The Effect of Afrocentric Missense variations on the Structural Dynamics of *CYP2B6*

SHAYLYN ASHLEY GOVENDER (G19G2631)

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

of

Master of Science in Bioinformatics and Computational Molecular Biology

(Coursework and Thesis)

of

Rhodes University, South Africa

Research Unit in Bioinformatics (RUBi)

Department of Biochemistry, Microbiology and Bioinformatics

Faculty of Science

Supervisor: Professor Özlem Tastan Bishop

Co-supervisor: Dr Allan Sanyanga

December 2024

ABSTRACT

Cytochrome P450s are a superfamily of enzymes with over 50 members involved in metabolizing 90% of xenobiotics. Among the these, families 1, 2, and 3 are responsible for approximately 80% of clinical drug metabolism. This study investigates the effect of Afrocentric missense variants on the structural dynamics of *CYP2B6*. Molecular dynamic simulations reveal that specific variants affect the enzyme's flexibility and stability, potentially altering catalytic activity and drug binding properties. These findings highlight the importance of considering genetic variants in personalized medicine and drug development. By investigating *CYP2B6*'s function and structural changes induced by missense variants, this research advances our understanding of the enzyme's role in drug metabolism. The study utilized computational tools such as GROMACS and AMBER for pre- and post-simulation analysis, with clustering and DSSP used to assess protein structures. Variants *I328T*, *K282R*, *P428T* and *R140Q* exhibited significant deviations in enzyme dynamics, while other variants caused minor shifts. Overall, the findings provide insight into the relationship between genetic variants and enzyme function, contributing to bioinformatics and molecular modelling approaches in drug discovery. Future studies could explore the structural and fuctional impacts of CYP2B6 bound to substrates such as antimalarials, expanding the investigation to a broader range of missense variants.

Keywords: *CYP2B6*, variations, structural dynamics, computational, MD, metabolism

DECLARATION

As part of my MSc degree in Bioinformatics, I, **Shaylyn Ashley Govender**, humbly attest that the dissertation I have submitted to Rhodes University is the result of my own independent intellectual work. I declare that this work is completely unique and hasn't been submitted or presented to any other educational institution before. I adequately acknowledge all outside sources and ideas that are not my own, using references and in-text citations when appropriate.

Signature: ***Shaylyn Govender***
Date: 19/11/24

ACKNOWLEDGEMENTS

First, I want to acknowledge my supervisor, Professor Özlem Tastan Bishop, who continuously guided and supported me throughout my MSc study. Her dedication and professional attitude to work always inspired me to work harder. Moreover, I want to thank her for all the professional skills that I have learned from her during my MSc. An even bigger thanks must be said to her for not giving up on me.

I want to thank my co-supervisor, Dr. Allan Sanyanga for his continuous support and guidance throughout the time of my MSc. His contribution towards the experimentation throughout, this includes but is not limited to, metal parameterisation, scripts for clustering and scripts which could be used within PyMOL.

Acknowledgements made towards all the peers within the RUBi group for their support towards the work and assistance of scripts that were provided for data analysis within the paper. I thank the past and present lab members for all the questions and the constructive criticism that I have received from them during our group meetings. A special thank you to Rehema and Chiratidzo Chamboko for supplying material for the homology modelling and DSSP scripts, respectively. The rest of the MSc students, thank you for your endless time and help throughout the year, special mentions to Queenie Mondile and Chipo Makombe for their assistance. Lastly a very special thank you towards the GRADIENT team for their unconditional support and the funding that was provided. The team has offered many insights into the work that was done and helped substantially.

A massive thank you to the Centre for High Performance Computing (CHPC) as they provided substantial support and resources that allowed for this project to be completed.

DEDICATION

I humbly dedicate this work to my Family.

My mom for always being there and providing her unconditional love and support.
To my dad and stepmom for always believing in me and encouraging me to chase my dream.
My brother, your constant reminder to work hard.
To Robyn Hendricks. Your love and endless nights of encouragements.
Lastly, the Jattiem family for their continuous support

I express my humble gratitude to all of you, your endless words of encouragement and love has motivated me. Your support will always be remembered and appreciated.

'*For I know the plans I have for you," declares the Lord, "plans to prosper you and not to harm you, plans to give you hope and a future.*'

Jeremiah 29:11

TABLE OF CONTENTS

TABLE OF FIGURES

LIST OF TABLES

LIST OF WEB SERVERS

H++	http://biophysics.cs.vt.edu
RCSB	https://www.rcsb.org
UniProt	https://www.uniprot.org/

LIST OF ABBREVIATIONS

ADR	Adverse drug reactions
AMBER	Assisted Model Building and Energy Refinement
BG	Bergamottin
BPA	Tert-butylphenylacetylene
CAR	Constitutive androstane receptor
CHPC	Centre for High Performance Computing
Cpd 0	Compound 0
Cpd I	Compound I
CXR	Constitutive androstane receptor
CYP	Cytochrome P450
CYP2B	Cytochrome P450 family 2 subfamily B
CYP2B1	Cytochrome P450 family 2 subfamily B member 1
CYP2B4	Cytochrome P450 family 2 subfamily B member 4
CYP2B6	Cytochrome P450 family 2 subfamily B member 6
DDI	Drug-drug interactions
DDGI	Drug-drug-gene interaction
DGI	Drug-Gene interaction
DOPE	Discrete Optimised Protein Energy
DSSP	Define Secondary Structure of Proteins
FASTA	FAST-Alignment
FeIV	Ferryl
GPU	Graphical processing units
GROMACS	Groningen Machine for Chemical Simulation
MD	Molecular Dynamics
MTS	Multiple timescale
MUSCLE	MUltiple Sequence Comparison by Log- Expectation
NMR	Nuclear magnetic resonance
NPT	Constant pressure
NVT	Constant volume
ns	nanoseconds
O_2	molecular oxygen
PharmVar	Pharmacogene Variation
PDB	Protein Data Bank
PME	Particle Mesh Ewald
PXR	Pregnane × receptor
RCSB	RCSB Protein Data Bank
Rg	Radius of Gyration
RMSD	Root Mean Square Deviation
RMSF	Root-mean-square fluctuations

RUBI	Rhodes University Bioinformatics Unit
SNP	Single nucleotide polymorphism
TM	Transmembrane
UDP	Uridine diphosphate
UNIPROT	Universal Protein Resource
3D	Three-dimensional
7-EFC	7-ethoxy-4-trifluoromethylcoumarin

THESIS OVERVIEW

Chapter 1

This chapter introduces the protein of interest, *CYP2B6*. The structure, function and mechanism of action is explained within this chapter including the catalytic cycle that occurs. In addition to this, the substrate specificity of the enzyme and the involvement in drug metabolism is explained. The overall aim and objectives are explained and detailed within the chapter.

Chapter 2

Within this chapter, an overview of what homology modelling is provided and how 3D structures are generated. In addition to this, is the understanding of genetic variations and the potential impact they have on the protein of interest. Methodologies of how models were produced, and the criteria used to select allele variations are explained. The chapter explains how various literature sources as well as databases containing genetic variations give insights into selecting variations for *CYP2B6* with prime focus in the African population.

Chapter 3

In this chapter, the understanding and application of MD simulations is explored. The principle is then applied to that of the *CYP2B6* enzyme, and the variants associated within this study. This is done to investigate the structural impact the variations have on the *CYP2B6* enzyme. Included in this chapter is the brief description of the major software packages and their use in this study. The method for each step is incorporated and the results display how the trajectories were analysed, processed, and visually represented.

Chapter 4

Analysis conducted in Chapter 3 gives the results and interpretations that can be understood on a global scale with regards to the impact each variation has on the structure of the *CYP2B6* enzyme. To move a step further and pay attention to the small changes in a detailed image, methods such as clustering and determining secondary structures is implemented. In this chapter, the different methods will be explored to assist in understanding if any differences occur in the *CYP2B6* structure when magnified to certain sections of the protein.

1 Chapter 1: Literature Review

1.1 Introduction

The healthcare sector is rapidly advancing, with precision medicine emerging as a transformative paradigm that seeks to individualise medical treatments by recognizing the profound impact of genetic factors on responses to medications [1]. In the context, Cytochrome P450 (CYP) enzymes play a crucial role in drug metabolism, serving as key enzymes that determine how drugs are metabolized in the body. Understanding these enzymes are essential for optimizing therapeutic outcomes, particularly through pharmacogenomics, which explores the relationship between an individual's genetic makeup and their response to drugs [2]. This field emphasizes the significance of tailored treatment plans to achieve optimal results.

Pharmacogenomic insights reveal that genetic variations significantly contribute to the diversity of drug responses among individuals. By identifying specific genetic markers, healthcare providers can predict how patients will respond to medications, enabling customized treatment plans [3]. Pharmacogenomic testing is particularly valuable in identifying adverse drug reactions (ADRs) stemming from genetic predispositions, allowing healthcare providers to make informed decisions and enhance patient safety [4]. Precision medicine mandates the selection of the most effective and least harmful treatment for each individual. Pharmacogenomics plays a pivotal role in identifying drugs likely to be most effective based on an individual's genetic profile, minimising the trial-and-error approach to finding suitable medications [5]. Understanding how genetic variations influence drug metabolism and efficacy is important for optimising therapeutic outcomes. Pharmacogenomic information facilitates the adjustment of drug dosages, or the selection of alternative medications based on individual genetic characteristics to ensure maximum effectiveness [6].

In chronic diseases like cancer, long-term treatment success relies on the identification of effective drug combinations. Pharmacogenomics guides the selection of optimal treatments, minimising the risk of drug resistance and enhancing the prospects of long-term treatment success [7]. It contributes to predictive medicine by identifying individuals at risk of not responding well to

specific drugs. This proactive approach allows for anticipating challenges and adjusting treatment plans before adverse effects occur, optimising patient care [8]. The integration of pharmacogenomics into precision medicine holds the potential for economic benefits by avoiding unnecessary treatments, reducing hospital admissions due to adverse drug reactions, and mitigating the financial burden associated with ineffective medications [9].

The thesis focuses on the critical role of CYP enzymes, with a specialized emphasis on CYP2B6. The CYP superfamily, particularly *CYP2B6*, is fundamental in processing various drugs, including antimalarial agents [10,11]. Given the global health challenge posed by malaria, exploring CYP2B6 is urgent and underscores the need to optimize treatment strategies.

Genetic variations within the *CYP2B6* gene introduce interindividual variability in drug metabolism, alongside other factors like variability in mRNA expression and protein activity levels in the human liver [10,12,13]. Advanced methodologies, such as molecular dynamics (MD) simulations, are becoming instrumental in providing nuanced insights into the structural dynamics, drug interactions, and functional consequences of CYP allele variations, particularly in antimalarial drug processing.

1.1.1 Cytochrome P450s

CYPs are essential enzymes in biological processes [14]. Initially discovered by Osamu Hayaishi in 1957, CYP enzymes belong to the oxygenase family, catalysing the insertion of oxygen atoms into carbon-hydrogen bonds [13]. Within this family, CYPs are monooxygenases crucial for hydroxylation across a wide range of substrates [15]. These enzymes play a pivotal role in the metabolism of a wide array of endogenous and exogenous compounds [16].

Drug metabolism is a complex process involving two crucial phases, illustrated in Figure 1.1 below, phase I and phase II [11]. The diagram underscores the pivotal role of CYP enzymes in Phase I drug metabolism. Various reactions occur within the initial phase, including the incorporation of reactive or polar groups (-OH, -COOH, $-NH_2$, -SH, etc.) into drugs through processes such as oxidation, reduction, and hydrolysis [11]. These modifications are essential to ensure that drugs can be effectively eliminated from the body. Phase II reactions, where modified drugs undergo conjugation with polar compounds. This conjugation renders drugs more

hydrophilic, facilitating their elimination from the body. The process involves a range of transferase enzymes, among them are: uridine diphosphate (UDP)-glucuronosyltransferases, sulfotransferases, and glutathione S-transferases [11,17]. Through this conjugation, the drugs undergo a further refinement that prepares them for recognition by efflux transporters within cells. Once identified, these compounds are subsequently expelled from the cells as part of the body's detoxification mechanism. It is important to note that this metabolic pathway, while crucial for drug elimination, also poses potential risks such as the creation of reactive metabolites [11]. The bioactivation of drugs and its occurrence is a phenomenon that hinges on specific structural features. Therefore, understanding and managing the intricacies of drug metabolism is essential to optimise therapeutic outcomes while mitigating potential adverse effects.

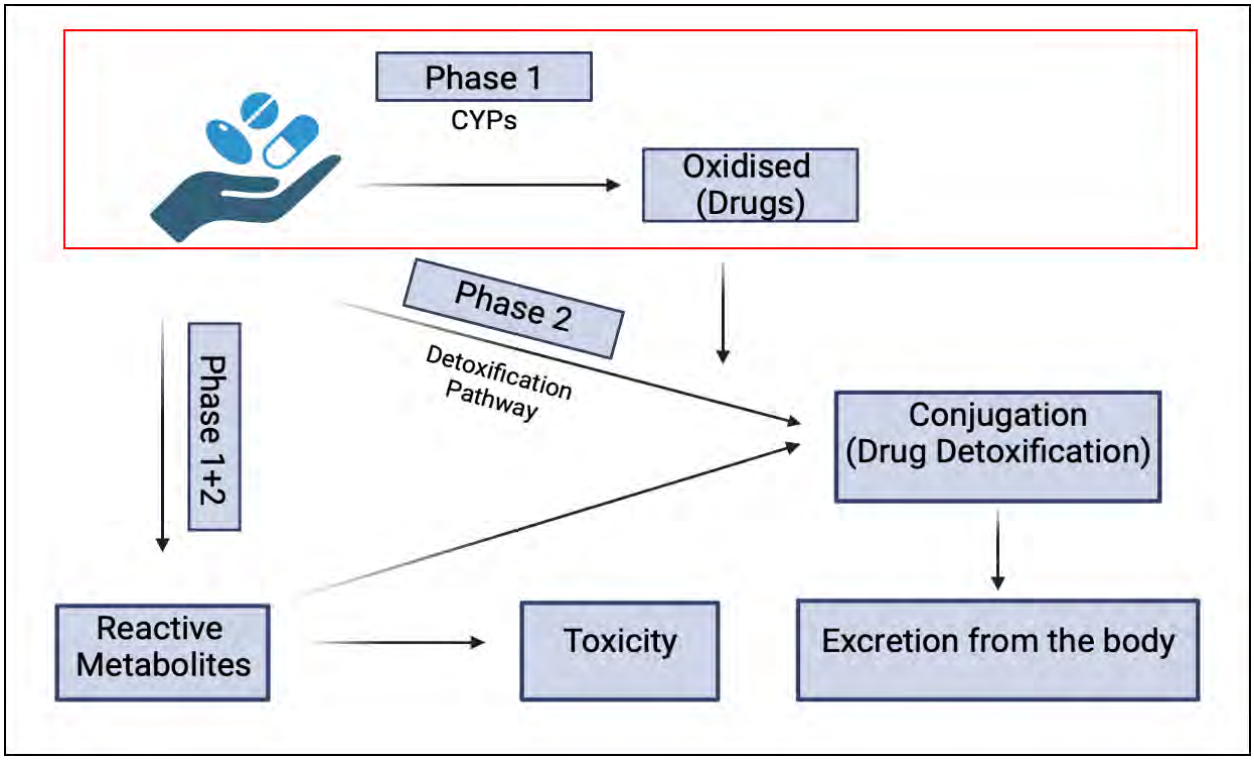


Figure 1.1: General pathways of drug metabolism. CYPs are involved in Phase I metabolism, the main process of CYPs is outlined in red.

