied variants. This may include assessments such as assessing the effect of the variants on the catalytic pocket volume, residue interaction network, dynamic cross-correlation and binding affinity. Assessing the impact on the catalytic pocket volume may provide understanding on how variants may affect accessibility of the catalytic site by substrates. Evaluating the impact of the variants on residue interaction network may provide insight on how the mutations may affect the interaction between the residues of the enzyme which has implications on the efficiency of the enzyme. Dynamic cross-correlation analysis may provide greater understanding of the implications of variants on the collective motion. To further understand the impact of the variants at molecular level, assessing the binding affinity of the substrate may provide understanding on how variants may alter enzyme activity.

Future work may also incorporate assessing other drugs metabolised by CYP2D6 by molecular dynamics. This may provide insight on which variants may potentially exhibit

substrate specific variant effect, given that CYP2D6 substrates have some unique chemical properties. In addition, the studies may also include assessing the overall core variants in uncharacterised star alleles. To provide a more valid representation of a biological system, future studies may also include assessing the enzyme with the membrane in the system and use a structure from the AlphaFold 2 database which may possibly consist of more complete protein structures. This will provide a better representation of a biological system.

6.2 Future Directions

Precision medicine is a complex model and, thus, there are several other factors that need to be taken into account to optimise the efficiency of the application. To achieve this, more studies involving other assays such as functional assays, *in vitro* and *in vivo* studies should be performed to verify or provide clarification to the findings of this study in order to draw more robust conclusions and enhance precision medicine. Studying other types of variants and other ADME genes may reveal potential epistatic effects and this will expand the genetic information pertaining to African populations. Furthermore, studies incorporating other factors involved in drug response such as environmental factors, sex, age, gut microbiome and lifestyle should be performed to enhance precision medicine applications which aim to reduce adverse drug reactions by optimising efficacy (beneficence) and minimising toxicity (non-maleficence).

References

Abduljaleel, Z. (2019). Structural and functional analysis of human lung cancer risk associated hOGG1 variant Ser326Cys in DNA repair gene by molecular dynamics simulation. *Non-coding RNA Research*. 4(2019): 109-119. doi: 10.1016/j.ncrna.2019.10.002.

Ampadu, H.H., Hoekman, J., de Bruin, M.L., Pal, S.N., Olsson, S., Sartori, D., Leufkens, H.G. and Dodoo, A.N. (2016). Adverse Drug Reaction Reporting in Africa and a Comparison of Individual Case Safety Report Characteristics Between Africa and the Rest of the World: Analyses of Spontaneous Reports in VigiBase®. *Drug Safety*. 39(4): 335-45. doi: 10.1007/s40264-015-0387-4.

Andrade, E.L., Bento, A.F., Cavalli, J., Oiliveira, S.K., Freitas, C.S., Marcon, R., Schwanke, R.C., Siqueira, J.M. and Calixto, J.B. (2016). Non-clinical studies, required for new drug development – Part 1: early in silico and in vitro studies, new target discovery and validation, proof of pribciples and robustness of animal studies. *Brazilian Journal of Medical and Biological Research*. 49(11):1-9. doi: 10.1590/1414-431X20165644.

Baig, M.S., Roy, A., Saqib, U., Rajpoot, S., Srivastava, M., Naim, A., Liu, D., Saluja, R., Faisal, S.M., Pan, Q., Turkowski, K., Darwhekar, G.N. and Savai, R. (2018). Repurposing Thioridazine (TDZ) as an anti-inflammatory agent. *Scientific Reports*. 8(1): 12471. doi: 10.1038/s41598-018-30763-5.

Bartlett, G.J., Porter, CT., Borkakoti, N. and Thornton, J.M. (2002). Analysis of catalytic residues in enzyme active sites. *Journal of Molecular Biology*. 324(1):105-21. doi: 10.1016/s0022-2836(02)01036-7.

Beauchamp, T. and Childress, J. (2019). Principles of Biomedical Ethics: Marking Its Fortieth Anniversary. *The American Journal of Bioethics.* 19 (11): 9-12. doi: 10.1080/15265161.2019.1665402.

Berezhkovskiy, L.M., Wong, S. and Halladay, J.S. (2013). On the maintenance of hepatocyte intracellular pH 7.0 in the in-vitro metabolic stability assay. *Journal of Pharmacokinetics and Pharmacodynamics*. 40(6): 683-689. doi: 10.1007/s10928-013-9339-8.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*. 28: 235-242. doi: 10.1093/nar/28.1.235.

Blazewicz, J., Frohmberg, W., Kierzynka, M., Pesch, E. and Wojciechowski, P. (2011). Protein alignment algorithms with an efficient backtracking routine on multiple GPUs. *BMC Bioinformatics*. 12(181). https://doi.org/10.1186/1471-2105-12-181.

Boras, B.W., Hirakis, S.P., Votapka, L.W., Malmstrom, R.D., Amaro, R.E. and McCulloch, A.D. (2015). Bridging scales through multiscale modeling: A case study on protein kinase A. *Frontiers in Physiology*. 6(250): 1-15. doi: 10.3389/fphys.2015.00250.

Brown, D.K. and Bishop, O.T. (2017). The role of structural bioinformatics in drug discovery via computational SNP analysis – a proposed protocol for analysing variation at the protein level. *Global Heart*. 12(2): 151-161. doi: 10.1016/j.gheart.2017.01.009.

Cacabelos, R., Martínez, R., Fernández-Novoa, L., Carril, J. C., Lombardi, V., Carrera, I., Corzo, L., Tellado, I., Leszek, J., McKay, A. and Takeda, M. (2012). Genomics of Dementia: APOE- and CYP2D6-Related Pharmacogenetics. *International Journal of Alzheimer's Disease*. 2012: 518901: 1-38. https://doi.org/10.1155/2012/518901.

Case, D.A., Ben-Shalom, I.Y., Brozell, S.R., Cerutti, D.S., Cheatham, T.E., Cruzeiro, V.W.D., Darden, T.A., Duke, R.E., Ghoreishi, D., Gilson, M.K., Gohlke, H., Goetz, A.W., Greene, D., Harris, R., Homeyer, N., Izadi, S. Kovolenko, A., Kurtzman, T., Lee, T.S., LeGrand, P., Li, P. Lin, C., Liu, J., Luchko, T., Luo, R., Mermelstein, D.J., Mers, K.M., Miao, Y., Monard, G., Nguyen, C., Nguyen, H., Omelyan, I., Onufriev, .A., Pan, F., Qi, R., Roe, D.R., Roitberg, A., Sagui, C., Schott-Verdugo, S., Shen, J., Simmerling, C.L., Smith, J., Salomon-Ferrer, R., Swails, J., Walker, R.C., Wang, J., Wei, H., Wolf, R.M., Wu, X., Xiao, L., York, D.M. and Kollman, P.A. (2018). AMBER 2018, University of California, San Fransisco.

Chang, C. C., Hsieh, M. H., Wang, J. Y., Chiu, N. Y., Wang, Y. H., Chiou, J. Y., Huang, H. H. and Ju, P. C. (2018). Association between Thioridazine Use and Cancer Risk in Adult Patients with Schizophrenia-A Population-Based Study. *Psychiatry Investigation*. 15(11): 1064–1070. doi: 10.30773/pi.2018.10.10.1.

Chen, J., Wang, J. and Zhu, W. (2016). Molecular Mechanism and Energy Basis of Conformational Diversity of Antibody SPE7 Revealed by Molecular Dynamics Simulation and Principal Component Analysis. *Scientific Reports*. 6: 36900. https://doi.org/10.1038/srep36900.

Choong YS, Tye GJ, Lim TS. (2013). Minireview: applied structural bioinformatics in proteomics. *The Protein Journal*. 32(7): 505-11. doi: 10.1007/s10930-013-9514-1.

David, C. C. and Jacobs, D. J. (2014). Principal component analysis: a method for determining the essential dynamics of proteins. *Methods in Molecular Biology (Clifton, N.J.)*. 1084: 193–226. https://doi.org/10.1007/978-1-62703-658-0_11.

de Waal, P.W., Sunden, K.F. and Furge, L.L. (2014). Molecular dynamics of CYP2D6 Polymorphisms in the absence and presence of a mechanism-based inactivator reveals changes in local flexibility and dominant substrate access channels. *Plos one*. 9(10): 1-12. doi: 10.1371/journal.pone.0108607.