CYP enzymes mediate over 90% of drug metabolism reactions, with significant expression in the liver and some presence in the kidney, intestines, and other organs[11,18] The CYP1, CYP2, and CYP3 families account for the metabolism of approximately 80% of clinical drugs, contributing

to individual variability in drug responses due to genetic, epigenetic, and environmental factors like age, diet, and disease states [11,19]. Moreover,
CYPs are further influenced by co-administered drugs or metabolites, leading to drug–drug interactions (DDIs), drug–gene interactions (DGIs), and drug–drug–gene interactions (DDGIs). [11,20]

CYP enzymes belong to the superfamily of heme-thiolate proteins and are widely distributed across all kingdoms of life [21,22,23,24]. The basic structural unit of a CYP is a monomer, typically composed of around 500 amino acids arranged in a specific sequence that determines the enzyme's three-dimensional structure and function[25]. The core of CYPs includes a heme cofactor. The heme is a complex of iron and a porphyrin ring which plays a crucial role in the enzyme's catalytic activity [26]. The heme cofactor is coordinated to the protein through a cysteine residue. This interaction involves a conserved thiolate ligand, forming the active site of the enzyme where the substrate molecule (the molecule upon which the enzyme acts) binds and where the chemical reaction takes place [22,23,25]. The interaction between the heme cofactor and the protein (see Section 1.1.2) involves a cysteine residue with a thiol group (—SH) that forms a thiolate ligand in CYPs. [22,25]. The thiolate ligand is crucial for the catalytic function of the enzyme as it actively participates in the redox reactions that CYP enzymes facilitate [25,26,27]. The enzyme's heme-thiolate structure enables it to serve as a versatile catalyst in various biochemical reactions, especially oxidation reactions. [22,24].

Among the various CYP enzymes, *CYP2B6* holds unique relevance due to its significant genetic variability, especially in populations of African descent, where polymorphisms are more prevalent. This genetic diversity influences how individuals metabolize drugs such as efavirenz and bupropion, impacting therapeutic outcomes and the risk of adverse effects. As a result, *CYP2B6* is particularly important for pharmacogenomic studies and precision medicine efforts aimed at optimizing drug efficacy and safety across diverse populations. In the upcoming sections, we will explore the distinct features and contributions of CYP2B6 within the expansive field of CYP biology, focusing on its role in drug metabolism and implications for personalized medicine.

1.1.2 Catalytic Cycle

Monooxygenases of the CYP are key catalysts that mediate the oxidation of sp3 and sp2 C-H bonds [28], playing a critical role in the hepatic breakdown of diverse compounds. These enzymes mediate over 95% of documented reactions in the literature, underscoring their significance in both oxidative and reductive transformations [29]. The catalytic cycle of CYP enzymes follows a precise sequence, illustrating their complex function [30]. This cycle is typically described as a series of chemical reactions in the form of a loop

Figure 1.2 below describes the several steps that are involved in the CYP catalytic cycle. Before the cycle can commence, the enzyme is characterised by a low-spin ferric configuration referred to as the resting state, where the heme iron is coordinated to a water molecule [25,31]. Step 1: The substrate binds with the resting state enzyme which leads to a transformative process where the substrate triggers the detachment of the aqua ligand as water molecules leave the pocket, inducing a shift to a high-spin Fe^{3+}-heme complex [32]. Step 2: Involves electron transfer facilitated by a more positive redox potential [25,32,33], reducing the heme to a ferrous Fe^{2+} complex. Step 3: Molecular oxygen (O_2) binding to this ferrous complex leads to the formation of an oxyferrous complex, this oxygen activation step is crucial in preparing the enzyme for further transformations [25,34]. Step 4: The oxyferrous complex undergoes reduction to form a peroxo complex. Step 5: Water molecules, which had temporarily left the active site, return to create a water channel [35]. Step 6: Protonation of the peroxo complex leads to the formation of Compound 0 (Cpd 0) and Compound I (Cpd I), a ferryl (FeIV)-oxo-π porphyrin cation radical [25],[35]. These steps represent the formation of intermediates in the catalytic cycle. Step 7: The activated Cpd I abstracts a hydrogen atom from the substrate, initiating its oxidation [36]. Upon successful completion of the catalytic reaction, the product is released from the active site, and a water molecule takes its place. To complete the catalytic cycle, the enzyme regenerates to its resting state. This regeneration is achieved by releasing the product and accepting another electron from NADPH-cytochrome P450 reductase.

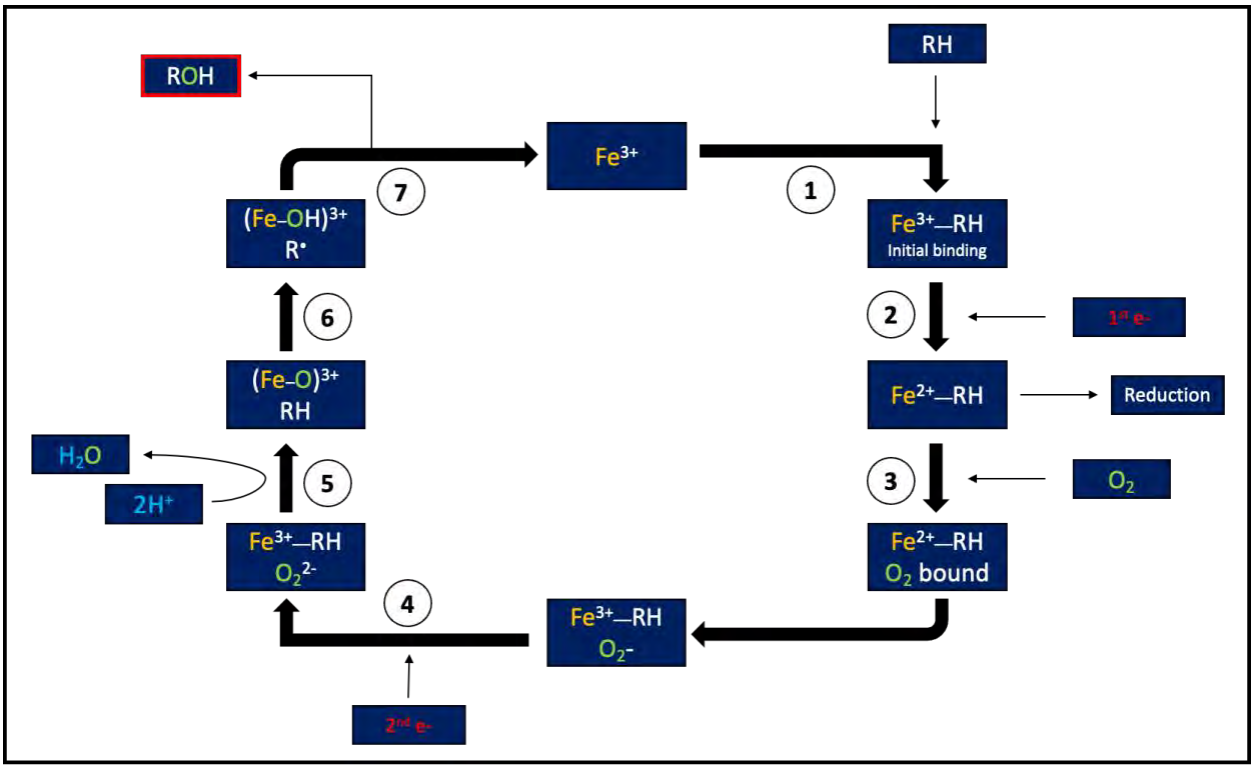


Figure 1.2: Seven step catalytic cycle demonstrating the simplified mechanism of CYP enzymes. Each step of the cycle is described within the text, indicated by circled black numbers. The substrate, hydrogens, electrons, and oxygen are represented by white, blue, red, and green, respectively. The iron is represented in orange. The released product is outlined in a red box. Adapted from Hart et al., (2008).

These enzymatic reactions are important to biological significance, as they contribute to essential cellular processes [37]. There are several factors that regulate the catalytic activity of CYP enzymes, including substrate concentration, cofactor availability, and feedback inhibition [37,38]. Understanding these factors that are part of regulatory mechanisms is important for modulating enzymatic activity for specific physiological needs or therapeutic interventions.

1.2 CYP2B6 Enzyme

1.2.1 Structure and function of CYP2B6.

The enzyme *CYP2B6 is* located on chromosome 19q13.2. and plays a crucial role in the metabolism of a wide range of endogenous and exogenous compounds [16, 39]. Its complex structure and versatile functions make it a subject of significant interest in the fields of pharmacology, toxicology, and drug development. This section will delve into the structural aspects of *CYP2B6*, exploring its overall architecture, active sites, substrate specificity, and the implications of its polymorphic nature [40,41].

The proportion of *CYP2B6* in relation to all CYPs ranges between 1-6% which was identified through antibody-based expression analysis [42]. A previous study performed an investigation to identify the substrates that were associated with *CYP2B6* [43]. The study included over 1300 substances associated with CYP systems in which *CYP2B6* played a role in 7.6% of drug metabolism [43]. This highly polymorphic gene has at least 38 allelic variants of which 25 are considered important; of that, it was found that 8 are common within a single racial or ethnic population [44]. *CYP2B6* is a heme-containing monooxygenase located in the endoplasmic reticulum of liver cells, primarily hepatocytes however can also be found in the brain, lungs and intestines which are extrahepatic tissues [12,39]. Its three-dimensional (3D) structure is organised into several domains, each contributing to its catalytic activity and substrate specificity [40], as illustrated in Figure 1.3. The enzyme consists of a core structure with a heme prosthetic group nestled in a hydrophobic pocket, crucial for its catalytic function [11,25] The heme group undergoes redox cycling, transitioning between the ferric and ferrous states during the catalytic cycle, facilitating the oxygenation of substrates [25,31,32].

Throughout the CYP superfamily, the 3D structure is considered to be moderately conserved [45,46]. Generally, CYP enzymes' shape consists of 14 major helices and loops [47], with some CYP enzymes having extra or less helices and loops [48]. *CYP2B6* consists of 12 major α-helices and loops denoted from A-L (Figure 1.3A and 1.3C) and includes 4 major Beta (β) sheets. There are three highly conserved regions that have been identified shedding light on the structural intricacies of the protein which are explained by Chamboko *et al.,* [25]. Helices A and B are situated in the N-terminal region and are key contributors to the overall protein folding and stability of *CYP2B6*. Though not located at the N-terminal, helix C also assists helices A and B with maintaining the protein structure and stability.

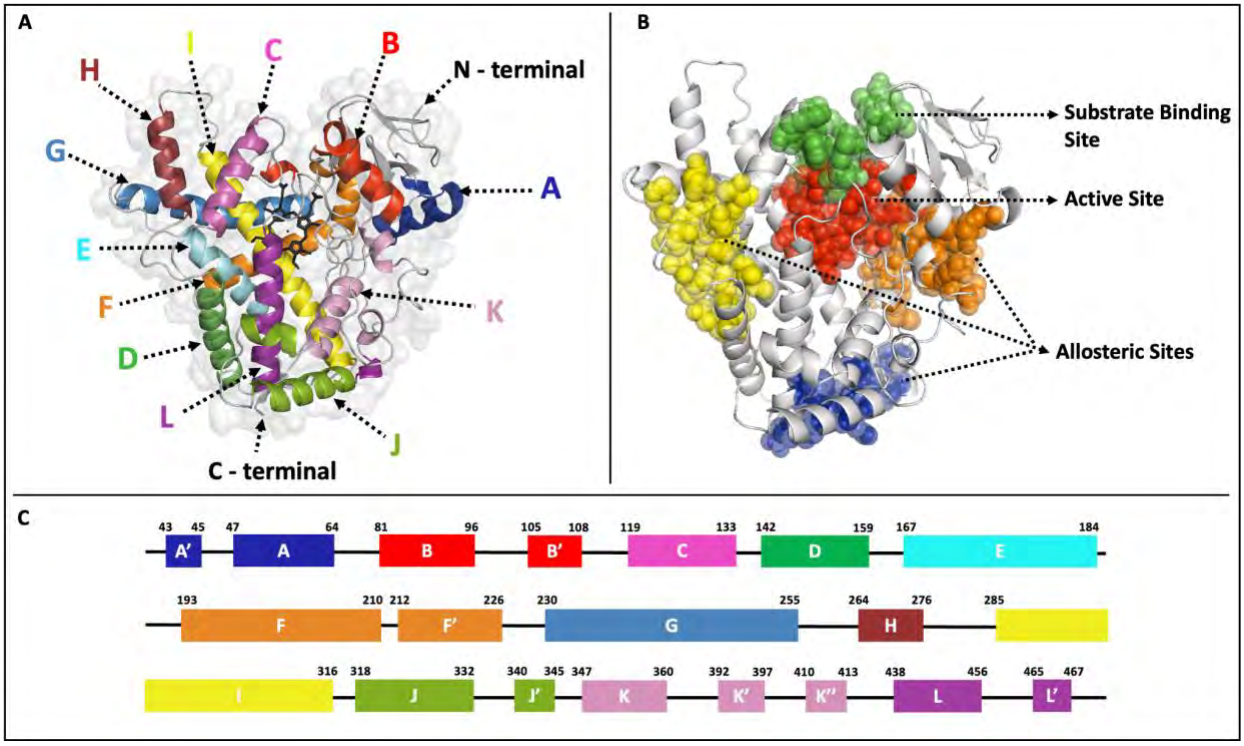


Figure 1.3: *CYP2B6* information of structural properties and sites. (A) Structure of CYP2B6 Enzyme (PDB ID: 5UFG). Major helices are denoted A-L in different colours. N- and C-terminals are labelled. **(B)** The substrate binding site, active site and allosteric sites/ binding sites are shown on the structure of CYP2B6 in green, red, (orange, yellow and blue), respectively. **(C)** Enzyme residue numbers with blocks representing the helices and the corresponding colours to the structure of CYP2B6

Membranes play a fundamental role in cellular organisation and compartmentalization, providing structural support and facilitating communication between different cellular compartments. Within this context, helix A, also known as the transmembrane (TM) helix serves to anchor *CYP2B6* to the membrane, ensuring its proper localisation and functionality [49]. Helix F' assists helix A with the anchoring process. Aside from the 12 major helices, an important loop is the BC loop, which serves the role for substrate specificity. Conformational changes that occur within this region can influence the shape and size of the active site which affects the enzyme's ability to bind and metabolise specific substrates [49]. This particular loop is regarded as a flexible region which leads it to having a dynamic nature which is essential for ligand entry and exit, facilitating substrates during the catalytic cycle [49]. Helices F and G are essential for substrate recognition and binding,

contributing significantly to the formation of the active site pocket. Assisting with the formation of this pocket are helices D and E, where the pocket is responsible for substrate accommodation during catalysis [50,51]. Together the BC loop and FC loop form the top (roof) and bottom (floor) of the active site, respectively [25], regulating access and contributing to the overall structural adaptations for substrate metabolism. Despite potential sequence diversity, the BC loop is a region of interest for understanding the catalytic processes and substrate interactions of CYP enzymes due to its flexibility and role in substrate binding.

Within the active site, the heme iron centre of *CYP2B6* (Figure 1.3A) in black serves as the site for oxygen binding and activation [25,26,31]. The active site also contains amino acid residues that form hydrogen bonds and hydrophobic interactions with substrates [22], contributing to the enzyme's specificity [51]. The flexibility of the active site allows *CYP2B6* to accommodate a diverse array of substrates, ranging from drugs and environmental toxins to endogenous compounds [52]. There are several residues within the active site and the substrate binding site that are of particular interest as they play a critical role with regards to the enzyme's function, the table below displays which residue numbers belong to the active site, substrate binding site and the allosteric sites/binding sites (Table 1.1).

Table 1.1: Sites of interest in CYP2B6 - with each sites corresponding residue numbers and colour represented in Figure 1.3.

Sites	**Residue Numbers**	**Colour**
Active Site	101,104,108,114,115,206,209,238,294,297,298, 301,302,363,365,366,367,368,389,477,492	Red
Substrate Binding Site	93,96,97,98,99,100,105,108,109,110,115,116,117,118,121,122,370,371,372,385,433,434	Green
Allosteric Site 1	130,131,134,135,136,139,140,145,148,149,177,178,181,182,185,187,261,262,263,264,267	Yellow
Allosteric Site 2	49,50,51,52,53,56,59,80,210,211,212,215,359,361,364,393,394,397,398,399,472,473,474,475,476,479	Orange
Allosteric site 3	314,315,318,321,322,325,454,458,459,460,461,463,465,468,491	Blue

Thoroughly examined by Shah et al. (2011) [55] and other contributors, the importance of the active site for substrate recognition and catalysis was displayed [56-59]. The residues of particular interest include: Phe206, Phe297, Val367, Thr205 and Thr302. The orientation of the side chain of flexible residues Phe206 and Phe297 were shown to be important for accommodation of inhibitors such as 4-(4-chlorophenyl)imidazole and 4-benzylpyridine [58]. This flexibility aids in the maintenance of appropriate substrate orientation and improving catalytic efficiency.

Val367 exhibits similar characteristics as Thr302, as they are both essential for ligand stabilisation within the active site. This was observed with the inhibitor tert-butylphenylacetylene (BPA), as it is situated close to the heme iron and acetylene triple bonds (The simplest hydrocarbon that contains a triple bond). Docking simulations were performed by Gay et al. [54], which revealed the significant tactical placement of Val367 and Thr302 during the ligand-binding process. Thr205's contribution to the active site dynamics is more complex as Gay et al. [54] demonstrated its effect on substrate selectivity and sensitivity to mechanism-based inactivation.

Across the *CYP2B* isoforms, such as *CYP2B1*, *CYP2B4* and *CYP2B6*, the two residues, Phe206 and Val367, show an exceptionally high degree of conservation, highlighting their functional importance in the larger context of the enzyme family. Phe206 and Val367's functions in androgen metabolism are clarified by comparing mechanism-based inactivation, especially with BPA and bergamottin (BG). The effects of these residues to substrate selectivity and catalytic efficiency are better understood through substitution experiments by Lin et al., [51]. Studies have shown that residues like *Phe206* and *Val367* play critical roles in substrate recognition and binding, influencing *CYP2B6's* metabolic efficiency and selectivity [52,54].

Lastly, the significant cysteine residue functions as an essential thiolate ligand to the heme iron in the active site [25]. The catalytic activity of the enzyme depends on this contact. During the catalytic cycle, the thiolate ligand is essential for the activation of molecular oxygen, which facilitates the oxidation processes that follow. Cysteine and heme iron work in tandem to maintain the integrity of the heme prosthetic group and to guarantee that *CYP2B6* can properly metabolise a wide range of substrates. These residues' minor contributions together shape the environment of the active site, enabling the *CYP2B6* enzyme to identify, bind, and catalyse a wide range of substrates [25,35].

1.2.2 Substrate specificity and Genetic Polymorphism in *CYP2B6*

Understanding the factors that affect *CYP2B6* substrate binding and selectivity is essential for predicting its role in the metabolism of new drugs and their metabolites. This understanding is crucial for assessing potential drug-drug interactions, optimizing drug dosing, and predicting adverse reactions. The substrate specificity of *CYP2B6* is influenced by several factors, including molecular size, shape, electron density, and interactions with specific amino acid residues in the enzyme's active site. These interactions determine how efficiently the enzyme metabolizes a wide variety of substrates, from therapeutic drugs to environmental toxins.

It is essential to comprehend *CYP2B6's* function in drug metabolism to determine how it will affect treatment outcomes and possible drug interactions. From the paper done by Hedrich et al., (2016) an extensive list of important substrates for *CYP2B6*, which includes antiretrovirals such as nevirapine and efavirenz, bupropion, ketamine, as well as cyclophosphamide and artemisinin are provided [59]. *CYP2B6*'s polymorphic nature further emphasises the role in drug metabolism. Genetic variations in the enzyme can result in significant interindividual differences in enzyme activity. These polymorphisms, especially loss-of-function variants, can lead to higher plasma concentrations of certain medications, increasing the risk of toxicity [60]. For example, variations in *CYP2B6* polymorphisms influence the metabolism of the antiretroviral drug efavirenz, which can affect its efficacy and toxicity, particularly in different ethnic groups. This variability emphasizes the importance of personalized medicine, where pharmacogenomic information can guide treatment plans and optimize drug effectiveness for individual patients [59].

Additionally, distinguishing between enzyme induction and inhibition is crucial for understanding drug metabolism. Induction occurs when a drug or substance increases enzyme activity, often by activating nuclear receptors such as the pregnane X receptor (PXR) and constitutive androstane receptor (CAR). This can lead to enhanced drug metabolism, reducing the therapeutic efficacy of some drugs. Conversely, inhibition occurs when a drug reduces enzyme activity, which can lead to increased drug concentrations and heightened toxicity [61,62,63]. Comprehending these pathways offers valuable perspectives on the complex interactions between medications and enzymes, impacting plasma levels and the results of treatment.

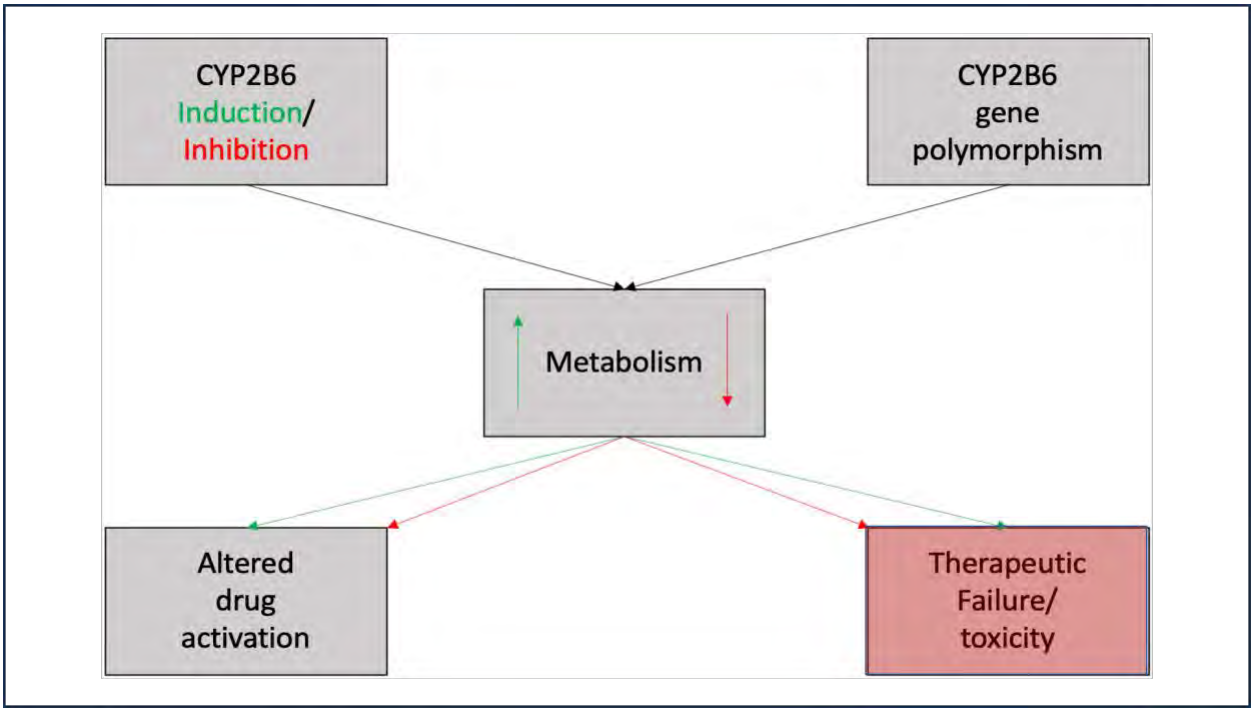


Figure 1.4: Effect of genetic polymorphisms in combination with inhibition and induction giving rise to possible outcomes. Text and arrows in green represent induction while those in red represent inhibition. Therapeutic failure or toxicity is represented in a red box.

Figure 1.4 illustrates the effects of genetic polymorphisms in combination with enzyme induction and inhibition, showing how these factors can result in therapeutic failure or toxicity. For example, genetic variants like CYP2B6*6 can reduce enzyme activity and lead to increased plasma concentrations of particular medications such as efavirenz, potentially causing side effects [61,63]. Moreover, drug-drug interactions—including those involving inhibitors like fluconazole—may raise the plasma concentrations of medications that are metabolised by *CYP2B6* and raise the risk of side effects [63,64,65]. In patients with liver impairment, disease-related alterations in drug metabolism can further complicate treatment, increasing the risk of toxicity[65,10].

This information clarifies the complex role of CYP2B6, highlighting the impact of genetic variants and other factors on drug metabolism. By supporting personalized medication plans tailored to an individual's genetic profile and considering drug interactions and health conditions, these insights advance personalized medicine. As pharmacogenomics progresses, integrating these findings into clinical practice will improve treatment outcomes and ensure safer, more effective therapies.

1.3 Project Aims, Objectives and Motivation

1.3.1 Problem Statement, Aim and Hypothesis

Problem Statement

Genetic variability within the drug-metabolising enzyme *CYP2B6* introduces complexity into drug metabolism and responses. Understanding the functional impact of genetic variations in *CYP2B6* is crucial for devising optimised treatment strategies.

Aim

The aim of this study is to use molecular dynamics techniques to investigate the effect of Afrocentric missense variations on the structure of CYP2B6.

Hypothesis

Our hypothesis is that specific missense variations in *CYP2B6* can cause changes to the structure and function of the enzyme and may induce changes to the active site. These structural changes may then impact substrate binding, resulting in changes in drug metabolism. By investigating the molecular basis of these variations, we aim to uncover insights into the functional consequences that drive variations in drug response.

1.3.2 1.3.2. Objectives

1) Identifying Afrocentric variants for *CYP2B6*
2) Model the 3D structures of wild-type and variants of *CYP2B6*.
3) Conduct molecular dynamics simulations to explore dynamic changes in the enzyme's structure.
4) Analyse the functional consequences of identified missense variations on drug metabolism.
5) Investigate the correlation between wild-type and mutated *CYP2B6* in the context of drug metabolism.

These objectives collectively form a structured approach to address the identified knowledge gaps and contribute to a more comprehensive understanding of the genetic variations in *CYP2B6* and their implications for structure and function of the enzyme.

1.3.3 Knowledge gap

CYP2B6 is an important member of the cytochrome P450 superfamily, as it is essential for the metabolism of certain drugs. Although there are extensive studies done, there remains a clear knowledge gap about how certain missense variations affect the structure and function of the enzyme including the effect on drug metabolism. Understanding the impacts of genetic variations is essential for tailoring drug therapies to individual patients, enhancing therapeutic results and more so for minimising potential side effects.

While the importance of genetic variations in *CYP2B6* is acknowledged, the lack of thorough research on the structural and functional consequences of missense variations hinders the ability for prediction of their impact on drug metabolism. Minimising this gap is crucial for advancing pharmacogenomics and precision medicine, especially when it comes to developing individual genetic profile regimes.

Existing literature has mostly concentrated on genetic differences in drug metabolising as well as structure and function for a specific population, much less focus has been directed to the African population. Comprehending the influence of genetic variations on drug metabolism, structure and function within the African populace is imperative in order to formulate treatment approaches.

1.4 Conclusion

In conclusion, this chapter offers a detailed exploration of the critical role of pharmacogenomics, with a specific focus on the cytochrome P450 (CYP) enzyme, particularly CYP2B6, in drug metabolism and responses. It underscores the importance of understanding genetic variations for tailoring medical treatments to individual patients, thereby optimising therapeutic outcomes and minimising potential adverse effects.

The chapter begins by emphasising the transformative paradigm of precision medicine, which recognizes the impact of genetic factors on responses to medications. Pharmacogenomics emerges as a pivotal field within this landscape, aiming to individualise treatment plans based on genetic makeup. Specifically, CYP enzymes, particularly CYP2B6, are highlighted for their central role in drug metabolism.

Detailed insights into the structure and function of CYP2B6 are provided, along with explanations of how genetic polymorphisms can influence enzyme activity and subsequently impact drug metabolism. The significance of customising medication regimens based on individual genetic composition is underscored, as it can maximise therapeutic results while ensuring patient safety. The chapter concludes by outlining the aims and objectives of a study aimed at investigating the effect of Afrocentric missense variations on the structure and function of CYP2B6 using molecular dynamics techniques. This research aims to address knowledge gaps in understanding the functional consequences of genetic variations in CYP2B6, particularly within the African population.

Overall, this chapter serves as a comprehensive overview of the intricate interplay between pharmacogenomics, CYP2B6, and drug metabolism, highlighting the importance of personalised medicine approaches in optimising patient care.

2 Chapter 2: Data Retrieval and Allele Identification

2.1 Introduction to Homology Modelling

Structural biology has rapidly evolved, progressively revealing complex details of biomolecular architecture since the mid-20th century. Homology modelling (also known as comparative modelling) is the computational technique that is widely used in structural bioinformatics to assist in the prediction of a 3D structure of a protein, based on its sequence similarity to one or more proteins of known structure [66,67,68].

The process of homology modelling relies on the observation that the structural conformation of a protein is more highly conserved than its amino acid sequence. Using tools like Primo and Modeller, homology modelling involves four main steps: template identification, alignment, model building and refinement, and validation (discussed in detail later in this chapter). This approach offers a fast and cost-effective means of protein structure prediction, with the added benefit of its increasing significance in drug discovery and industrial applications over the past decade. Additionally, methods such as Nuclear Magnetic Resonance Spectroscopy (NMR) and X-ray Crystallography provide high-quality structural templates that can be used to build accurate 3D models through homology modelling.

2.2 Principles of Homology Modelling

Homology modelling follows a systematic four-step workflow, starting with identifying a suitable target sequence and aligning it to a resolved template structure. As shown by Edgar (2004), protein structures tend to be more conserved than protein sequences[68]. Consequently, an alignment with over 20% sequence similarity typically increases the likelihood of producing an accurate model. The following subsections detail the methods applied in this study.

2.2.1 Sequence alignment.

A large part of homology modelling relies on the uncovering of evolutionary connections embedded in protein structures. At its core is sequence alignment, this is regarded as a fundamental process in which intricate patterns are revealed about amino acid sequences. Within the

Bioinformatics field, there are specialised tools that consist of advanced algorithms that can identify shared residues and reveal evolutionary patterns. Such tools that employ these algorithms are MUSCLE (Multiple Sequence Comparison by Log-Expectation) and PROMALS3D (PROfile Multiple Alignment with predicted Local Structures and 3D constraints) [68,69]. One commonly used approach in these tools is the progressive alignment algorithm. In progressive alignment, sequences are first aligned in pairs to create a guide tree, which represents their relative evolutionary distances. This tree then directs the hierarchical alignment of all sequences, starting from the most similar pairs and moving towards the inclusion of all sequences in the final alignment. By building alignments in a stepwise manner, progressive algorithms enable the gradual integration of evolutionary information, facilitating more accurate sequence comparisons.

2.2.2 Template Selection

Template selection involves carefully taking certain properties or factors into considerations such as: sequence identity, structural coverage and the reliability of experimental data [70] The careful consideration of parameters during template selection is critical to guaranteeing the accuracy of the modelling process. Sequence identity is important because a higher similarity between the target and template proteins suggests a tighter evolutionary link [70]. Structural coverage is required to capture crucial properties of the target protein such as structural coverage, sequence identity and validation of experimental data. These properties help ensure that the template used, accurately represents its whole structure. Common issues include incomplete structural coverage—where only part of the protein's structure is resolved—or low-resolution data, which can limit the reliability of the structural information. Templates obtained from experimental techniques like X-ray crystallography and NMR spectroscopy are generally preferred, but even high-quality crystallographic data may vary in resolution, impacting the clarity and detail of the model. For instance, resolution values lower than 2 Å provide high detail, whereas values above 3 Å may obscure finer structural features. Structural coverage issues may also arise when certain protein domains are flexible or difficult to crystallize, resulting in models with missing or poorly defined regions. The reliability of experimental data further complicates template selection. Templates with high-resolution data (often in Ångström, Å) yield finer detail, but even well-validated structures may contain uncertainties, particularly in flexible regions [71,72]. The variability can impact the accuracy of the final model, especially when constructing models for

drug-binding or highly specific functional studies. A low-resolution score, typically expressed as a higher numerical value, reflects reduced clarity, which may limit confidence in the resulting model. By carefully considering these factors, homology modelling can leverage high-quality templates to create accurate 3D models that reflect the complex structural characteristics of target proteins.