Deepak, R. N. and Sankararamakrishnan, R. (2016). Unconventional N-H…N Hydrogen Bonds Involving Proline Backbone Nitrogen in Protein Structures. *Biophysical Journal*. 110(9): 1967–1979. https://doi.org/10.1016/j.bpj.2016.03.034.

Del Tredici, A.L., Malhotra, A., Dedek, M., Espin, F., Roach, D., Zhu, G.D., Voland, J. and Moreno, T.A. (2018). Frequency of *CYP2D6* Alleles Including Structural Variants in the United States. *Frontiers in Pharmacology*. 9(305): 1-13. doi: 10.3389/fphar.2018.00305.

Dines, J.N., Shirts, B.H., Slavin, T.P., Walsh, T., King, M., Fowler, D.M. and Pritchard, C.C. (2020). Systematic misclassification of missense variants in BRCA1 and BRCA2 "coldspots". *Genetics in Medicine*. 1-6. doi: 10.1038/s41436-019-0740-6.

Dodgen, T.M., Hochfeld, W.E., Fickl, H., Asfaha, S.M., Durandt, C., Rheeder, P., Drögemöller, B.I., Wright, G.E., Warnich, L., Labuschagne, C.D., van Schalkwyk, A., Gaedigk, A. and Pepper, M.S. (2013). Introduction of the AmpliChip CYP450 Test to a

South African cohort: a platform comparative prospective cohort study. *BMC Medical Genetics*. 29;14:20. doi: 10.1186/1471-2350-14-20.

Dolinsky, T. J., Czodrowski, P., Li, H., Nielsen, J. E., Jensen, J. H., Klebe, G. and Baker, N. A. (2007). PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Research*. 35: W522–W525. doi: 10.1093/nar/gkm276.

Duncan, I., Maxwell, T.L., Huynh, N. and Todd, M. (2020). Polypharmacy, Medication Possession, and Deprescribing of Potentially Non-Beneficial Drugs in Hospice Patients. *American Journal of Hospice & Palliative Medicine*. 37(12): 1076-1085. doi: 10.1177/1049909120939091.

Frohlich, E. and Salar-Behzadi, S. (2014). Toxicological assessment of inhaled nanoparticles: Role of in vivo, ex vivo, in vitro and in silico studies. *International Journal of Molecular Science*. 15: 4795-4822. doi: 10.3390/ijms15034795.

Fukuyoshi, S., Kometani, M., Watanabe, Y., Hiratsuka, M., Yamaotsu, N., Hirono, S., Manabe, N., Takahashi, O. and Oda, A. (2016). Molecular Dynamics Simulations to Investigate the Influences of Amino Acid Mutations on Protein Three-Dimensional Structures of Cytochrome P450 2D6.1, 2, 10, 14A, 51, and 62. *Plos one*. 11(4): 1-16. doi: 10.1371/journal.pone.0152946.

Gaedigk, A., Ingelman-Sundberg, M., Miller, N.A., Leeder, J.S. Whirl-Carrillo, M., Klein, T.E. and the PharmVar Steering Committee. (2018). The Pharmacogene Variation (PharmVar) Consortium: Incorporation of the human cytochrome P450 (*CYP*) allele nomenclature database. *Clinical Pharmacology and Therapeutics*. 103(3): 399-401. doi: 10.1002/cpt.910.

Gaedigk A, Twist GP, Farrow EG, Lowry JA, Soden SE, Miller NA. (2017). In vivo characterization of CYP2D6*12, *29 and *84 using dextromethorphan as a probe drug: a case report. *Pharmacogenetics*. 18(5):427-431. doi: 10.2217/pgs-2016-0192.

Gameiro G.R., Sinkunas V., Liguori G.R. and Auler-Júnior J.O.C. (2018). Precision Medicine: Changing the way we think about healthcare. *Clinics.* 37(3): 1-6. doi: 10.6061/clinics/2017/e723.

Gill, P. S., Yu, F. B., Porter-Gill, P. A., Boyanton, B. L., Allen, J. C., Farrar, J. E., Veerapandiyan, A., Prodhan, P., Bielamowicz, K. J., Sellars, E., Burrow, A., Kennedy, J. L., Clothier, J. L., Becton, D. L., Rule, D. and Schaefer, G. B. (2021). Implementing Pharmacogenomics Testing: Single Center Experience at Arkansas Children's Hospital. *Journal of Personalized Medicine*. 11(5), 394. https://doi.org/10.3390/jpm11050394.

Grimsley, G. R., Scholtz, J. M. and Pace, C. N. (2009). A summary of the measured pK values of the ionizable groups in folded proteins. *Protein Science: A Publication of the Protein Society*. 18(1): 247–251. https://doi.org/10.1002/pro.19.

Haider, S., Parkinson, G. N. and Neidle, S. (2008). Molecular dynamics and principal components analysis of human telomeric quadruplex multimers. *Biophysical Journal*. 95(1): 296–311. https://doi.org/10.1529/biophysj.107.120501.

Holliday, G.L., Mitchell, J.B.O. and Thornton, J.M. (2009). Understanding the Functional Roles of Amino Acid Residues in Enzyme Catalysis. *Journal of Molecular Biology*. 390(3): 560-577. doi: 10.1016/j.jmb.2009.05.015.

Hollingsworth, S.A. and Dror, R.O. (2018). Molecular dynamics simulation for all. *Neuron*. 99(6): 1129-1143. doi: 10.1016/j.neuron.2018.08.011.

Hongkaew, Y., Gaedigk, A., Wilffert, B., Ngamsamut, N., Kittitharaphan, W., Limsila, P. and Sukasem, C. (2021). Relationship between CYP2D6 genotype, activity score and phenotype in a pediatric Thai population treated with risperidone. *Scientific Reports*. 11(1): 1-8. doi: 10.1038/s41598-021-83570-w.

Humphrey, W., Dalke, A. and Schulten, K. (1996). "VMD - Visual Molecular Dynamics", *Journal of Molecular Graphics and Modelling*. 14:33-38. doi: 10.1016/0263-7855(96)00018-5.

Huff, H. C., Vasan, A., Roy, P., Kaul, A., Tajkhorshid, E., and Das, A. (2021). Differential Interactions of Selected Phytocannabinoids with Human CYP2D6 Polymorphisms. *Biochemistry*. 60(37): 2749–2760. https://doi.org/10.1021/acs.biochem.1c00158

Ingelman-Sundberg M. (2005). Genetic polymorphisms of cytochrome P450 2D6 (CYP2D6): clinical consequences, evolutionary aspects and functional diversity. *The Pharmacogenomics Journal*. 5(1):6-13. doi: 10.1038/sj.tpj.6500285.

Ito, Y., Kondo, H., Goldfarb, P.S. and Lewis, D.F. (2008). Analysis of CYP2D6 substrate interactions by computational methods. *Journal of Molecular Graphics and Modelling*. 26(6): 947-56. doi: 10.1016/j.jmgm.2007.07.004.

Ittiosoponpisan, S., Islam, S.A., Khanna, T., Alhuzimi, E., David, A. and Sternberg, M.J.E. (2019). Can predicted protein 3D structures provide reliable insights into whether missense variants are disease associated? *Journal of Molecular Biology*. 431: 2197-2212. doi: 10.1016/j.jmb.2019.04.009.

Jandova Z, Gill SC, Lim NM, Mobley DL, Oostenbrink C. (2019). Binding Modes and Metabolism of Caffeine. *Chemical Research in Toxicology*. 32(7): 1374-1383. doi: 10.1021/acs.chemrestox.9b00030.

Jarvis, J.P., Peter, A.P. and Shaman, J.A. (2019). Consequences of *CYP2D6* Copy-Number Variation for Pharmacogenomics in Psychiatry. *Frontiers in Psychiatry*. 10(432): 1-14. doi: 10.3389/fpsyt.2019.00432.

Kalman, L. V., Agúndez, J., Appell, M. L., Black, J. L., Bell, G. C., Boukouvala, S., Bruckner, C., Bruford, E., Caudle, K., Coulthard, S. A., Daly, A. K., Del Tredici, A., den Dunnen, J. T., Drozda, K., Everts, R. E., Flockhart, D., Freimuth, R. R., Gaedigk, A., Hachad, H., Hartshorne, T., Ingelman-Sundberg, M., Klein, T.E., Lauschke, V.M., Maglott, D.R., McLeod, H.L., McMillin, G.A., Meyer, U.A., Müller, D.J., Nickerson, D.A., Oetting, W.S., Pacanowski, M., Pratt, V.M., Relling, M.V., Roberts, A., Rubinstein, W.S., Sangkuhl, K., Schwab, M., Scott, S.A., Sim, S.C., Thirumaran, R.K., Toji, L.H., Tyndale, R.F., van Schaik, R., Whirl-Carrillo, M., Yeo, K. and Zanger, U. M. (2016). Pharmacogenetic allele nomenclature: International workgroup recommendations for test result reporting. *Clinical Pharmacology and Therapeutics*. 99(2): 172–185. doi: 10.1002/cpt.280.

Kamaraj, B. and Purohit, R. (2013). In silico screening and molecular dynamics simulation of disease-associated nsSNP in TYRP1 gene and its structural consequences in OCA3. *BioMed Research International*. 2013: 1-14. doi: 10.1155/2013/697051.

Karczewski, K.J., Francoli, L.C., Tiao, G., Cummings, B.B., Alfoldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., Gauthier, L.D., Brand, H., Solomonson, M., Watts, N.A., Rhodes, D., Singer-Berk, M., Seaby, E.G., Kosmicki, J.A. Walters, R.K., Tashman, K., Farjoun, Y., Banks, E., Poterba, T., Wang, A., Seed, C., Whiffin, N., Chong, J.X., Samocha, K.E., Pierce-Hoffman, E., Zappala, Z., O'Donnell-Luria, Minikel, E.V., Bergelson, L., Cibulskis, K., Connolly, K.M., Covarrubias, M., Donnelly, S., Ferriera, S., Gabriel, S., Gentry, J., Gupta, N., Jeandet, T., Kaplan, D., Llanwarne, C., Munshi, R., Novod, S., Petrillo, N., Roazen, D., Ruano-Rubio, V., Saltzman, A., Schleicher, M., Soto, J., Tibbetts, K., Tolonen, C., Wade, G., Talkowski, M.E., The Genome Aggregation Database Consortium, Neale, B.M., Daly, M.J. and MacArthur, D.G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 581(7809), 434–443. doi: 10.1038/s41586-020-2308-7.

Khalil, H. and Huang, C. (2020). Adverse drug reactions in primary care: a scoping review. *BMC*. 20(1): 5. doi: 10.1186/s12913-019-4651-7.

Kim, J., Lim, Y.R., Han, S., Han, J.S., Chun Y.J., Yun, CH., Lee C.H. and Kim D. (2013). Functional influence of human CYP2D6 allelic variations: P34S, E418K, S486T, and R296C. *Archives of Pharmacal Research*. 36(12): 1500-1506. doi: 10.1007/s12272-013-0212-5.

Knapp, B., Frantal, S., Cibena, M., Schreiner, W. and Bauer, P. (2011). Is an intuitive convergence definition of molecular dynamics simulations solely based on the root mean square deviation possible?. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*. 18(8): 997–1005. https://doi.org/10.1089/cmb.2010.0237.

Lamb, D. C. and Waterman, M. R. (2013). Unusual properties of the cytochrome P450 superfamily. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*. 368 (1612): 1-13. doi: 10.1098/rstb.2012.0434.

Leitão, L., Souza, T. P., Rodrigues, J., Fernandes, M. R., Santos, S. and Santos, N. (2020). The Metabolization Profile of the *CYP2D6* Gene in Amerindian Populations: A Review. *Genes*. 11(3): 1-14. doi: 10.3390/genes11030262.

Li, H., Dawood, M., Khayat, M.M., Farek, J.R., Jhangiani, S.N., Khan, Z.M., Mitani, T., Coban-Akdemir, Z., Lupski, J.R., Venner, E., Posey, J.E., Sabo, A., Gibbs, R.A. (2021). Exome variant discrepancies due to reference-genome differences. *American Journal of Human Genetics*. 108(7): 1239-1250. doi: 10.1016/j.ajhg.2021.05.011.

Li, J. and Koehl, P. (2014). 3D representations of amino acids-applications to protein sequence comparison and classification. *Computational and Structural Biotechnology Journal*. 11(18): 47-58. doi: 10.1016/j.csbj.2014.09.001.

Likhachev, I. V., Balabaev, N. K. and Galzitskaya, O. V. (2016). Available Instruments for Analyzing Molecular Dynamics Trajectories. *The Open Biochemistry Journal*. 10: 1–11. https://doi.org/10.2174/1874091X01610010001.

Lim, J. M., Kim, G. and Levine, R. L. (2019). Methionine in Proteins: It's Not Just for Protein Initiation Anymore. *Neurochemical research*. 44(1): 247–257. https://doi.org/10.1007/s11064-017-2460-0.

Lonsdale, R., Rouse, S. L., Sansom, M. S. and Mulholland, A. J. (2014). A multiscale approach to modelling drug metabolism by membrane-bound cytochrome P450 enzymes. *Plos Computational Biology*. 10(7): 1-16. https://doi.org/10.1371/journal.pcbi.1003714.

Lopes, P. E., Guvench, O. and MacKerell, A. D. Jr (2015). Current status of protein force fields for molecular dynamics simulations. *Methods in Molecular Biology (Clifton, N.J.)*. 1215: 47–71. doi: 10.1007/978-1-4939-1465-4_3.

Lynch, T. and Price, A. (2007). The effect of cytochrome P450 metabolism on drug response, interactions, and adverse effects. *American Family Physician*. 76(3): 391-396. PMID: 17708140.

Kampmann, T., Mueller, D.S., Mark, A.E., Young, P.R. and Kobe B. (2006). The Role of histidine residues in low-pH-mediated viral membrane fusion. *Structure*. 14(10): 1481-7. doi: 10.1016/j.str.2006.07.011.

Ma, Y., Liu, Y. and Cheng, J. (2018). Protein Secondary Structure Prediction Based on Data Partition and Semi-Random Subspace Method. *Scientific Reports*. 8(9856): 1-10. https://doi.org/10.1038/s41598-018-28084-8.

Madeira, F., mi Park, Y., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivery, A.R.N., Potter, S.C., Finn, R.D. and Lopez, R. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Research*. 47: W636-W641. doi: 10.1093/nar/gkz268.

Madian, A.G., Wheeler, H.E., Jones, R.B. and Dolan, M.E. (2012). Relating human genetic variation to variation in drug responses. *Trends in Genetics*. 28(10): 487-495. doi: 10.1016/j.tig.2012.06.008.

Marcath, L.A., Pasternak, A.L. and Hertz, D.L. (2019). Challenges to assess substrate-dependent allelic effects in CYP450 enzymes and the potential clinical implications. *Pharmacogenomics Journal*. 19(6): 501-515. doi: 10.1038/s41397-019-0105-1.

Maréchal, J. D., Kemp, C. A., Roberts, G. C., Paine, M. J., Wolf, C. R. and Sutcliffe, M. J. (2008). Insights into drug metabolism by cytochromes P450 from modelling studies of CYP2D6-drug interactions. *British Journal of Pharmacology*. 155(1): 582-589. doi: 10.1038/sj.bjp.0707570.

Márquez, A.Y.V., Briceño, I., Aristizábal, F., Niño, L.F. and Reyes, Y.J. (2019). Dynamic Effects of CYP2D6 Genetic Variants in a Set of Poor Metaboliser Patients with Infiltrating Ductal Cancer Under Treatment with Tamoxifen. *Scientific Reports*. 9(1): 1-12. doi: 10.1038/s41598-018-38340-6.

Martínez L. (2015). Automatic identification of mobile and rigid substructures in molecular dynamics simulations and fractional structural fluctuation analysis. *Plos one*. 10(3): 1-10. https://doi.org/10.1371/journal.pone.0119264.

McGibbon, R. T., Beauchamp, K. A., Harrigan, M. P., Klein, C., Swails, J. M., Hernández, C. X., Schwantes, C. R., Wang, L. P., Lane, T. J. and Pande, V. S. (2015). MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophysical Journal*. 109(8): 1528–1532. https://doi.org/10.1016/j.bpj.2015.08.015.

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thornman, A., Flicek, P. and Cunningham, Fiona. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*. 17(122): 1-14. doi: 10.1186/s13059-016-0974-4.

Miao, B., Xiao, Q., Chen, W., Li, Y. and Wang, Z. (2018). Evaluation of functionality for serine and threonine phosphorylation with different evolutionary ages in human and mouse. *BMC Genomics*. 19(1): 1-9. doi: 10.1186/s12864-018-4661-6.