2.2.3 Model building and Model Validation with z-DOPE score.

When generating a protein structure, homology modelling transitions from theory to reality. Molecular mechanics and optimisation algorithms are utilised, transforming the basic framework (sequence alignment and template selection) into a 3D model [70]. During model refinement, stereochemical quality, energetics, and structural alignment are assessed to ensure the model closely reflects the properties of a native protein. The refining procedure, which is similar to sculpting precise features, brings the model closer to experimental results. This part breaks down the complexities of model development and refining, providing a thorough grasp of the processes that convert sequence information into physical molecular structures [74].

Once a model has been generated, there needs to be some type of validation of the model that occurs. Common tools utilised for validation include PROCHECK and Verify3D amongst others [75,76]. Other steps used for validation include a Ramachandran plot and the normalised Discrete Optimised Protein Energy (z-Dope) score. These validation steps check for the accuracy and quality of the structure produced, respectively. Within this study we make use of the z-DOPE score to validate our generated model.

To understand the z-DOPE score, we first need to understand what the DOPE score represents. The DOPE score is a statistical potential that evaluates the quality of a protein model based on its energetics, stereochemistry, and overall structural characteristics. Stereochemistry plays a significant role in evaluating model quality, with parameters like bond angles, bond lengths, and torsional angles (assessed through tools such as Ramachandran plots) revealing how closely the model conforms to typical protein structures. Additionally, energetic stability is assessed to determine whether the modelled structure resides in a favourable energy conformation, with lower potential energy indicating a more realistic conformation.

The z-DOPE score extends this evaluation by normalizing the DOPE score against a dataset of known protein structures, enabling comparisons across different models. A z-DOPE score below -1 is considered indicative of a model that closely resembles native protein conformations, making it a reliable metric for assessing structural quality. This score is particularly sensitive to structural deviations, making it a priority in evaluating protein models

.. A lower DOPE score indicates a better model since it implies that the modelled structure is more energetically favourable and closer to native-like conformations [77]. The z-DOPE score is frequently used to rank alternative models produced by homology modelling or other structure prediction methods. The z-DOPE score is a normalised version of the DOPE score that compares the resulting DOPE score to the scores of a large number of known protein structures [78]. It evaluates how well the model compares to structures in the protein structure database. The z-DOPE score is advantageous because it normalizes DOPE scores across various protein structures, allowing for an objective comparison to known structures in the protein database. By providing a score in relation to a range of native-like conformations, it helps distinguish between realistic models and those likely to have inaccuracies. A more negative z-DOPE score indicates a higher likelihood of native-like quality, whereas a less negative score suggests areas where the model may diverge from expected structural norms. This prioritization was chosen due to the z-DOPE's sensitivity to subtle variations in structural accuracy, making it suitable for detecting both favourable and potentially erroneous regions in homology models, particularly in the absence of high-resolution experimental data [78].

2.3 Genetic Variations

CYP2B6's genetic variation creates a complicated tapestry of drug metabolism, resulting in significant consequences for therapeutic outcomes and personalised medicine. This chapter narrow down its focus to *CYP2B6* alleles that have traditionally been found in a variety of African communities highlighting their potential influence on medication responses, pharmacokinetics, and the development of adverse drug reactions.

Genetic polymorphisms within the *CYP2B6* gene play a major role in determining interindividual variability in medication metabolism and response. Several well-studied single nucleotide polymorphisms (SNPs), including rs3745274 (CYP2B6*6), rs2279343 (CYP2B6*4), rs3211371 (CYP2B6*5), rs34223104 (CYP2B6*22) , rs28399499 (CYP2B6*18) and rs34826503 (CYP2B6*19) introduce variations in amino acid sequences and can significantly impact enzyme expression levels [87,88]. This genetic variability, which has been reported throughout populations, contributes substantially to the wide range of functions exhibited by *CYP2B6*. For instance, certain polymorphisms are associated with increased transcriptional activity, resulting in elevated *CYP2B6* concentrations in the liver, while others may lead to reduced expression. This diversity in enzyme expression modulates the metabolic capacity of individuals with distinct genotypes, affecting their ability to metabolise specific substrates [23].

The catalytic activity of *CYP2B6* is influenced by these genetic polymorphisms, resulting in altered substrate specificity and varying metabolic rates. Individuals with specific *CYP2B6* variants may be classified as slow or fast metabolizers, affecting the clearance rates of drugs metabolised by this enzyme [81]. This has profound implications for drug pharmacokinetics, as slow metabolizers may require lower drug doses to prevent toxicity, while fast metabolizers may need higher doses for therapeutic efficacy (Figure 1.4.).

2.3.1 Genetic Diversity of *CYP2B6* in African Populations

Previous studies have emphasised the clinical impact of *CYP2B6* polymorphisms, and in conjunction with the metabolism of efavirenz, a key component in the treatment of HIV that was investigated in Papua New Guinea [40,90]. *CYP2B6* is responsible for the metabolism of efavirenz, and individuals with certain polymorphisms may exhibit significantly altered plasma concentrations of the drug. This not only affects treatment efficacy but also influences the risk of adverse effects associated with efavirenz, emphasising the importance of considering genetic variations in treatment planning.

Beyond efavirenz, the clinical implications extend to various drugs metabolised by *CYP2B6*, including bupropion and cyclophosphamide [82]. The polymorphic nature of *CYP2B6* contributes to variable drug responses, necessitating a personalised approach to drug dosing. Integrating

genetic information into clinical decision-making becomes crucial for optimising therapeutic strategies and minimising the risk of adverse drug reactions [90,91]. The importance of understanding the impact of these genetic variations on drug pharmacokinetics and patient responses assists in producing more advanced HIV treatments. By elucidating the unique genetic variation within CYP2B6, the study done by [82] underscores the potential for personalised medicine approaches tailored to the genetic profiles of individuals from African populations [77].

2.3.2 Recent history of CYP2B6 alleles

There is a prevalence of CYP2B6*6 allele in Papua New Guinea and West African populations [89,91], where these studies found significant differences in the allele and genotype frequencies between the two populations. This suggests that these populations are significantly different at the *CYP2B6* locus [84], which may also influence treatment outcomes with efavirenz. Several studies highlight the need for further research to fully understand the impact of *CYP2B6* alleles and polymorphisms on HIV/AIDS treatment in different populations as well as the conjunction with efavirenz. This includes comparative pharmacokinetic studies to determine the differences in plasma efavirenz concentrations and treatment outcomes between populations with different *CYP2B6* genotypes.

Another significant variation, rs28399499 (CYP2B6*18) , appears mostly in Africans (4-12%) and does not express functional protein [85]. A 1000 Genomes project that was established has become of significant interest as a large number of uncharacterised variants are emerging, each from multiple different races and populations [86]. According to Zanger (2013) [85], a study of several alleles were investigated for their functional impact on drug metabolism, particularly for their effects on the metabolism of drugs such as bupropion, efavirenz, 7-ethoxycoumarin, selegiline, and artemether. The study aimed to understand how genetic variability in *CYP2B6* affects drug efficacy and toxicity, and to identify potential pharmacogenetic markers for personalised drug therapy. Additionally, the study aimed to identify novel variants of *CYP2B6* in different ethnicities and to characterise their functional impact.

The study revealed that CYP2B6*5 expresses very low levels of protein [87] , which does not translate into similarly reduced activities as measured with bupropion as well as efavirenz [88].

The variant was shown to have higher specific activity compared to reference. In the human liver, this leads to partial compensation of low expression, finally resulting in a phenotype with moderately decreased activity with bupropion and efavirenz in vivo [89]. The next allele was CYP2B6*6, which is common in most studies in most populations to date. It harbours two amino acid changes, *Q172H*, and *K262R*. Although additional variants occur in the promoter and in introns, their functional impact appears to be of limited relevance according to what is found. Followed was CYP2B6*9, which occurs in isolation, although at much lower frequencies. In most pharmacogenetic studies, it is not being determined, and functional data is therefore rare. Next was CYP2B6*18, this allele occurs predominantly in African subjects with allele frequencies of 4–11%. The *I328T* variant expressed no detectable protein or activity toward bupropion, 7-ethoxy-4-trifluoromethylcoumarin (7-EFC), selegiline, and artemether in COS-1 cells, demonstrating another example of expression system-dependent differences.

Allele CYP2B6*22, is a gain-of-function variant associated with increased transcription in vitro and with increased activity in vivo. It was shown that a -82T>C exchange alters the TATA-box into a functional CCAAT/enhancer-binding protein binding site that causes increased transcription from an alternative downstream initiation site [90]. Next was CYP2B6*30 where this allele is a structural variant including a novel *CYP2B6*/*2B7P1* duplicated fusion allele. CYP2B6*33 until *37, are considered as novel and previously uncharacterized amino acid variants in different combinations that were identified by resequencing the *CYP2B6* gene in a Rwandese cohort of HIV-1-infected patients. At least four of the variants were shown to result in complete or almost complete loss of function with bupropion and efavirenz as substrates. CYP2B6*35, a splice variant that results in a truncated protein with a premature stop codon. It is associated with decreased catalytic activity. CYP2B6*37 *40, *42, *46, *47, *48: These alleles are missense variations that result in a protein with decreased catalytic activity [94,95].

Although these conclusions were established from the said study, the enzyme was not looked at individually to assess the variations on its own without being bound to a specific substrate. For the purpose of this study, this chapter sets out to comprehensively analyse the identification of *CYP2B6* alleles known in the African population. As we scrutinise specific alleles, we embark on a journey to unravel the genetic tapestry of *CYP2B6* within the African population.

2.4 Methods

2.4.1 Structure Retrieval

The initial step involved retrieving the *CYP2B6* enzyme from the Protein Data Bank (PDB) (UniProt accession:P20813) [91] through its authoritative platform RCSB [92]. The search criteria were strategically defined: encompassing the enzyme name (*CYP2B6*), *Homo sapiens* as the source organism, X-ray diffraction as the experimental method, refinement resolution below 2.5 Å, and enzyme classification as oxidoreductase. These stringent parameters aimed to select a structure that aligns with high-resolution standards and accurately represents the *Homo sapiens CYP2B6* enzyme.

As a result of the search criteria, 10 structures of *CYP2B6* were available. The structure with the best resolution (5UFG) was chosen. A single PyMOL (version 2.4) session was employed for examination, ensuring conformational accuracy, identifying, and rectifying missing residues, and eliminating extraneous compounds such as water and ligands. To validate the structure of the protein sequence, a multiple sequence alignment was conducted using MUSCLE [68]. The wild-type sequence, extracted from UniProt with the accession ID P20813 in FASTA format [93], served as the reference. The chosen structure underwent refinement using MODELLER v10.4 [94]. The wild-type sequence, extracted from UniProt, served as the template for this process, ensuring that the resulting model sequence precisely mirrored the UniProt reference. Notably, specific residues in the model sequence, specifically the 29th and the last residue before the heme (residue 492), were adjusted to align with the wild-type reference sequence. This involved changing the lysine residue to arginine (L29A) and removing the histidine residue (H492).

2.4.2 Data collection

The PharmVar database [86], served as the primary source for acquiring comprehensive information on *CYP2B6* alleles along with relevant literature sources. The “genes” tab on the PharmVar database was selected, thereafter selection of the *CYP2B6* enzyme. The selection criteria for data inclusion involved a review of studies and data sets within PharmVar, and studies mentioned earlier in this chapter, focusing on the information relevant to *CYP2B6* alleles. The criteria for allele selection included: Pharmacogenomic relevance (ensuring that the alleles were specifically associated with pharmacogenomic implications), African population frequency

(prioritised alleles with documented prevalence or relevance within the African population) and data availability (selected alleles for which frequency data was available as well as overall enzyme function that was accessible). In addition to PharmVar, literature sources contributing to prevalence of Afrocentric alleles were used to select alleles for the study,

2.4.3 Introduction of Variations and Protonation of Structures

The final refinement stage involved the introduction of variations using Discovery Studio [95], with the refined model structure serving as the reference (reference). Each *pdb* structure file was separated into protein, heme and iron (Fe) counterparts. The protein was submitted to H++ for protonation at a pH of 7.4 [96]. Thereafter the two corresponding files (*.top* and *.cro*) produced made up the *pdb* protonated protein file after being combined together using the program AMBER *ambpdb* [97]. The *reduce* program in Amber facilitated the protonation of the heme. The protein, heme and Fe were concatenated to form a single protonated *pdb* structure file. The final step included cleaning up the file and renumbering the structure using the *pdb4amber* program found in AMBER [97].

2.5 Results and Discussion

2.5.1 Modelling of Reference Structure

From criteria set out for structural selection of *CYP2B6* on PDB, 10 structures were identified. All identified structures for *CYP2B6* contained variations. The structure that had the best resolution was chosen for down processing applications. In Table 2.1 below, the 10 structures are shown with their Entry ID, resolution and factors that were considered when filtering out the criteria.

Table 2.1: List of *CYP2B6* structures, from criteria set out, displaying properties of importance such as Entry ID, Structure determination Methods, Resolution (Å), Molecular weight (kDa), Experimental Method and Structural Keywords.

Entry ID	Structure Determinatio	Resolution (Å)	Molecular Weight (kDa)	Experimental Method	Structure Keywords

	n Methods				
5UFG	experimental	1.76	56.95	X-Ray Diffraction	Oxidoreductase
5UDA	experimental	1.93	113.77	X-Ray Diffraction	Oxidoreductase
3IBD	experimental	2	56.66	X-Ray Diffraction	Oxidoreductase/ Oxidoreductase Inhibitor
4I91	experimental	2	57.06	X-Ray Diffraction	Oxidoreductase
5UAP	experimental	2.03	113.93	X-Ray Diffraction	Oxidoreductase
3QOA	experimental	2.1	56.42	X-Ray Diffraction	Oxidoreductase/ Oxidoreductase Inhibitor
4RQL	experimental	2.105	113.89	X-Ray Diffraction	Oxidoreductase
4RRT	experimental	2.2	113.95	X-Ray Diffraction	Oxidoreductase
4ZV8	experimental	2.24	56.41	X-Ray Diffraction	Oxidoreductase
5UEC	experimental	2.27	56.96	X-Ray Diffraction	Oxidoreductase

Among the identified structures, PDB ID 5UFG was selected due to its superior resolution (1.76 Å). Notably, all structures in Table 2.1 shared Y226H/K262R variations, with 5UFG containing an additional I114V variant (can be found here). As 5UFG is to be the template sequence, a multiple sequence alignment was run using MUSCLE against the UniProt sequence [69,75,81].

Alignment with the UniProt reference sequence revealed minor discrepancies at the termini (Arginine 29 and Histidine 492), which were manually corrected to ensure consistency with the reference sequence.

Four different structures were produced, Table S1 in the Supplementary data reveals the z-DOPE score of the structures. The model 5UFG_WT.B99990002.pdb achieved a z-DOPE score of -1.715, indicating high accuracy and suitability for downstream applications (Table S1). The achieved z-DOPE score suggests that our model represents an accurate depiction of the native protein structure. The final modelled structure as well as the multiple sequence alignment performed can be found below in Figure 2.1.