Monte, A. A., West, K., McDaniel, K. T., Flaten, H. K., Saben, J., Shelton, S., Abdelmawla, F., Bushman, L. R., Williamson, K., Abbott, D. and Anderson, P. L. (2018). CYP2D6 Genotype Phenotype Discordance Due to Drug-Drug Interaction. *Clinical Pharmacology and Therapeutics*. 104(5): 933–939. doi: 10.1002/cpt.1135.

Nagy, G. and Oostenbrink, C. (2014). Dihedral-based segment identification and classification of biopolymers I: proteins. *Journal of Chemical Information and Modelling*. 54(1): 266–277. https://doi.org/10.1021/ci400541d.

Nagy, G. and Oostenbrink, C. (2012). Rationalization of stereospecific binding of propranolol to cytochrome P450 2D6 by free energy calculations. *European Biophysics Journal*. 41(12): 1065–1076. https://doi.org/10.1007/s00249-012-0865-x

Nguyen, H., Case, D. A. and Rose, A. S. (2018). NGLview-interactive molecular graphics for Jupyter notebooks. *Bioinformatics (Oxford, England)*. 34(7): 1241–1242. https://doi.org/10.1093/bioinformatics/btx789.

Oesch-Bartlomowicz, B. and Oesch, F. (2005). Phosphorylation of cytochromes P450: first discovery of a posttranslational modification of a drug-metabolizing enzyme. *Biochemical and Biophysical Research Communications*. 338(1): 446-9. doi: 10.1016/j.bbrc.2005.08.092.

Othman, H., da Rocha, J. and Hazelhurst, S. (2020). Variant Annotation of ADME genes using *SWAAT*: A Structural Bioinformatics Approach [version 1; not peer reviewed]. F1000Research 2020, 9(ISCB Comm J):812 (poster) https://doi.org/10.7490/f1000research.1118112.1.

Pace, C. N., Grimsley, G. R. and Scholtz, J. M. (2009). Protein ionizable groups: pK values and their contribution to protein stability and solubility. *The Journal of Biological Chemistry*. 284(20): 13285–13289. doi: 10.1074/jbc.R800080200.

Pace, C.N., Horn, G., Hebert, E.J., Bechert, J., Shaw, K., Urbanikova, L., Scholtz, J.M. and Sevcik, J (2001). Tyrosine hydrogen bonds make a large contribution to protein stability. *Journal of Molecular Biology*. 312(2): 393-404. doi: 10.1006/jmbi.2001.4956.

Paine, M.J., McLaughlin, L.A., Flanagan, J.U., Kemp, C.A., Sutcliffe, M.J., Roberts, G.C. and Wolf, C.R. (2003). Residues glutamate 216 and aspartate 301 are key determinants of

substrate specificity and product regioselectivity in cytochrome P450 2D6. *Journal of Biological Chemistry*. 278(6): 4021-4027. doi: 10.1074/jbc.M209519200.

Pandya, A., Howard, M.J., Zloh, M. and Dalby, P.A. (2018). An Evaluation of the Potential of NMR Spectroscopy and Computational Modelling Methods to Inform Biopharmaceutical Formulations. *Pharmaceutics*. 10(4):165. doi: 10.3390/pharmaceutics10040165.

Preissner, S.C., Hoffmann, M.F., Preissner, R., Dunkel, M., Gewiese, A. and Preissner, S. (2013). Polymorphic Cytochrome P450 Enzymes (CYPs) and Their Role in Personalised Medicine. *Plos one*. 8(12): 1-12. doi: 10.1371/journal.pone.0082562.

Rodrigues, C. H., Pires, D. E. and Ascher, D. B. (2018). DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. *Nucleic acids research*. 46(W1): W350–W355. https://doi.org/10.1093/nar/gky300.

Roe, D.R. and Cheatham, T.E. (2013). PYTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *Journal of Chemical Theory and Computation.* 9(7): 3084–3095. doi: 10.1021/ct400341p.

Sakuyama, K., Sasaki, T., Ujiie, S., Obata, K., Mizugaki, M., Ishikawa, M. and Hiratsuka, M. (2008). Functional characterization of 17 CYP2D6 allelic variants (CYP2D6.2, 10, 14A-B, 18, 27, 36, 39, 47-51, 53-55, and 57). *Drug Metabolism and Disposition*. 36(12): 2460-2467. doi: 10.1124/dmd.108.023242.

Salmaso, V. and Moror, S. (2018). Bridging molecular docking to molecular dynamics in exploring ligand-protein recognition process: An overview. *Frontiers in Pharmacology*. 9(923): 1-16. doi: 10.3389/fphar.2018.00923.

Salomon-Ferrer, R., Case, D.A. and Walker, R.C. (2012). An overview of the Amber biomolecular simulation package. *WIRES Computational Molecular Science*. 2012: 1-13. https://doi.org/10.1002/wcms.1121.

Santos, L. A., Prandi, I. G. and Ramalho, T. C. (2019). Could Quantum Mechanical Properties Be Reflected on Classical Molecular Dynamics? The Case of Halogenated Organic Compounds of Biological Interest. *Frontiers in Chemistry*. 7(848): 1-10. https://doi.org/10.3389/fchem.2019.00848.

Schrödinger, LLC. (2010). The PyMOL molecular graphics system, version 2.4.1.

Shahrokh, K., Orendt, A., Yost, G. S. and Cheatham, T. E., 3rd (2012). Quantum mechanically derived AMBER-compatible heme parameters for various states of the cytochrome P450 catalytic cycle. *Journal of Computational Chemistry*. 33(2): 119–133. doi: 10.1002/jcc.21922.

Silvino, A.C., Costa, G.L., Araújo, F.C., Ascher, D.B., Pires, D.E., Fontes, C.J., Carvalho, L.H., Brito, C.F. and Sousa, T.N. (2016). Variation in Human Cytochrome P-450 Drug-Metabolism Genes: A Gateway to the Understanding of Plasmodium vivax Relapses. *Plos One*. 13(2): 1-14. doi: 10.1371/journal.pone.0160172.

Sindelar, C. V., Hendsch, Z. S. and Tidor, B. (1998). Effects of salt bridges on protein structure and design. *Protein Science: A Publication of The Protein Society*. *7*(9): 1898–1914. https://doi.org/10.1002/pro.5560070906.

Skaric-Juric, T., Tomas, Z., Petranovic, M.Z., Bozina, N., Narancic, N.S., Janiccijevic, B. and Salihovic. (2018). Characterisation of ADME genes variation in Roma and 20 populations worldwide. *Plos One*. 13(11): 1-15. doi: 10.1371/journal.pone.0207671.

Šrejber, M., Navrátilová, V., Paloncýová, M., Bazgier, V., Berka, K., Anzenbacher, P. and Otyepka, M. (2018). Membrane-attached mammalian cytochromes P450: An overview of the membrane's effects on structure, drug binding, and interactions with redox partners. *Journal of Inorganic Biochemistry*. 183:117-136. doi: 10.1016/j.jinorgbio.2018.03.002.

Taylor, C., Crosby, I., Yip, V., Maguire, P., Pirmohamed, M. and Turner, R. M. (2020). A Review of the Important Role of *CYP2D6* in Pharmacogenomics. *Genes*. 11(1295): 1-22. https://doi.org/10.3390/genes11111295.

Terblanche, A. (2018) Pharmacovigilance and the reporting of adverse drug reactions. *South African Pharmaceutical Journal*. 85(6): 65-68. doi: 10.4103/0976-500X.142436.

Thanacoody, H. K. (2007). Thioridazine: resurrection as an antimicrobial agent?. *British Journal of Clinical Pharmacology*. 64(5), 566–574. doi: 10.1111/j.1365-2125.2007.03021.x.

The UniProt Consortium. (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*. 47(D1): D508-D515. doi: 10.1093/nar/gky1049.

Thibert, B., Bredesen, D. E., & del Rio, G. (2005). Improved prediction of critical residues for protein function based on network and phylogenetic analyses. *BMC Bioinformatics*. 6(213): 1-15. doi: 10.1186/1471-2105-6-213.

Tsuzuki, D., Hichiya, H., Okuda, Y., Yamamoto, S., Tamagake, K., Shinoda, S. and Narimatsu, S. (2003). Alteration in catalytic properties of human CYP2D6 caused by substitution of glycine-42 with arginine, lysine and glutamic acid. *Drug Metabolism and Pharmacokinetics*. 18(1): 79-85. doi: 10.2133/dmpk.18.79.

Tucci, S. and Akey, J.M. (2019). The long walk to African genomics. *Genome Biology*. 20(130): 1-3. doi: 10.1186/s13059-019-1740-1.

van Hooft, P., Greyling, B. J., Getz, W. M., van Helden, P. D., Zwaan, B. J. and Bastos, A. D. (2014). Positive selection of deleterious alleles through interaction with a sex-ratio suppressor gene in African Buffalo: a plausible new mechanism for a high frequency anomaly. *Plos one*. 9(11): e111778. https://doi.org/10.1371/journal.pone.0111778.

Vnučec, D., Kutnar, A. and Goršek, A. (2016). Soy-based adhesives for wood-bonding? a review. *Journal of Adhesion Science and Technology*. 31: 1-22. https://doi.org/10.1080/01694243.2016.1237278.

Vollmayr-Lee, K. (2020). Introduction to molecular dynamics simulations. *American Journal of Physics*. 88(5): 401-422. https://doi.org/10.1119/10.0000654.

Wang, A., Stout, C. D., Zhang, Q. and Johnson, E. F. (2015). Contributions of ionic interactions and protein dynamics to cytochrome P450 2D6 (CYP2D6) substrate and inhibitor binding. *The Journal of Biological Chemistry*. 290(8): 5092–5104. doi: 10.1074/jbc.M114.627661.

Wang, D., Poi, M.J., Sun, X., Gaedigk, A., Leeder, J.S. and Sadee, W. (2014). Common CYP2D6 polymorphisms affecting alternative splicing and transcription: long-range haplotypes with two regulatory variants modulate CYP2D6 activity. *Human Molecular Genetics*. 23(1): 268-78. doi: 10.1093/hmg/ddt417.

Wang, S., Yang, S., An, B., Wang, S., Yin, Y., Lu, Y., Xu, Y. and Hao, D. (2011). Molecular dynamics analysis reveals structural insights into mechanism of nicotine N-demethylation catalyzed by tobacco cytochrome P450 mono-oxygenase. *Plos one*. 6(8): 1-11. https://doi.org/10.1371/journal.pone.0023342.

Webb, B. and Sali, A. (2016). Comparative protein structure modelling using Modeller. *Current Protocols in Bioinformatics*. 54: 5.6.1-5.6.37. doi: 10.1002/cpbi.3.

Whirl-Carillo, M., McDonagh, E.M., Hebert, J.M., Gong, L., Sangkuhl, K., Thorn, C.F., Altman, R.B. and Klein, T.E. (2012). Pharmacogenomics knowledge for personalised medicine. *Clinical Pharmacology and Therapeutics*. 92(4): 414-417. doi: 10.1038/clpt.2012.96.

Wishart, D.S., Knox, C., Guo, A.C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z. and Woolsey, J. (2006). Drugbank: A comprehensive resource for *in silico* drug discovery and exploration. *Nucleic Acids Research*. 34: D668-672. doi: 10.1093/nar/gkj067.

Wright, G.E., Niehaus, D.J., Drögemöller, B.I., Koen, L., Gaedigk, A. and Warnich, L. (2010). Elucidation of CYP2D6 genetic diversity in a unique African population: implications for the future application of pharmacogenetics in the Xhosa population. *Annals of Human Genetics*. 74(4):340-350. doi: 10.1111/j.1469-1809.2010.00585.x.

Xie, N. Z., Du, Q. S., Li, J. X. and Huang, R. B. (2015). Exploring Strong Interactions in Proteins with Quantum Chemistry and Examples of Their Applications in Drug Design. *Plos one*. 10(9): e0137113. https://doi.org/10.1371/journal.pone.0137113.

Xin, J., Yuan, M., Peng, Y. and Wang, J. (2020). Analysis of the Deleterious Single-Nucleotide Polymorphisms Associated With Antidepressant Efficacy in Major Depressive Disorder. *Frontiers in Psychiatry*. 11(151): 1-11. doi: 10.3389/fpsyt.2020.00151.

Zanger, U.M., Momoi, K., Hofmann, U., Schwab, M. and Klein, K (2021). Tri-Allelic Haplotypes Determine and Differentiate Functionally Normal Allele CYP2D6*2 and Impaired Allele CYP2D6*41. *Clinical Pharmacology Therapeutics*. 109(5): 1256-1264. doi: 10.1002/cpt.2078.

Zhou, Y., Ingelman-Sundberg, M. and Lauschke, V. M. (2017). Worldwide Distribution of Cytochrome P450 Alleles: A Meta-analysis of Population-scale Sequencing Projects. *Clinical Pharmacology and Therapeutics*. 102(4): 688–700. doi: 10.1002/cpt.690.

Zia, A., Kiani, A., Bhatti, A. and John, P. (2013). Genetic Susceptibility to Type 2 Diabetes and Implications for Therapy. *Journal of Diabetes and Metabolism*. 4(3): 1-7. doi: 10.1146/annurev.med.59.090706.135315.

Appendices

Appendix A: Ethical Clearance Certificate

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG	HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

Office of the Deputy Vice-Chancellor (Research & Post Graduate Affairs)

TO: Mr BR Sitabule
School of Pathology
Depoartment of Human Genetics
Sydney Brenner Insitute for Molecular Bioscience
Medical School
University

E-mail: rotundwasitabule@gmail.com

CC: Supervisor: Dr H Othman and Professor S Hazelhurst
<houcemoo@gmail.com>
and <HREC-Medical.ResearchOffice@wits.ac.za>

FROM: Iain Burns
Human Research Ethics Committee (Medical)
Tel: 011 717 1252

E-mail: Iain.Burns@wits.ac.za

DATE: 2020/09/17

REF: *R14/49*

PROTOCOL NO: **M200711** *(This is your ethics application study reference number. Please quote this reference number in all correspondence relating to this study)*

PROJECT TITLE: *Structural bioinformatics analysis of CYP2D6 pharmacogenetic variation relevant to Sub-Saharan African populations*

Please find attached the Clearance Certificate for the above project. I hope it goes well and that an article in a recognized publication comes out of it. This will reflect well on your professional standing and contribute to the Government funding of the University.

MSWorks2000/Iain0007/Clearscan.wps

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

R14/49 Mr BR Sitabule

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)
CLEARANCE CERTIFICATE NO. M200711

NAME: **(Principal Investigator)**	Mr BR Sitabule
DEPARTMENT:	School of Pathology Depoartment of Human Genetics Sydney Brenner Insitute for Molecular Bioscience Medical School University
PROJECT TITLE:	Structural bioinformatics analysis of CYP2D6 pharmacogenetic variation relevant to Sub-Saharan African populations
DATE CONSIDERED:	2020/07/31
DECISION:	Approved unconditionally
CONDITIONS:	
SUPERVISOR:	Dr H Othman and Professor S Hazelhurst
APPROVED BY:	CB Penny Dr CB Penny, Chairperson, HREC (Medical)
DATE OF APPROVAL:	2020/09/17

This clearance certificate is valid for 5 years from the date of approval. Extension may be applied for.

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Research Office Secretary on the 3rd Floor, Phillip Tobias Building, Parktown, University of the Witwatersrand, Johannesburg.
I/we fully understand the conditions under which I am/we are authorized to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions. Should any departure be contemplated, from the research protocol as approved, I/we undertake to submit details to the Committee. **I agree to submit a yearly progress report**. When a funder requires annual re-certification, the application date will be one year after the date when the study was initially reviewed. In this case, the study was initially reviewed in **July** and will therefore reports and re-certification will be due early in the month of **July** each year. Unreported changes to the application may invalidate the clearance given by the HREC (Medical).