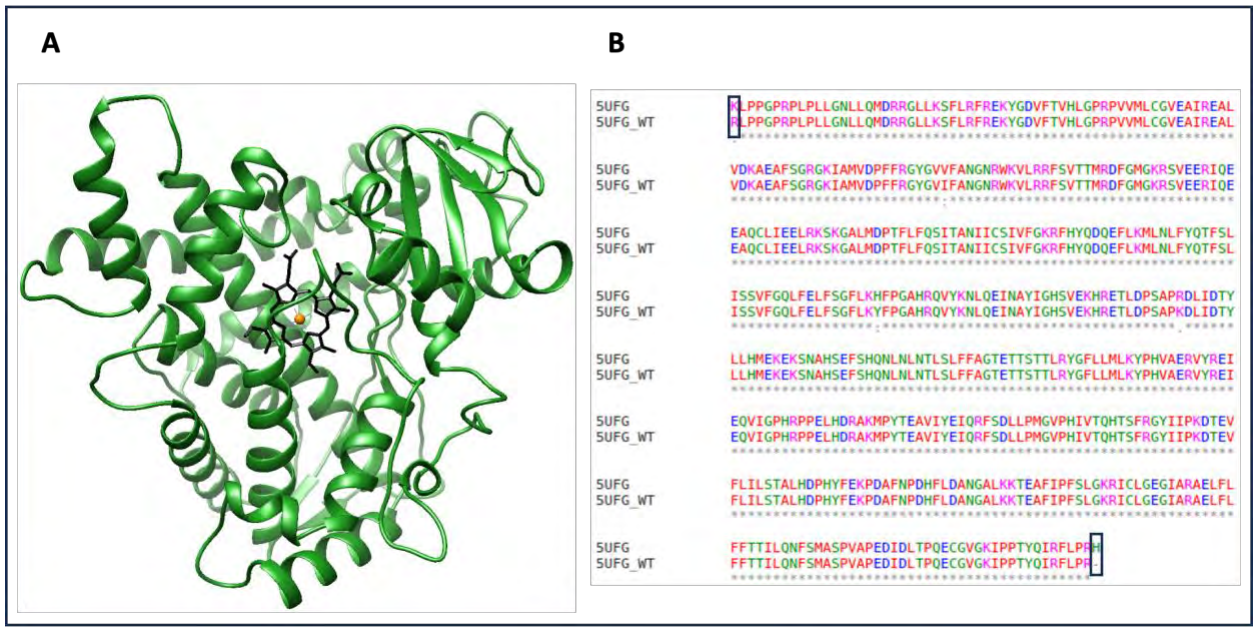


Figure 2.1: Model and sequence of the *CYP2B6* enzyme. **A)** Structure of modelled CYP2B6 (PDB ID:5UFG) **B)** Multiple sequence alignment with MUSCLE. The black boxes indicate the manual corrections that had to be made according to the UniProt reference sequence. 5UFG represents the sequence retrieved from UniProt. 5UFG_WT represents the sequence retrieved from PDB.

Mentioned above, a significant change that had to be made was the lysine residue to the arginine residue. Because of proline's unique structural properties, changing a residue before a proline-rich region is an important issue. Proline is an amino acid with a unique cyclic structure that forms a

closed ring around its amino group [99]. This structural characteristic constrains the phi (ϕ) dihedral angle, restricting the flexibility of the backbone relative to other amino acids [100]. Proline's rigidity has the potential to modify the protein's overall conformation.[101].

The identity of the residue immediately preceding a proline-rich region can have an impact on local structure and conformational dynamics therefore a change that differs from that of the reference may have a different fold around that proline-rich region altering the way the protein behaves [101]. An abrupt shift in amino acid type before a proline might cause the protein structure to become irregular and unstable, however in certain instances the variation of a residue contributes to the protein being crystallised. The transfer from one amino acid to another modifies the immediate environment, causing changes to sidechain interactions and hydrogen bonding patterns [102]. Proline residues frequently influence local structural characteristics in proteins, including turns and bends as a result [97,98], any change to the residue preceding a proline-rich region has the potential to impair conformational stability, disrupt secondary structure elements, and influence protein folding overall. This is especially true in the context of homology modelling, where proper preservation of local structural motifs is critical for producing valid 3D models.

2.5.2 Allele Selections and Locations

This study focused on 11 CYP2B6 missense variants frequently observed in African populations. These variants, selected from PharmVar and relevant literature [86-94], display diverse enzyme activity profiles, providing a basis for understanding allele-specific structural impacts on drug metabolism. In some cases, the alleles had certain effects which contributed to their activity status, whereas some of the alleles had no activity associated with its function.

Table 2.2: *CYP2B6* Alleles - associated effects of enzyme activity for variations

Allele [95]	**variation**	**Frequency**	**Enzyme Function**
CYP2B6*3	S259R	A=0.6695 | G=0.3305	Uncertain
CYP2B6*4	K262R	A =0.6695 | G=0.3305	Increased
CYP2B6*5	R487C	C=0.9855 | T=0.0151	Neutral
CYP2B6*8	K139E	A=0.9995 | G=0.0005	No function

CYP2B6*11	M46V	A=0.9983 | G=0.0017	Uncertain
CYP2B6*12	G99E	G =1.0000 | A = 0.00	No Function
CYP2B6*14	R140Q	G=0.9997 | A=0.0003	Uncertain
CYP2B6*15	I391N	T=0.9983 | A=0.0017	Uncertain
CYP2B6*18	I328T	T=0.93175 | C=0.06843	No Function
CYP2B6*21	P428T	C=0.99970 | A=0.0003	Uncertain
CYP2B6*28	T360S|R378K	C=0.9995 | T=0.0005	No Function

For this study, the listed the alleles were selected (Table 2.2). The study of the *CYP2B6* genetic landscape in the African population revealed a varied picture of allelic variations, each with its own frequency and enzyme function profile. This will allow us to examine the discovered *CYP2B6* alleles in depth, offering information on their frequency and potential influence on enzyme activity. Keeping in mind that the effect of the variation on the enzyme's structure will be of prime focus. The location of each variation is displayed below on the structure of *CYP2B6* in Figure 2.2.

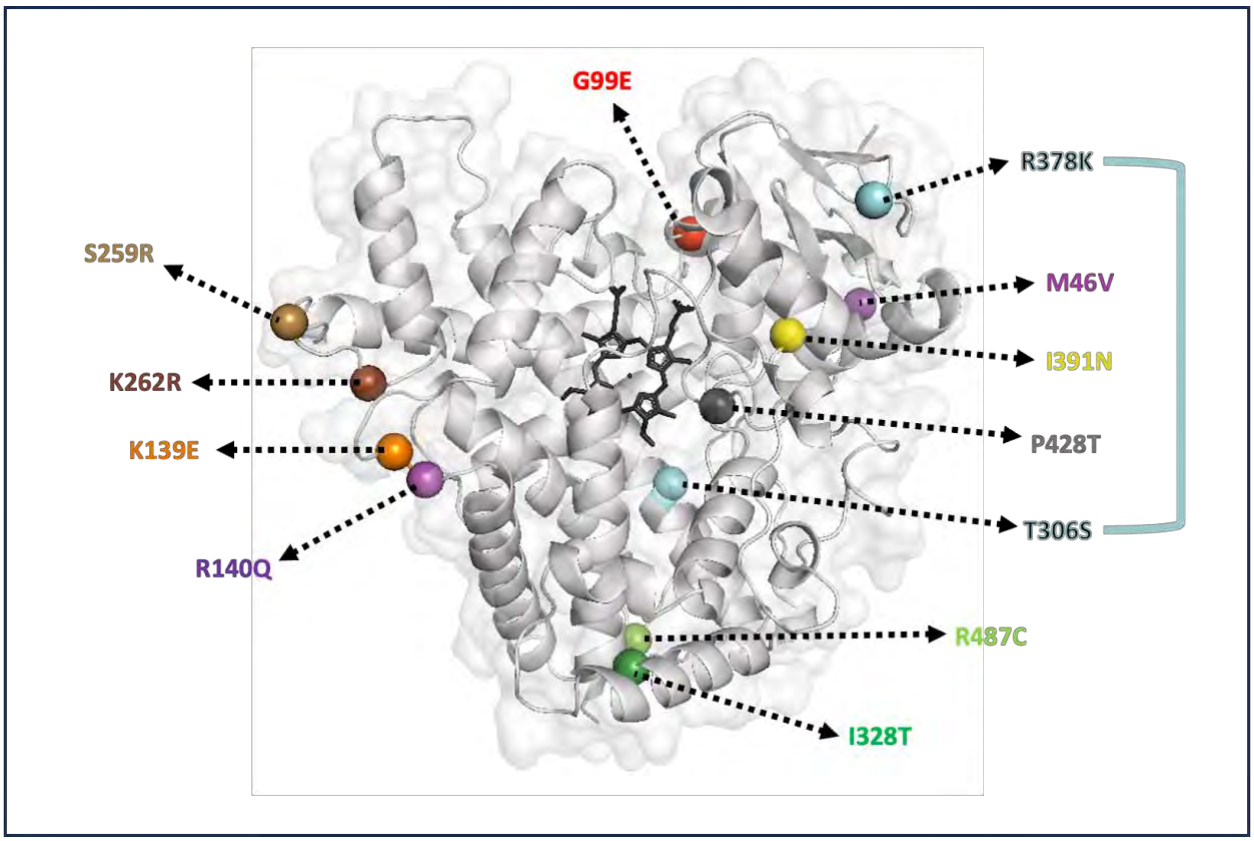


Figure 2.2: Structure of *CYP2B6* (PDB-5UFG) representing the location of each variation. Each variation is represented with a wheat-coloured sphere and black arrows used to show the specific variation. The cyan line represents the double variation that occurs within CYP2B6*28.

This visual representation allows us to identify where the variations occur and gives insights on the potential effect some variations may have. Referring to Table 2.2 and Figure 2.2, we are able to see how there is a diverse array of alleles within the African population [86]. Amongst these variants, the *S259R* variation that is marked for CYP2B6*3 results in a frequency distribution that favours adenine (A=0.6695) over guanine (G=0.3305). However, the functional repercussions of this variation are unknown, necessitating more research to determine its involvement in drug metabolism [89]. In terms of its enzymatic function, the same can be seen in CYP2B6*11, CYP2B6*14, CYP2B6*15 and CYP2B6*21. The *M46V* variation in CYP2B6*11 is largely represented by the adenine allele (A=0.9983) while the CYP2B6*14, which has the *R140Q* variation, has a high proportion of the guanine allele (G=0.9997). The *I391N* variation in CYP2B6*15 results in a high frequency of the thymine allele (T=0.9983) while the cytosine allele (C=0.99970) dominates CYP2B6*21, which has the P428T variation [100,92].

The allele frequency of adenine (A=0.9995) greatly favours the *K139E* variation in CYP2B6*8. This variation, however, is associated with a significant decrease of enzyme activity, posing possible issues in medication metabolism for those harbouring this allele [100,101]. The *G99E* variation distinguishes CYP2B6*12, which only has the guanine allele (G=1.0000). This allele is linked to a loss of enzyme activity, implying that people who carry this variation may face difficulties with medication metabolism, a similar occurrence is seen with CYP2B6*28 which has a mainly cytosine allele (C=0.9995) due to the *T360S|R378K* variation [100,102].

The CYP2B6*4 variation, which has the *K262R* variation, has a more balanced frequency distribution of adenine and guanine alleles (A=0.6695 | G=0.3305). Notably, this variation is linked to enhanced enzyme function, implying possible changes in drug metabolism and pharmacokinetics. [88,103,104,99]. In contrast to the ambiguity of CYP2B6*15, the *I328T* variation in CYP2B6*18 results in a mainly thymine allele (T=0.93175). This genotype is linked to impaired enzyme activity, implying a possible influence on drug metabolism and stressing the importance of personalised medication dosage [92,103] Moving ahead, the CYP2B6*5 variation, defined by the *R487C* variation, has a unique profile with a cytosine allele (C=0.9855) as the dominant allele. [92,90]. Despite its frequency, the variation is thought to be neutral in terms of enzyme activity, indicating that it has no effect on *CYP2B6* function in the African population.

In summary, the *CYP2B6* alleles discovered in the African population represent a diverse landscape, each with its own frequency distribution and possible repercussions for enzyme performance. More research and functional investigations are needed to completely understand the influence of these alleles on medication metabolism, allowing for individualised treatment methods matched to the genetic makeup of people in this community.

2.6 Conclusion

In conclusion, this chapter has highlighted the interplay between homology modelling techniques and *CYP2B6* genetic variations, focusing on alleles like *CYP2B6**4, *6, and *18, which hold significant implications for drug metabolism, particularly in the context of HIV therapy with efavirenz. The systematic approach employed—from structure retrieval to analysis—provided a

model that aids in predicting *CYP2B6's* role in individual drug response, especially among African populations where these alleles are prevalent.

Our findings underscore the importance of genetic variability in therapeutic planning, with *CYP2B6**4 linked to enhanced enzyme function and potentially altered drug metabolism, while *CYP2B6**18 may reduce enzyme activity, impacting drug efficacy and safety. By combining computational modeling with pharmacogenomic data, we establish a basis for tailored therapies that align with individual genetic profiles.

Continued research is necessary to deepen our understanding of how specific genetic variations impact drug metabolism and therapeutic responses, especially for personalized treatment strategies. This chapter illustrates the potential of bridging structural biology and pharmacogenomics, advancing precision medicine toward improved healthcare outcomes globally.

3 Chapter 3: Molecular Dynamic Simulations

3.1 Introduction

Molecular dynamics (MD) simulations have become of particular interest in computational biology, propelled by significant advances in computational power, especially the integration of graphical processing units (GPUs). The application of GPUs has extended the temporal reach of MD simulations [108], which allows many researchers to investigate several different processes at the microsecond time scale. While MD simulations have found applications across multiple disciplines, including medicinal chemistry, biophysics, and biochemistry, this study specifically focuses on their application to protein dynamics, with particular emphasis on *CYP2B6*, a crucial enzyme in drug metabolism.

Multiple aspects of biology can be investigated due to MD simulations, this diverse landscape of molecular research offers a molecular-level perspective which allows researchers to explore properties such as structural changes, and the stability of certain structures including chemical reactions and atomic-level dynamics [109]. The energies associated with simulation-related processes can be elucidated through MD simulations [108]. By simulating the interactions of molecules over time, MD simulations help predict how proteins like CYP2B6 will behave in different environments, thereby informing the development of more targeted therapeutics.

The concept of MD simulations lies in their capability to reveal the complexities of biological systems at the atomic level. By providing a virtual microscope into the dynamic behaviours of biomolecules, MD simulations enable the exploration of conformational changes, interactions, and dynamic fluctuations. The large amount of data generated by these simulations is then subjected to specialised analysis tools, such as Root Mean Square Deviation (RMSD), Root Mean Square Fluctuation (RMSF), and radius of gyration, to discern structural changes, stability, and flexibility.

At its core, MD simulation is founded on numerically solving Newton's equations of motion for a system of interacting particles, allowing the tracking of atomic movements over time [110]. The predictive capabilities of MD simulations have advanced with computational power and force field

development, enabling researchers to make informed predictions about the behaviour of biomolecules under varying conditions.

The ability to simulate protein dynamics is particularly significant for understanding enzymes like CYP2B6, whose conformational flexibility is central to its function in drug metabolism. By incorporating structural and dynamical information derived from MD simulations, researchers can obtain deeper insights into the molecular mechanisms of drug interactions, substrate binding, and enzymatic activity. This level of detail is essential for advancing our understanding of how genetic variations within CYP2B6 influence its function, especially in the context of personalized medicine

3.1.1 Evolution of Molecular Dynamics Simulations

The history of MD simulations mimics that of the progression of computational technologies. Initially, the first attempts consisted of investigating ways to model simple atomic interactions done by Alder and Wainwright in 1957-1959, this also included the study of hard spheres [107,108]. As computational power grew, the scope of MD expanded from simple systems to more complex and realistic models, thanks to key advancements in both hardware and numerical integration algorithms, such as the Verlet algorithm [112]. This stable integration method was instrumental in enabling longer, more accurate simulations, setting the stage for studying intricate biomolecular dynamics .

The development of multiple timescale (MTS) integration algorithms further enhanced MD efficiency by allowing different force components to act on different time scales, improving computational efficiency without compromising the full physics modeled by modern atomistic force fields [112]. This innovation facilitated larger time steps in biomolecular simulations, an essential step toward achieving more detailed and temporally extended simulations[110,111].

The application of MD simulations has expanded dramatically due to the convergence of classical many-body systems and quantum many-body systems [98], this was contributed by techniques such as the variational Monte Carlo, which brought new insights as the growth of computational power increased. The development of new algorithms not only increased the largest time step but

also gave insights for addressing quantum systems in which nuclear quantum effects were incorporated through the Feynman path integrals [109,110]. The paper by Ciccotti et al., (2022) gives better insight into how MD simulations initially began to advancements that improved over the last few decades [112]. The continuous evolution of MD simulations is thus deeply tied to the interplay of computational power, algorithmic innovations, and their applications, which have provided unparalleled insights into molecular dynamics that are often difficult or impossible to observe experimentally.