Principal Investigator Signature

09 Ocober 2020
Date

Appendix B: Minimisation plots

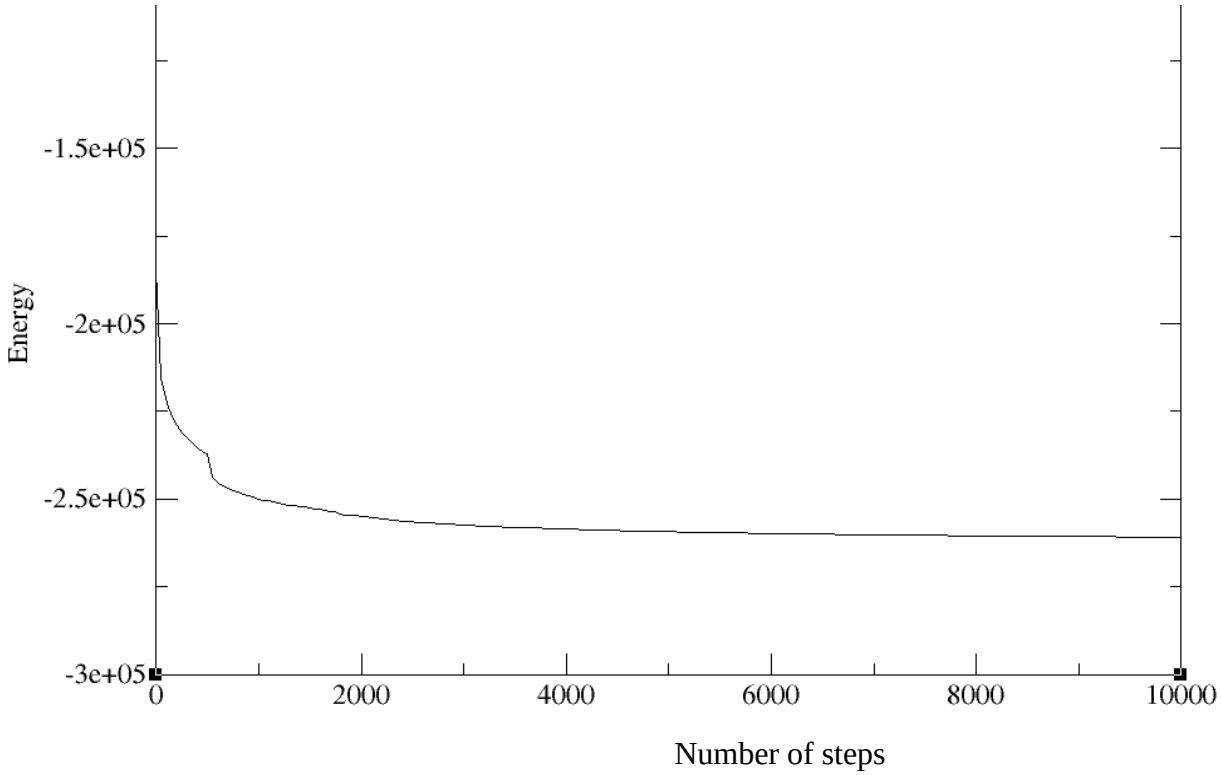


Figure 1: Minimisation plot showing the energy level in 20 000 steps for the first stage. The first 2 000 steps involved the steepest decent thus the energy decreases in a drastic manner. For the subsequent steps the conjugate gradient was applied, thus there is a gradual decrease and a plateau is reached.

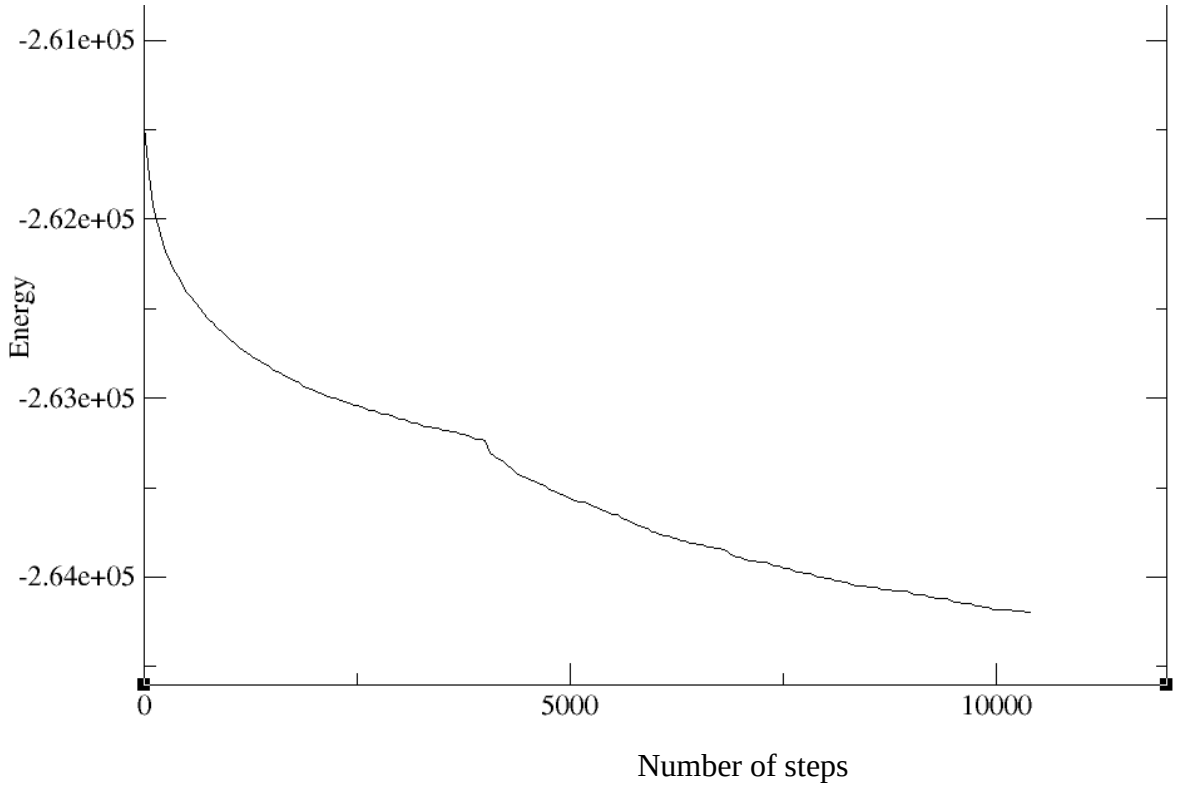


Figure 2: Minimisation plot showing the energy level in the 20 000 steps for the second stage. The first 4000 steps involved the steepest decent thus the energy decreases in a drastic manner. For the subsequent steps the conjugate gradient was applied, thus there is a gradual decrease approaching a plateau.

Appendix C: Heating and Equilibration

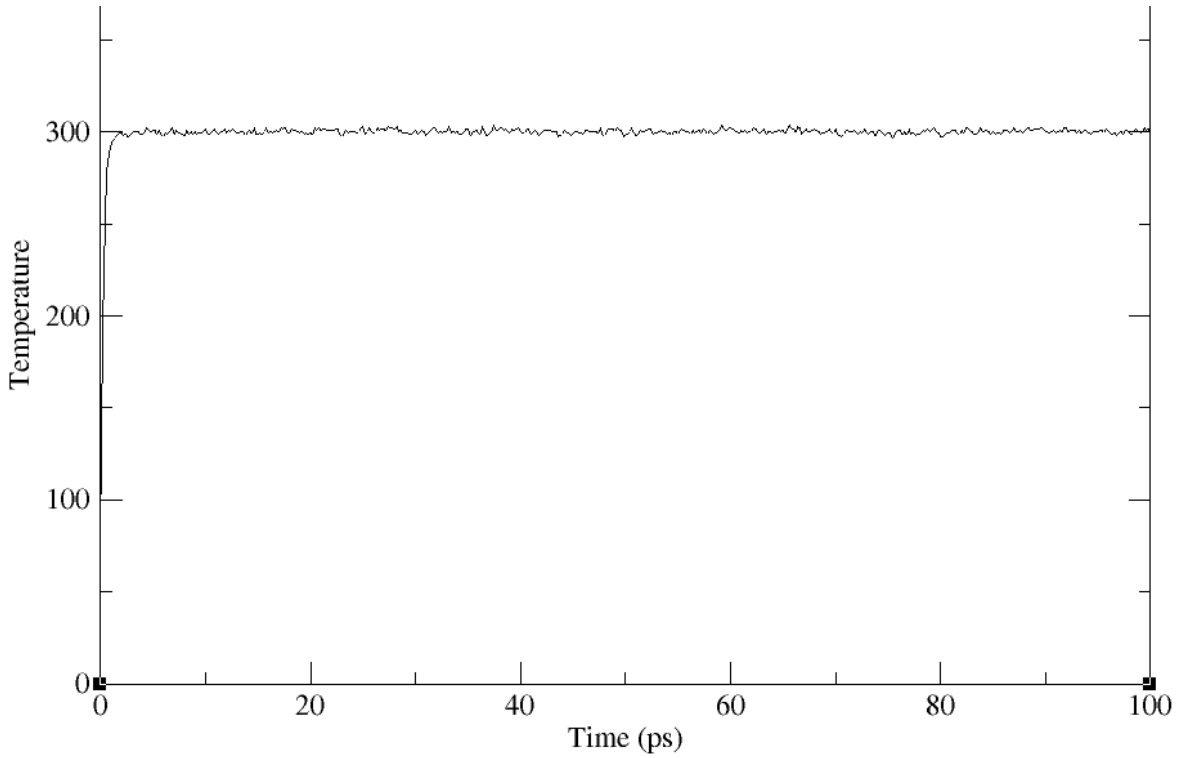


Figure 3: Plot showing the temperature during the heating and equilibration phase in 100 ps. The heating stage was done for the first 20 ps, which shows the increase in

temperature until it reached 300 K, the subsequent stage was the equilibration stage which is shown by the stabilised temperature.