MD simulations play a pivotal role in the discovery of the dynamic nature of proteins, those related to the essential components of cellular function [115]. Of particular interest, proteins can undergo dynamic fluctuations which in turn can affect their structure, function and even interactions. This corresponds to that of protein dynamics which is defined as the time-dependent fluctuations of atomic coordinates [116]. It is important to note that understanding these dynamic behaviours becomes crucial when investigating the impact of genetic variations. The insights into how proteins move and adapt on a molecular level are largely possible due to MD simulations which have proven to be an invaluable tool for deciphering the consequences of genetic variations.

3.1.2 Software and Force Field Parameters

It is now routine to simulate proteins with hundreds of amino acid residues ranging anywhere from 10-100 ns surrounded by water and salts that are appropriate for the experimental setup [114,115]. Multiple platforms are available such as AMBER (Assisted Model Building and Energy Refinement), vCHARM, DL_POLY, NAMD, and LAMMPS [115]. In this study, GROMACS (GROningen MAchine for Chemical Simulation) was chosen as it is widely recognized for its efficiency, high accuracy, and extensive array of simulation tools, particularly suited for biomolecular systems like *CYP2B6*. GROMACS's performance is enhanced through its ability to handle parallel computing effectively, which is critical when conducting simulations that need to capture fine molecular details over extensive timescales [116,117]. Additionally, GROMACS is a free, open-source software with a large, active community of users and developers who contribute to its improvement and provide extensive libraries for analysis. As a result, numerous libraries are compiled together to carry out MD simulations [120]. Calculations for both bonded and non-bonded interactions are accommodated, with bonded interactions being more favourable. Most of

the GROMACS program's language was written in C programming language which consists of numerous executable programs that make use of tools designed for analysis of trajectories.

The choice of force fields heavily impacts the accuracy and outcome of MD simulations, more specifically the type of interactions between atoms and bonds. The AMBER force field consists of many subsets of force fields, specifically the Amber ff14SB force field which has sparked interest [121]. The ff14SB force field holds a crucial role in MD simulations with particular interest in the context of protein structure and dynamic studies [121]. The force field is considered a substantial improvement when compared to its predecessor, the ff99SB force field. This is primarily due to its enhanced accuracy in identifying and capturing the intricacies of protein behaviour. The outlining aspect of improvement for the ff14SB force field is the refinement of parameters associated with both the backbone and side chains of amino acids [118,119]. This improvement was driven by the need for a better alignment concerning experimental data such as NMR scalar groupings and order parameters [118,119,120]. Order parameters reflect the degree of mobility or rigidity of a particular atom or bond within a molecule, offering insight into the flexibility of protein regions, while scalar groupings refer to scalar coupling constants that provide information about the structural conformation of proteins. The predictive power that allows simulations to mimic more real-life environments thus revealing the behaviour of proteins is drastically improved in the ff14SB force field.

These improvements are significantly impactful when observing protein structures and dynamic simulations. The study done by Maier et al., (2015) looks at protein folding, where the authors investigate the process of a protein adopting its native and functional conformation, outlining the importance of accuracy [121]. The improved accuracy from the ff14SB force field allows researchers to look further into the insights provided and expand into the mechanism governing protein folding, investigating the complex pathways and dynamics involved. This study leverages these advancements to investigate *CYP2B6*, as understanding conformational flexibility is key to predicting functional variations.

In essence, the selection of GROMACS for MD simulations combined with the ff14SB force field represents a carefully considered choice aimed at achieving both computational efficiency and

accuracy. Its refined parameters and improved accuracy contribute to a more realistic portrayal of protein structures and dynamics. As a result, ff14SB has the potential to refine our understanding of the intricate relationship between protein structure and function, offering valuable insights with broad applications in both basic research and drug development.

3.1.3 MD Simulation Process

The production of MD simulations involves several key steps, and each step can be broken down into smaller steps that are essential to the whole process. The use of GROMACS for MD simulations, Figure 3.1 below visually describes the steps taken. For completing a GROMACS simulation, usually, three large steps are taken into consideration: a physical step which is subdivided into 4 steps, the simulation step which is subdivided into 3 steps and the analysis step which can comprise multiple analysis methods.

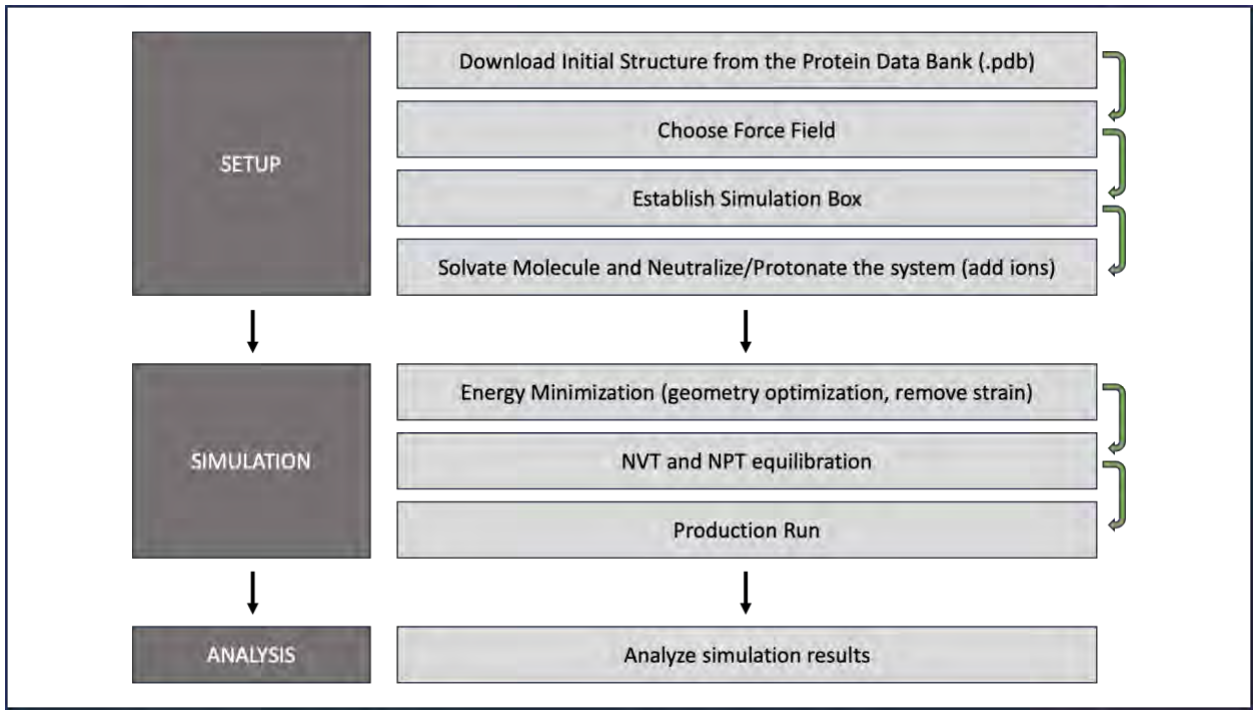


Figure 3.1: General working steps for an MD simulation using GROMACS. Arrows in black represent the workflow of stages while arrows in green display steps within stages.

The initial setup stage consists of retrieving the structure of the protein from the PDB [73,121], if no PDB structure can be retrieved from the PDB, a structure can be built from scratch using

computational packages, including ChemOffice and PRODRG [125]. The linkage of Chapter 2's homology modelling and its importance is seen within this chapter in the setup stage for MD simulations. The setup stage consists of generating a topology file which is followed by a *.gro* structure file and a position restraint file. The topology file encapsulates all the information that is essential about the protein, this includes atom masses, bond angles, lengths and even charges [126]. This file serves as a blueprint for subsequent calculations in GROMACS or any other software package that is chosen. Simultaneously, the created *.gro* file stores the protein's structure and details about its placement within the selected unit cell box which can be one of either of these: box, triclinic, hexagonal prism, dodecahedron or octahedron [127]. The position file that was created, restrains the substrate or ligand of desire, this concludes the setup stage.

The subsequent stage in MD simulation involves solvating the protein within the box, this involves the introduction of ions and water molecules into the structure and topology files [127]. Sodium and chloride ions are usually employed within the system to neutralise the system, ensuring a balance of positive and negative charges. The third stage consists of energy minimisation, where the overall aim is to eliminate or rid of steric clashes or unusual geometries that elevate the system's energy [127]. With this process, there are default options available, however within the *.mdp* file that outlines the parameters for energy minimisation, one can alter the parameters to the desired settings.

Following this is the fourth stage, which includes equilibration of the solvent around the protein which usually occurs within two steps. Initially, it is under constant temperature, constant volume (NVT) (default) ensemble which maintains the number of particles, volume, and temperature. The second step is constant temperature, constant pressure (NPT) ensemble, maintaining a constant number of particles, pressure and temperature [124,125]. It is important to note that if any restraints have been specified in any preceding steps, the protein will remain restrained in this process to allow solvent movement. Upon the completion of equilibration, the protein is then released from the restraints, in which it makes its way to the final stage, the actual MD simulation. From the simulation, trajectories are generated and outcomes from it can be analysed and assessed using visualisation tools such as VMD.

3.2 Methods

3.2.1 Metal Parameterisation

In the study conducted by Chebon-Bore et al.,[129], we incorporated the metal cofactor parameters that were established and based it on our reference (WT) structure. The Molecular Charge Potential Builder (MCPB) software, developed by Li and Merz (2016) [130], was utilised to create input files for Gaussian09 quantum mechanical (QM) calculations. A cutoff distance of 2.8 Å between the enzyme heme iron and the axial cysteine was set for these calculations. Following the generation of MCPB files, GaussView 5.0.9 validated the inclusion of cofactors and relevant residues in QM calculations [131]. QM calculations were executed using the B3LYP/6-31G* level of theory found within the Gaussian09 software, facilitated by the Centre for High-Performance Computing (CHPC) in Cape Town, South Africa. On the CHPC we utilised 240 CPU cores and 15 GB of memory with 24 cores per node.

Post-QM calculations, geometry optimization and bond angle/length analysis were conducted to refine the model further. Restrained electrostatic potential (RESP) charges were derived for surrounding atoms using the Merz-Kollman RESP fitting algorithm. These charges are critical for representing the accurate electrostatic environment in the force field. Finally, the force field parameters were adjusted using an equation to ensure consistency with the energy profiles computed using the AMBER energy function [129]. This accommodates atomic charges, bonds, angles, and dihedrals.

3.2.2 MD simulation

GROMACS 2018.6 orchestrated the 300ns MD simulations of the CYP2B6 enzymes which include both the reference and variant systems [127]. The simulations were performed on the CHPC using the GPU_1 queue which consisted of 10 CPU cores and 1 GPU core. The simulations were run in triplicate with the Berendsen barostat (equilibration and production) and then run in duplicate using the combination of the Berendsen and Parrinelo-Rahman barostat (equilibration and production, respectively). The AMBER ff14SB force field was harnessed making use of an octahedral box for computational efficiency. The box was solvated using TIP3P water molecules with the addition of Na+ and Cl- ions to establish an electroneutral and physiological environment.

The steepest descent algorithm was employed for energy minimization until the maximum force descended below 1000 kJ/mol/nm. During equilibration, position restraints were strategically applied to the *protein_cysteine_heme_iron* complex to prevent any interference. NVT and NPT equilibration of the systems unfolded over 0.2 ns, aligning the system to 300K and 1 bar, facilitated by the V-rescale thermostat, respectively. The production MD phase integrated the Particle Mesh Ewald (PME) algorithm for long-range electrostatic interactions with a cut-off of 1.2 nm, and hydrogen bond constraints were managed through the LINCS algorithm. The short-range interactions were set to a cut-off of 1 nm.

3.2.3 Post-MD calculations

After completion of the MD simulation, the initial steps involved periodicity correction which was performed using the *gmx trjconv* command line in GROMACS. This centred the protein in the simulation box and removed or corrected transitions and rotations (*-PBS -centre, rot+trans*). Each system was then carefully viewed using Visual Molecular Dynamics (VMD) version 1.9.3 [132] (The input files required for this included a *.xtc* and *.gro* file).

GROMACS was used to determine the RMSD with a focus on *Cα* atoms, the RMSF, and the Rg using the *gmx rms, gmx rmsf* and *gmx gyrate* GROMACS command lines, respectively. The plots were produced in Jupyter Notebook using Matplotlib.

3.3 Results and Discussion

The exploration of dynamic behaviours and conformational intricacies within the *CYP2B6* enzyme structures necessitated a specially designed Molecular Dynamics (MD) simulation protocol. For all the systems, the reference and Variant structures were protonated to accomplish a pH of 7.4, thereafter metal parameterization was employed to accommodate the heme and iron of the enzyme which is positioned within the active site. Employing GROMACS 2018-6 on the GPU queue at the CHPC, the simulations spanned an extensive 300 nanoseconds (ns), underpinned by the AMBER ff14SB force field for a mimicked representation of the protein's environment. The results of these MD simulations consisted of post-MD calculations, which included RMSD, RMSF as well as the Radius of gyration.

3.3.1 Berendsen-Berendsen barostat (Equilibration-Production run)

The Berendsen barostat is a parameter within MD simulations for controlling pressure by adjusting the volume of the simulation cell [139,140]. Weakly coupling the system to an external pressure bath is how this is done. The strength of the coupling is controlled by a relaxation time, and the barostat works by scaling the velocities of the particles based on the deviations from the target pressure [133]. While the Berendsen barostat is computationally efficient and easy to implement, it has been recently criticised [135].

To comprehensively evaluate the impact of the Berendsen barostat on our protein system, we employed a systematic investigation which we conducted during both the equilibration and production phases of the MD simulations. By doing this, we aim to understand the nuanced behaviour of how pressure control mechanisms influence the structural behaviour and stability of the protein, which can result in us finding valuable insights about the dynamic response of the system to varying pressure conditions. The figure below reveals the RMSD values throughout the 300ns simulation, each was done in triplicate.

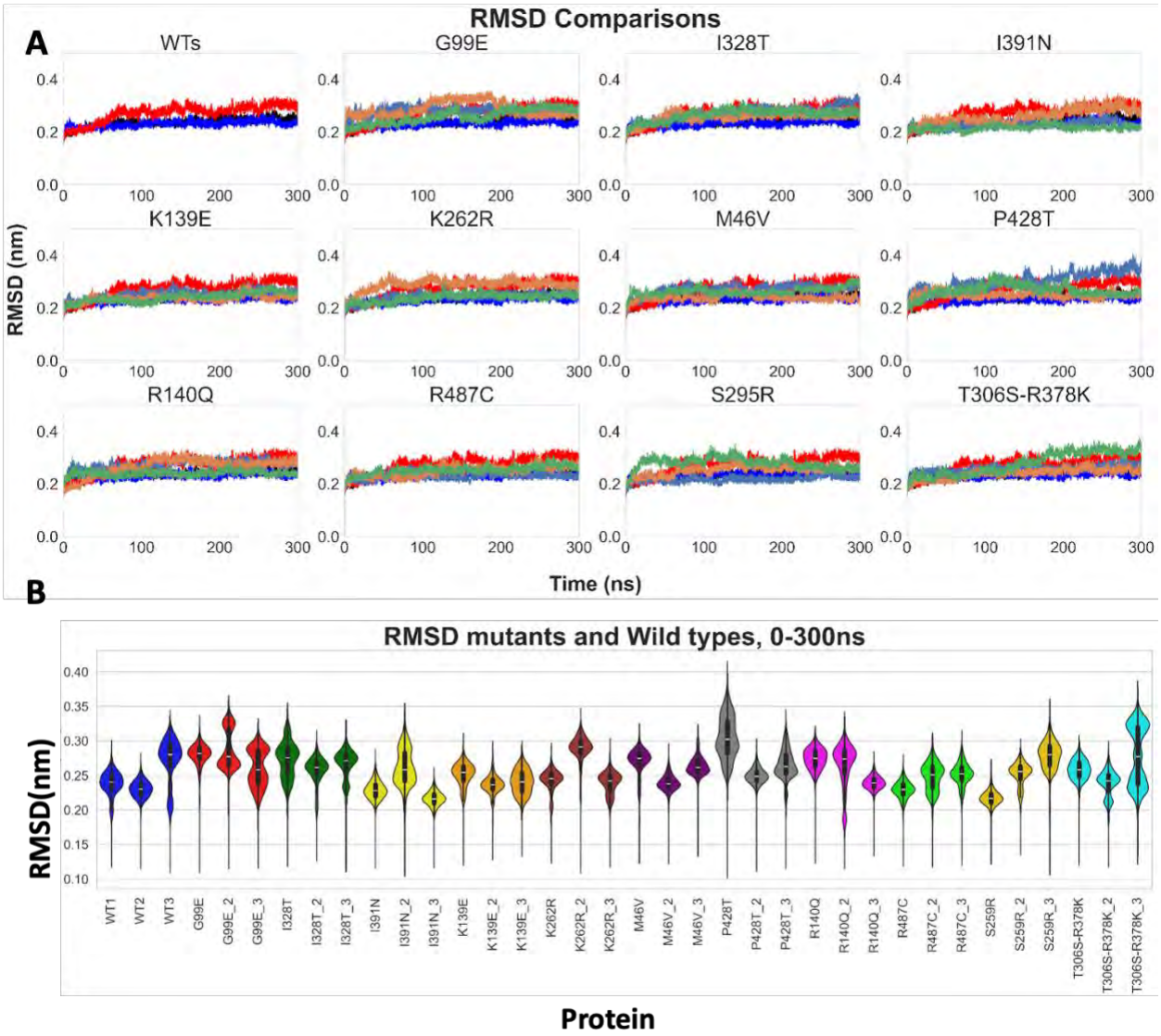


Figure 3.2: RMSD of all systems with incorporated Berendsen barostat for production run. **A)** Line plots showing the RMSD of the protein with respect to Cα for Reference and Variant Systems. Black: Reference 1, Blue: Reference 2, Green: Reference 3, Navy: Run 1, Orange: Run 2. Green: Run 3 **B)** Violin plots representing 0-300ns for all systems, each system is colour-coded.

In evaluating the reference simulations above, a noticeable coherence is discerned between the initial two runs, whereas the third run exhibits higher RMSD values, indicating increased structural deviations, and from observation less stability of the protein system. Despite these variations, a single conformation is observed for the reference systems simulations. Similarly, this can be seen in some of the variant systems during the first and second runs, however, most of the third runs

lack substantial correlation or consistency with the proceeding counterparts. Notably, variations *G99E, P428T, R140Q, S259R*, and *T306S-R378K* display heightened inconsistencies (Figure 3.2A). The violin plots display the inconsistencies between the variation triplicate runs, this is noticed where the conformation or RMSD values can be seen deviating (Figure 3.2B).

Although computationally efficient, the Berendsen barostat has limitations in capturing protein dynamics. A study done by Wong-ekkabut et al., (2016) and Ke et al, (2022) displayed reasons why the Berendsen barostat would be unfavourable [132,133]. Other reasons include the lack of equilibration, as it can cause the system to equilibrate too quickly to the target pressure, potentially leading to non-physical states. This rapid equilibration may not accurately mimic real-life environments. The Berendsen barostat has been shown to exhibit a non-canonical ensemble as it does not sample states from a canonical (NVT) or isothermal-isobaric (NVT) ensemble in a statistically correct manner [138]. This can make it unsuitable for certain types of simulations as the thermodynamic properties are either not calibrated or reached correctly which is perhaps what we are experiencing here. According to another study done by Braun et al., (2019)[139], the Berendsen barostat exhibits artificial damping and is incompatible with certain systems, this may result in incorrect pressure fluctuations with strong pressure-dependent properties which ultimately leads to failure to capture the true dynamics of the system.

The large inconsistencies noticed across the simulations suggested that the Berendsen barostat used for the production run was not efficient therefore downstream application did not go further than calculations of radius of gyration. The RMSD violin plots displaying 50ns-300ns, 100ns-300ns and150ns-300ns (Figure S1), Radius of Gyration line plots and violin plots (Figure S2), including the RMSF line plot and Heatmap (Figure S3), can all be found in Supplementary Material. Across all violin plots in Figure S1, it is observed that the conformations for each system (with replicates) are not consistent, and a wide distribution is observed. The RG line plots suggest that the protein is compact with certain variations displaying differences in their compactness such as: *I328T, M46V*, and *T306S-R378K*. Similarly, the violin plots display the same information however the *M46V* variation is more notable here for its differences in gyration values. The RMSF plots seen in Figure S3 display the deviations seen in flexible loop regions.

3.3.2 Berendsen-Parrinelo Rahman barostat (Equilibration-Production run)

The choice to use the Berendsen barostat for equilibration and the Parrinelo-Rahman barostat for production runs in molecular dynamics (MD) simulations is a deliberate strategy to balance computational efficiency and accuracy. During equilibration, the Berendsen barostat efficiently relaxes the system, this is due to its ability to quickly alter its volume to reach the target pressure [133]. This provides computational speed which is crucial for the initial stages. In the following stage, which is the production run, the Parrinelo-Rahman barostat takes over to ensure a more realistic sampling of the isothermal-isobaric ensemble, capturing pressure-induced effects and avoiding compressibility artefacts associated with the Berendsen barostat [137]. This combined approach enhances the stability, accuracy, and reproducibility of the MD simulations

The figures below represent the protein systems of *CYP2B6*, the references and the corresponding variations. The equilibration was performed using the Berendsen barostat while the production run utilised the Parrinelo-Rahman barostat.

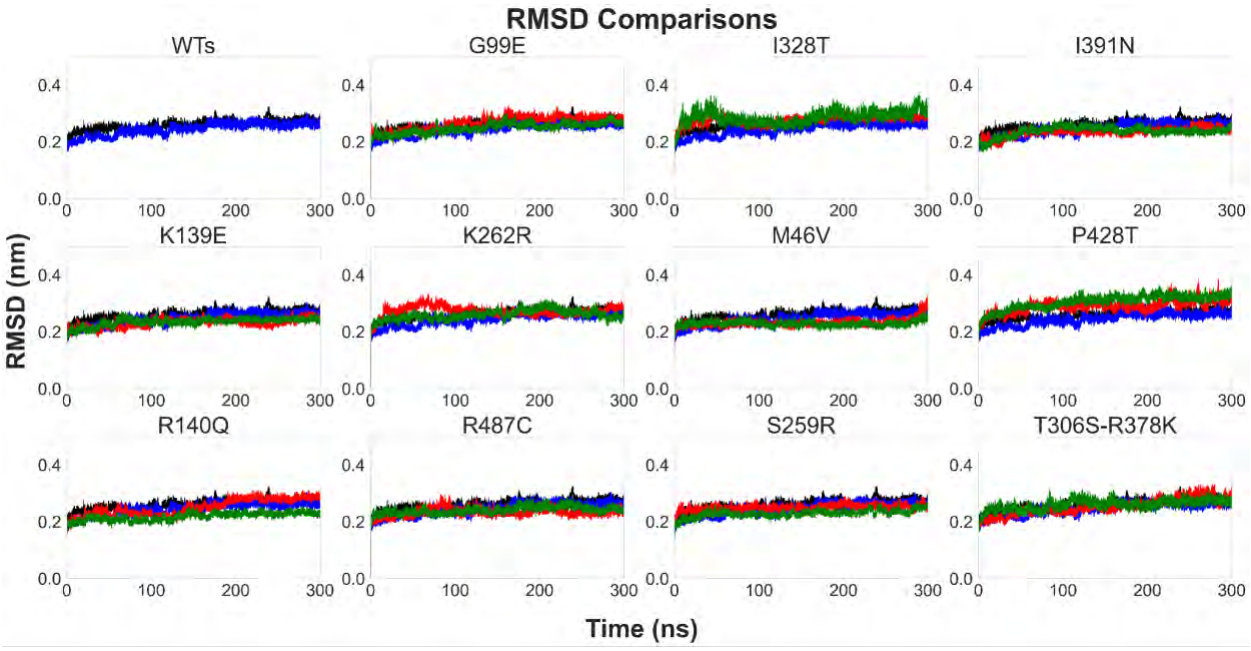


Figure 3.3: Line plots showing the RMSD of the protein, with respect to Cα for Reference and variant systems. Black: Reference 1, Blue: Reference 2, Red: Run 1, Green: Run 2.

Making use of the Parrinelo-Rahman barostat, more stabilised systems throughout were expected compared to that of the results observed in Figure 3.2. From the results shown in Figure 3.3, the line plots suggest that the references exhibit a similar conformational and start to stabilise approximately around the 100-150ns region. They lay over one another, from using the computational resources such as the GPU from the CHPC the results will not be the same.

Figure 3.4 below displays the violin plots of the RMSDs, ranging from 0-300ns, 50-300ns, 100-300ns, and 150-300 ns intervals. Similar observations are noticed with variations *G99E, I391N, K139E, M46V, R487C*, and *S259R* where runs 1 and 2 are in correlation with one another and are stabilising during the last 150 ns of the simulation. Their mean and RMSD values are similar and show a relatively close resemblance to one another. The outstanding variations experience some deviations concerning their first and second runs.

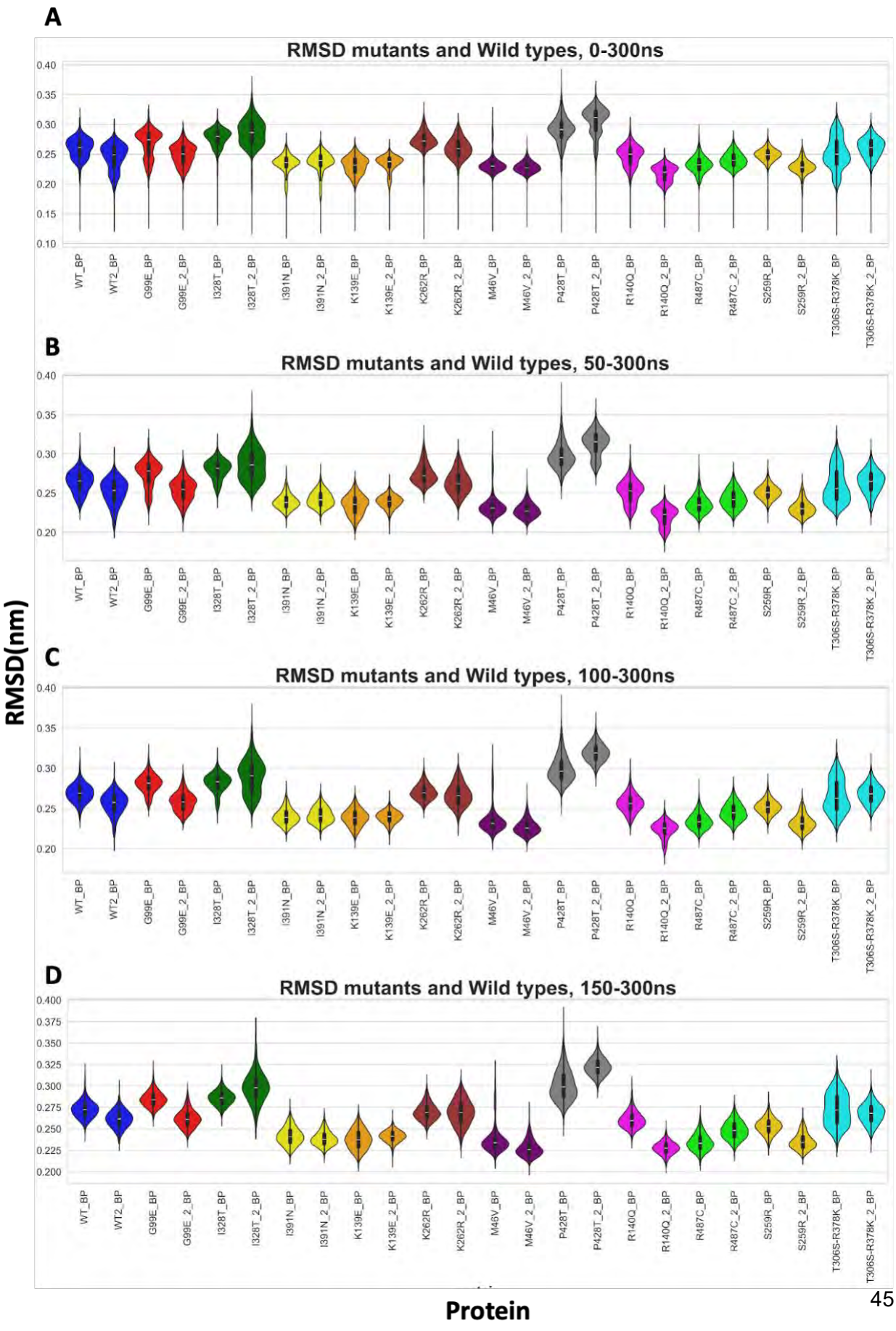

A
RMSD mutants and Wild types, 0-300ns
B
RMSD mutants and Wild types, 50-300ns
C
RMSD mutants and Wild types, 100-300ns
D
RMSD mutants and Wild types, 150-300ns
RMSD(nm)
Protein
WT_BP
WT2_BP
G99E_BP
G99E_2_BP
I328T_BP
I328T_2_BP
I391N_BP
I391N_2_BP
K139E_BP
K139E_2_BP
K262R_BP
K262R_2_BP
M46V_BP
M46V_2_BP
P428T_BP
P428T_2_BP
R140Q_BP
R140Q_2_BP
R487C_BP
R487C_2_BP
S259R_BP
S259R_2_BP
T306S-R378K_BP
T306S-R378K_2_BP

Figure 3.4: Violin plots showing the RMSD distribution. A) represents 0ns to 300ns **B)** represents 50ns to 300ns **C)** represents 100ns to 300ns D) represents 150ns to 300ns. Each system is colour-coded together. The underscore (_2) represents the second run.

The above violin plots support the observations made. According to a study conducted by Braun et al.,[140], most of the variations the authors selected converged except for *G99E, I391N* and *P428T*. The same can be seen with regards to *P428T* but not with the other two variations as they converged and stabilised in our study. It must be noted that the simulations run in the study performed by Braun et al., were 20ns whereas in our study we used 300ns simulations [140]. To investigate these changes in more detail, later in this chapter we explored the opportunities that were offered through RMSF.

3.3.2.1 Observed instability in variations *I328T* and *T306S-R378K*

For the *I328T* variant, both simulation runs initially follow a similar trajectory; however, the second run shows less stabilization than the first, suggesting differences in conformational stability. This difference is illustrated in Figure 3.4, where violin plots across time intervals (A, B, C, D) show that, despite comparable RMSD values, the two runs diverge in their conformational patterns. This discrepancy suggests that I328T may subtly affect the enzyme's structural dynamics, even if it does not alter enzymatic function. These observed differences in the runs indicate that the variation may contribute to distinct structural responses, potentially affecting protein flexibility or stability in a localized manner. This change in conformational behaviour is notable as this variation is associated with a no function enzyme activity. The connection between structural dynamics and functional implications is crucial in understanding the effects of variations on enzyme activity. Therefore, the observed differences in conformational stability in the simulations may be reflective of the variation's influence on the enzyme's functional properties.

Viewing the line plot in Figure 3.3 for the variation *T306S-R378K,* we noticed how the duplicate runs follow the general path of not only the reference but also each other. Both runs seemed to stabilise after 150ns, however roughly 20ns before the end of the simulation for the first run, there was a decrease in the RMSD values which significantly dented the stability that was noticed. The deviation noticed from the two runs suggests a dynamic structural response in terms of changes

either to the protein's conformation or that there are fluctuations that might be occurring locally. This is further supported in the violin plots provided in Figure 3.4, more specifically Figure 3.4D which displays the last 150ns. Here we see that the conformation for both runs are unimodal however the first run shows more of a distant distribution than that of the second run. This is interesting as this variation is associated with no function enzyme activity.