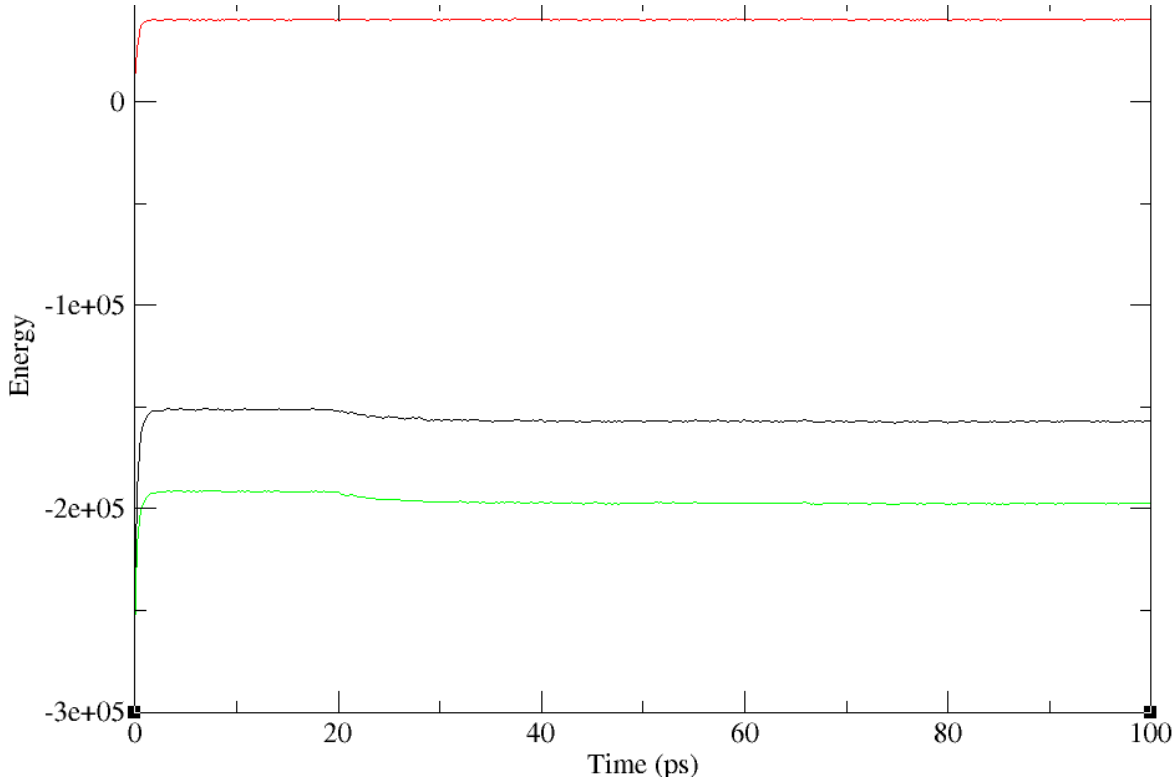


Figure 4: Plot showing the potential energy (red), kinetic energy (green) and the total energy (black) during the heating and equilibration. The energy increases as the temperature increased during the heating stage, then the potential energy stabilises, while the kinetic energy and total energy stabilise after the first few seconds followed by a decline during 20 ps and a final stabilisation from the two energies.

Appendix D: Table Showing Details of the 50 Variants Obtained from PharmVar and Their Star Alleles

rs_number	GRCh37 HGVS	Amino acid position	Amino acid change	*alleles harbouring the variant
rs1440526469	22:42522580 C>T	497	R/H	*139
rs568495591	22:42522607 G>A	488	S/F	*112
rs75467367	22:42522625 G>C	482	A/S	*36; *57; *83; *141
rs74478221	22:42522626 C>T	482	A/S	*36; *57; *83; *141
rs1135838	22:42522629 A>C	481	F/V	*36; *57; *83; *141
rs1135837	22:42522635 C>G	479	G/R	*36; *57; *83; *141
22_42522636_A/C	22:42522636 A>C	478	H/Q	*85
rs766507177	22:42522637	478	H/S	*36; *57; *83; *141

	T>G			
rs28371735	22:42522638 G>A	478	H/S	*36; *57; *83; *141
rs141756339	22:42522649 C>T	474	R/Q	*138
rs1135835	22:42522662 T>C	470	T/A	*36; *57; *83; *141
rs1135833	22:42522665 G>C	469	P/A	*36; *57; *83; *141
22_42522683_G/C	22:42522683 G>C	463	H/D	*98
22_42522699_G/T	22:42522699 G>T	457	F/L	*97
rs369177208	22:42522721 C>T	450	R/H	*125
rs751092905	22:42522737 C>T	445	G/R	*110
rs532668079	22:42522748 C>T	441	R/H	*75
rs730882251	22:42522749 G>A	441	R/C	*62
rs267608319	22:42522751 C>T	440	R/H	*31
rs569439709	22:42522754 C>T	439	G/D	*113
rs3021084	22:42522879 G>A	430	P/L	*136
rs28371733	22:42522916 C>T	418	E/K	*52; *106
rs747089665	22:42522928 G>A	414	R/C	*135
rs769157652	22:42522940 C>T	410	E/K	*27; *32; *141
22_42522958_T/G	22:42522958 T>G	404	K/Q	*55
rs77312092	22:42523459 C>T	388	R/H	*95
rs75386357	22:42523475 C>T	383	E/K	*72
rs61745683	22:42523514 C>T	370	V/I	*122
22_42523516_A/G	22:42523516 A>G	369	I/T	*26
rs1555888910	22:42523525 A>G	366	F/S	*105
rs1058172	22:42523528 C>T	365	R/H	*127; *139
rs202102799	22:42523558 T>C	355	Y/C	*108; *127
rs61736517	22:42523567 T>C	352	H/R	*108

rs76088846	22:42523591 C>T	344	R/Q	*134
rs267608295	22:42523595 G>C	343	R/G	*25
22_42523604_C/T	22:42523604 C>T	340	G/R	*133
rs59421388	22:42523610 C>T	338	V/M	*29; *70; *109
rs748712690	22:42523612 T>C	337	D/G	*94
rs78209835	22:42523613 C>T	337	D/N	*117
rs72549348	22:42523621 T>G	334	E/A	*51
rs141009491	22:42523633 C>G	330	R/P	*116
rs5030867	22:42523858 T>G	324	H/P	*7
22_42523901_T/C	22:42523901 T>C	310	T/A	*118
rs1406719554	22:42523924 A>G	302	L/P	*123
rs949717872	22:42523940 T>G	297	I/L	*24
rs1135829	22:42523975 T>C	285	N/S	*115; *132
rs1135828	22:42524183 A>T	279	M/K	*81; *86
rs77913725	22:42524187 C>T	278	E/K	*81; *86
rs148769737	22:42524219 G>T	267	P/H	*84
rs267608297	22:42524237 G>A	261	T/I	*54
22_42524274_T/G	22:42524274 T>G	249	T/P	*93
rs28371717	22:42524310 C>A	237	A/S	*33
rs17002853	22:42524327 A>G	231	L/P	*131
rs199535154	22:42524814 A>G	213	L/P	*20
rs745365204	22:42524850 C>T	201	R/H	*37
22_42524880_C/A	22:42524880 C>A	191	C/F	*130
22_42524881_A/G	22:42524881 A>G	191	C/F	*130
rs5030865	22:42525035 C>T	169	G/R	*14; *114
rs1135826	22:42525038 A>C	168	S/A	*121

rs1135825	22:42525039 G>T	167	H/Q	*121
rs1135824	22:42525044 T>C	166	N/D	*103; *121
22_42525058_C/G	22:42525058 C>G	161	C/S	*91
22_42525073_T/A	22:42525073 T>A	156	E/V	*104
rs267608302	22:42525073 T>G	156	E/A	*50
rs28371710	22:42525077 C>T	155	E/K	*45; *46
rs78482768	22:42525089 G>C	151	Q/E	*58
rs569229126	22:42525100 T>C	147	K/R	*90
rs375135093	22:42525115 A>G	142	L/S	*89
rs61736512	22:42525134 C>T	136	V/I	*29; *70
rs781457579	22:42525136 G>A	135	S/F	*126
rs1135823	22:42525176 C>A	122	A/S	*53
rs1135822	22:42525182 A>T	120	F/I	*49; *53
rs374616348	22:42525185 C>T	119	V/M	*70
rs535642512	22:42525761 C>T	111	G/S	*111
rs78459009	22:42525767 T>C	109	I/V	*82
rs28371706	22:42525772 G>A	107	T/I	*17; *40; *58; *64; *141
rs28371706	22:42525772 T>A	107	T/Y	*82
rs74802369	22:42525773 G>T	107	T/Y	*82
rs76187628	22:42525781 A>G	104	V/A	*82; *88
rs267608308	22:42525782 C>T	104	V/M	*73
rs28371704	22:42525811 T>C	94	H/R	*82
rs28371703	22:42525821 G>T	91	L/M	*74
rs267608309	22:42525823 G>A	90	A/V	*48; *102; *103
rs267608276	22:42525829 C>G	88	R/P	*99
rs267608310	22:42525838 G>A	85	A/V	*23

22_42525889_A/G	22:42525889 A>C	68	V/G	*128
rs267608311	22:42525908 G>A	62	R/W	*57
rs118203758	22:42526669 C>T	42	G/E	*71
rs5030862	22:42526670 C>T	42	G/R	*12
rs1065852	22:42526694 G>A	34	P/S	*10; *36; *37; *47; *49; *52; *54; *57; *64; *65; *69; *72; *87; *94; *95; *99; *100; *101; *114; *132; *142
rs138100349	22:42526712 G>A	28	R/C	*22; *44; *142
rs28371696	22:42526717 C>T	26	R/H	*43; *46
rs267608313	22:42526721 G>A	25	R/W	*47
rs769258	22:42526763 C>T	11	V/M	*35; *143
rs72549358	22:42526775 C>T	7	V/M	*28
rs773790593	22:42526780 G>A	5	A/V	*87

Appendix E: SWAAT results

Table 1: Non processed variants summary for *SWAAT* results

Reference Allele	Aternative allele	Reference residue	Residue position	Residue variant
G	A	R	28	C
C	T	R	26	H
C	T	V	11	M

Reference Allele	Aternative allele	Reference residue	Residue position	Residue variant
C	T	V	7	M

Table 2: Processed variants summary for *SWAAT* results

Reference Allele	Aternative allele	Reference residue	Residue position	Residue variant	dG (kcal/mol)	dS (kcal/mol/K)	ML prediction	Red Flags
G	A	P	34	S	4.575	0.447	2.0	⚑buried_exposed_switch, ⚑hotspotpatch
C	T	G	42	R	2.243	-0.012	2.0	⚑hotspotpatch
C	T	G	42	E	2.768	0.000	2.0	⚑hotspotpatch
C	G	R	88	P	4.920	0.136	2.0	

Reference Allele	Aternative allele	Reference residue	Residue position	Residue variant	dG (kcal/mol)	dS (kcal/mol/K)	ML prediction	Red Flags
G	T	L	91	M	-0.533	0.130	2.0	
T	C	H	94	R	-0.525	-0.046	0.0	
C	T	V	104	M	-0.796	-0.423	0.0	
A	G	V	104	A	1.685	0.101	1.0	
G	A	T	107	I	-0.803	-0.166	0.0	
C	T	V	119	M	-1.441	-0.423	2.0	
A	T	F	120	I	0.635	0.535	0.0	

Reference Allele	Aternative allele	Reference residue	Residue position	Residue variant	dG (kcal/mol)	dS (kcal/mol/K)	ML prediction	Red Flags
C	A	A	122	S	0.308	-0.319	0.0	
C	T	V	136	I	-0.236	-0.028	0.0	
G	C	Q	151	E	0.537	0.109	0.0	
C	T	E	155	K	-0.365	0.892	0.0	
T	C	N	166	D	1.007	0.099	0.0	
G	T	H	167	Q	-0.182	0.165	0.0	
A	C	S	168	A	0.611	0.000	0.0	

Reference Allele	Aternative allele	Reference residue	Residue position	Residue variant	dG (kcal/mol)	dS (kcal/mol/K)	ML prediction	Red Flags
C	T	R	201	H	0.240	0.060	0.0	
A	G	L	213	P	1.468	-0.092	2.0	
A	G	L	231	P	6.389	0.488	2.0	⚑buried_Pro_introduced, ⚑buried_exposed_switch ,⚑hotspotpatch
C	A	A	237	S	1.192	-0.079	2.0	
G	T	P	267	H	0.419	-0.318	0.0	
C	T	E	278	K	-0.435	-0.077	0.0	

Reference Allele	Aternative allele	Reference residue	Residue position	Residue variant	dG (kcal/mol)	dS (kcal/mol/K)	ML prediction	Red Flags
A	T	M	279	K	1.878	0.008	2.0	⚑Salt bridge formation
T	C	N	285	S	0.968	-0.002	0.0	
T	G	H	324	P	4.673	0.108	2.0	
C	G	R	330	P	3.289	0.791	2.0	
C	T	D	337	N	-0.031	1.094	0.0	
C	T	V	338	M	0.527	0.243	2.0	
C	T	R	344	Q	0.235	0.439	0.0	

Reference Allele	Aternative allele	Reference residue	Residue position	Residue variant	dG (kcal/mol)	dS (kcal/mol/K)	ML prediction	Red Flags
T	C	H	352	R	-0.232	-0.019	0.0	
T	C	Y	355	C	2.290	1.601	2.0	⚑hotspotpatch
C	T	R	365	H	8.525	0.132	2.0	⚑hotspotpatch
C	T	V	370	I	-0.431	-0.188	0.0	
C	T	E	383	K	-0.560	-0.033	0.0	
C	T	R	388	H	-0.204	0.181	0.0	⚑Salt bridge formation
C	T	E	410	K	-0.223	-0.060	0.0	

Reference Allele	Aternative allele	Reference residue	Residue position	Residue variant	dG (kcal/mol)	dS (kcal/mol/K)	ML prediction	Red Flags
C	T	E	418	K	-0.548	0.512	0.0	⚑Salt bridge breakage
C	T	R	440	H	-0.029	0.461	0.0	
G	A	R	441	C	-4.008	0.550	2.0	⚑exposed_hydrophilic_introduced
C	T	R	441	H	0.606	0.395	2.0	
C	T	R	450	H	0.234	-0.311	0.0	⚑Salt bridge formation
G	C	P	469	A	2.020	0.519	2.0	⚑buried_exposed_switch ,⚑hotspotpatch
T	C	T	470	A	-0.292	0.000	0.0	

Reference Allele	Aternative allele	Reference residue	Residue position	Residue variant	dG (kcal/mol)	dS (kcal/mol/K)	ML prediction	Red Flags
C	T	R	474	Q	0.477	0.037	0.0	

Appendix F: Plagiarism Documents

PLAGIARISM DECLARATION TO BE SIGNED BY ALL HIGHER DEGREE STUDENTS

SENATE PLAGIARISM POLICY: APPENDIX ONE

I Blessing Rotondwa Sitabule (Student number: 1439665) am a studentregistered for the degree of Master of Science in Medicine (Human Genetics) in the academic year 2021 .

I hereby declare the following:

- I am aware that plagiarism (the use of someone else's work without their permission and/orwithout acknowledging the original source) is wrong.
- I confirm that the work submitted for assessment for the above degree is my own unaided workexcept where I have explicitly indicated otherwise.
- I have followed the required conventions in referencing the thoughts and ideas of others.
- I understand that the University of the Witwatersrand may take disciplinary action against me if there is a belief that this is not my own unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.
- I have included as an appendix a report from "Turnitin" (or other approved plagiarism detection) software indicating the level of plagiarism in my research document.

Signature: Date: 11 April 2022

Corrected_MSc_Dissertation_1439665_Blessing_Sitabule-4.pdf

ORIGINALITY REPORT

12%	9%	7%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	hdl.handle.net Internet Source	1%
2	www.iscb.org Internet Source	1%
3	www.tandfonline.com Internet Source	<1%
4	ascpt.onlinelibrary.wiley.com Internet Source	<1%
5	link.springer.com Internet Source	<1%
6	www.biorxiv.org Internet Source	<1%
7	Submitted to Van Hal Larenstein (VHL) Student Paper	<1%
8	www.science.gov Internet Source	<1%
9	www.frontiersin.org Internet Source	<1%