In summary, both the I328T and T306S-R378K variants show differences in structural stability toward the end of the simulations, even though they do not affect enzyme activity. The I328T variant displays slightly reduced stabilization in its second run, hinting at minor conformational shifts, while T306S-R378K shows a late-stage deviation in the first run. These observations highlight that even non-functional variants can influence localized structural behavior, warranting further investigation into their potential impacts on protein stability.

3.3.2.2 Observing the uncertain function variations: *M46V, P428T* and *R140Q*

These variations share the property that they have an overall enzyme function that is uncertain in literature as of now [87]. When observing *M46V* in Figure 3.3, you notice that for the majority of the simulation period, both runs are relatively stable and of the same nature with similar mean and RMSD values, however towards the end of the simulation of the first run the RMSD values are seen to increase in the last 10ns. This may indicate that there was a conformational change that occurred and perhaps the structure was returning to its original structural confirmation. An extension on the simulation is then needed to make a more conclusive suggestion. The findings here are contradictory to what was found in the study done by Braun et al., [140].

Focusing on the *P428T* variation, some curiosity is invited. By first observation in Figure 3.3, the second run stabilises more effectively than that of the first run however they both follow the same general path with similar mean and RMSD values. Although only one run was done, we notice in Braun et al., [140] that this specific variation does not equilibrate, however it is challenging to compare a 20ns simulation with that of a 300ns. Nonetheless, the variation does not equilibrate, which is the case we observe in our study. Upon further observation in Figure 3.4A and B, we noticed that the second run exhibited bimodal distribution. This indicated that for this specific run, two distinct conformations occurred within the whole simulation. The first conformation likely

occurred in the first 100ns and the second conformation occurred in the last 150ns. This is supported by Figures 3.4C and D where the confirmations for the last 200ns and 150ns display unimodal distribution. Seeking insights into the RMSF can help identify any local structural changes that possibly occurred.

The difference occurring in *R140Q* is of interest as the first run stabilises throughout the simulation while the second run seems to be less stable as there is no distinct equilibration that occurs. This is quite interesting as they share the same RMSD values for the first 150ns, thereafter differ, not in conformation shape but in RMSD values. This can suggest that the structure shows slight structural deviations from one another but still share a lot of characteristics. Keeping in mind, that these are duplicate runs, it also suggests that this variation may affect the structure of *CYP2B6* as it causes the protein to behave unpredictably.

3.3.2.3 Deviations noted in *K262R* variation.

Similarly, what was noticed for the *I328T* variation can be observed in the *K262R* variation but with deepened deviations occurring. In Figure 3.3, we see that they follow the general path of the reference and the duplicates themselves lay over one another, however, they cross over initially and then again later in the simulation. They both deviate at the beginning of the simulation, displaying no clear sign of the system stabilising which gives rise to the possibility of fluctuations. This can also be viewed in Figures 3.4A, B, C and D where throughout the simulation, both runs maintain a single conformation that is slightly different in shape. They share similar means and RMSD values, but the second run exhibits a more widely distributed model. Unlike the other variations, *K262R* is associated with increased enzyme function, from the finding so far, it may suggest that this variation induces a structural change, affecting the protein more flexibly and adaptively. This possible change leads to further analysis and experimentation that would need to be done to fully understand the implications of these dynamic patterns observed and make a more conclusive claim.

Overall, there are only a few variations that have known enzymatic activity. These are *K262R* with an increase in enzyme activity, *R487C* with neutral enzyme activity and, *G99E, I328T* and *T306S-R378K* which exhibit no function with regards to enzymatic activity. The rest of the variations are

known to be uncertain. According to Kobayashi et al., (2014) and Honda et al., (2001) [136,137], the variations *M46V, I391N* and *R140Q* are associated with lower enzyme activity while *P428T* is associated with no functional enzymatic activity. This differs from what is recorded by Desta et al., (2021) [87]. At this point, we cannot conclusively decide on the variations that are unknown, but we can have some direction from what the line plots and violin plots provided in Figures 3.3 and 3.4 are suggesting.

From the information acquired, the below figure (Figure 3.5) displays the enzymatic activity gathered from Desta et al., study in the form of violin plots [87]. This offers information on how variations associated with uncertain enzymatic activity react and share similarities with other enzymatic functions. The information gathered is then compared to the results found in this study to observe if any correlation can be found between the studies.

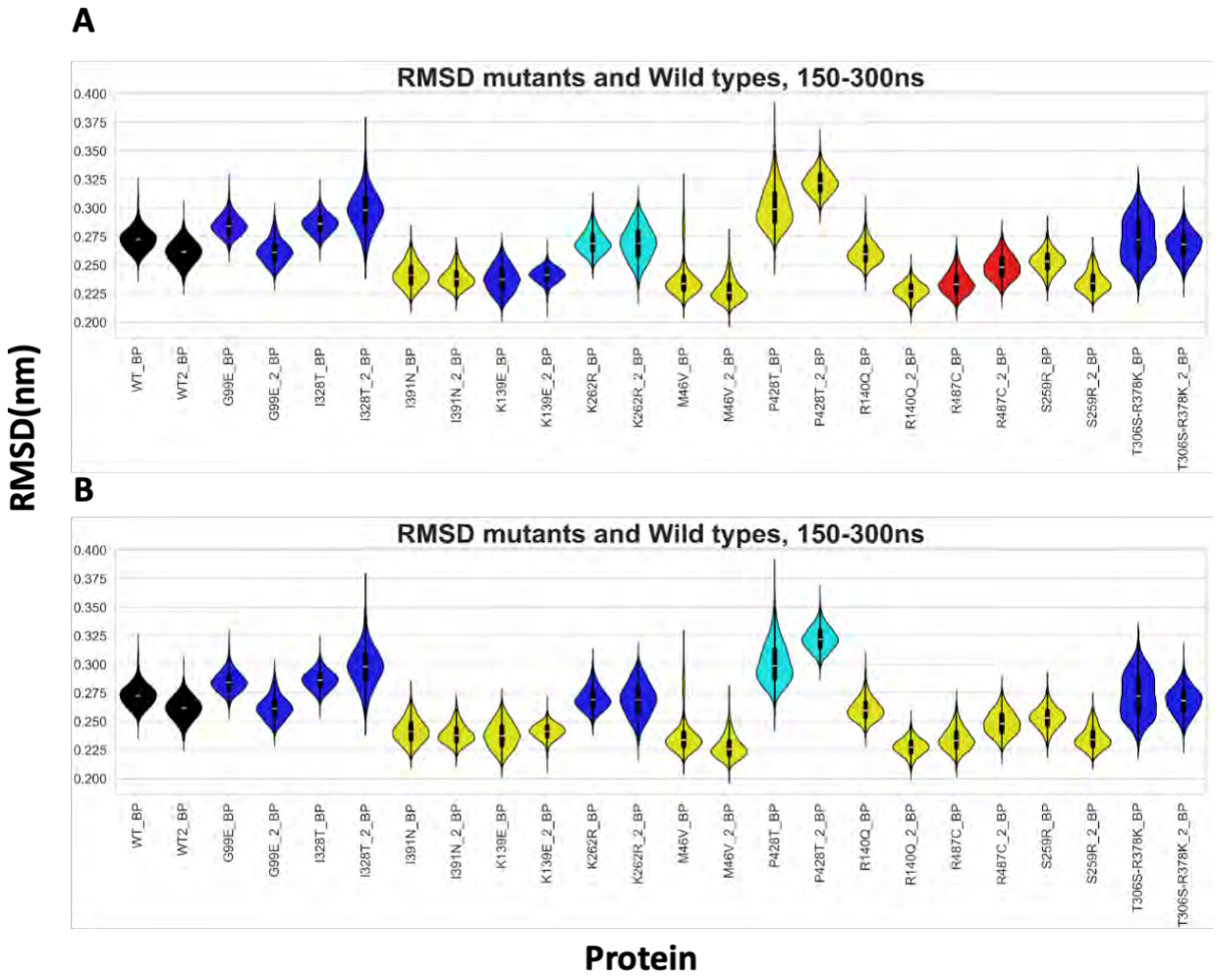


Figure 3.5: RMSD Violin plots representing 150-300ns of the simulation for all systems. The variation pairs as well as the references are colour-coded to signify increased, no function, neutral, and uncertain with regards to enzymatic function. **A)** Reference and variations represented as the enzymatic functions explain them. **B)** Reference and variations represented as the outcome from MD simulations. References are in black, no function variations are in blue, neutral variations are in red and uncertain enzymatic function variations are represented in yellow.

With assistance from Figure 3.5 above, displayed are the last 150 of the simulations across all the systems. As most of the variations are uncertain, we look to head in the direction that is provided to us by the aforementioned figures. The figure above displays the reference in black and the variations associated with no function for enzyme activity represented in blue, variations in yellow represent the uncertain variations, and the variations in cyan represent the variations that resemble

increased enzyme activity. The single variation in red represents the variation that is associated with neutral enzymatic function.

Focusing on Figure 3.5B, this is a hypothetical and interesting observation that can be noted. The colour coordination here is motivated by the similar RMSD values observed and where they fall in terms of the violin plots. Interestingly, most of the variations that are associated with uncertain enzyme function, are found to be with lower RMSD values except for variation *P428T*. This is partially coordinated with what was found in Honda et al., (2011) [141], as some of the uncertain variations were associated with lower enzymatic activity. In addition to this, *K262R* is associated with increased enzymatic function, however, the RMSD values are in similar accordance with that of the wild-type and the variations associated with no function enzymatic activity. Therefore, it was interesting to see how the variations of violin plots align with other variations and what if any sights can be offered as most of the variations are associated with uncertain enzymatic activity. This is not conclusive, however, from the data presented above, it would be interesting to investigate when bound to a substrate if similar results occur within complexes.

3.3.2.4 Assessing the compactness of all the systems

A significant property that if often looked past, is the compactness of the protein which assists in discovering impacts on structural stability. This is usually done by evaluating the spatial distribution of atoms within a protein which usually offers insight and information about the overall shape and dynamic behaviour [142]. In the context of our work, Rg is a valuable parameter that enables us to quantitatively estimate the compactness of *CYP2B6* and its relevant alleles [143]. Typically, the reference, like in most cases, sets the baseline for those to follow. Usually, a lower Rg value represents a more extended or open confirmation whereas a larger Rg value denotes the inverse - a more closed confirmation [143]. Assessing the Rg is crucial in interpreting the impact of variations on the structural dynamics of proteins. The accompanying figure below depicts the Rg line profiles as well as the violin plots for the reference and Variant systems.

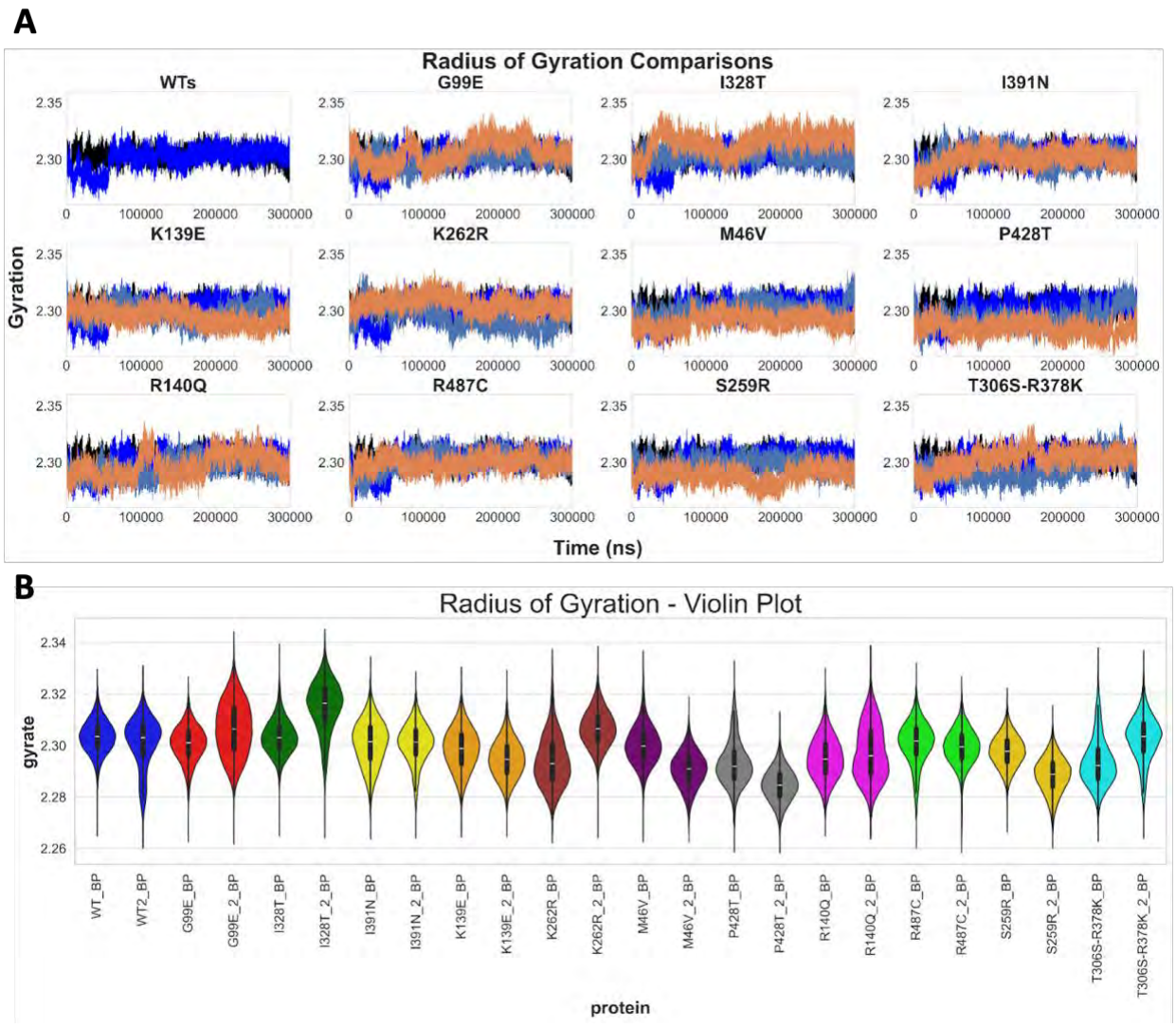


Figure 3.6: Radius of Gyration (Rg) profiles on the enzyme CYP2B6 - References and Variant Systems **A)** Line plots displaying the compactness of all systems. Reference1: Black, Reference2: Blue, Run1: Navy, Run2: Orange. **B)** Violin pots assessing the Rg conformation throughout the simulation - each variation and reference pair is colour-coded together.

The figure above displays the Rg values in two different representation styles, line plots and violin plots. Both give valuable insight, as for the reference in Figure 3.6A, the deviation of the second run from the first run is seen at the beginning of the simulation however the overall protein structure is compact throughout the simulation and is supported in Figure 3.6B. The same is seen throughout the Variant systems, in some cases the line plots show more deviations as the Rg values fluctuate between fixed levels as observed. This can be seen in the *G99E* variation system, where the line plot in Figure 3.6A displays the second run deviating from that of the first run, however

when observing the violin plot, it is noticeable that the mean is very similar, and that shape is slightly more elongated. This suggests that although differences are occurring the protein structure still can maintain its compactness and is relatively stable. Similar findings can be seen with variations *I328T, K262R, M46V* and the double variation *T306S-R378K*.

3.3.2.5 Investigating local changes

Understanding a protein's functional qualities and interaction capacities requires investigating its structural dynamic flexibility. RMSF is a useful statistic for investigating local fluctuations in atomic locations inside a protein during an MD simulation [144]. RMSF provides a nuanced perspective of residue-level variations, emphasising areas of increased flexibility or stability. This study enables us to identify key amino acids or structural motifs that are critical in the protein's response to its environment. In the following figure below, RMSF profiles and the accompanying heatmap reveal the complexities of local alterations, providing insights into varying degrees of flexibility between the variations and the reference. Interpreting RMSF data helps identify functionally significant areas and determine how variations or environmental variables affect the protein's conformational dynamics [143].

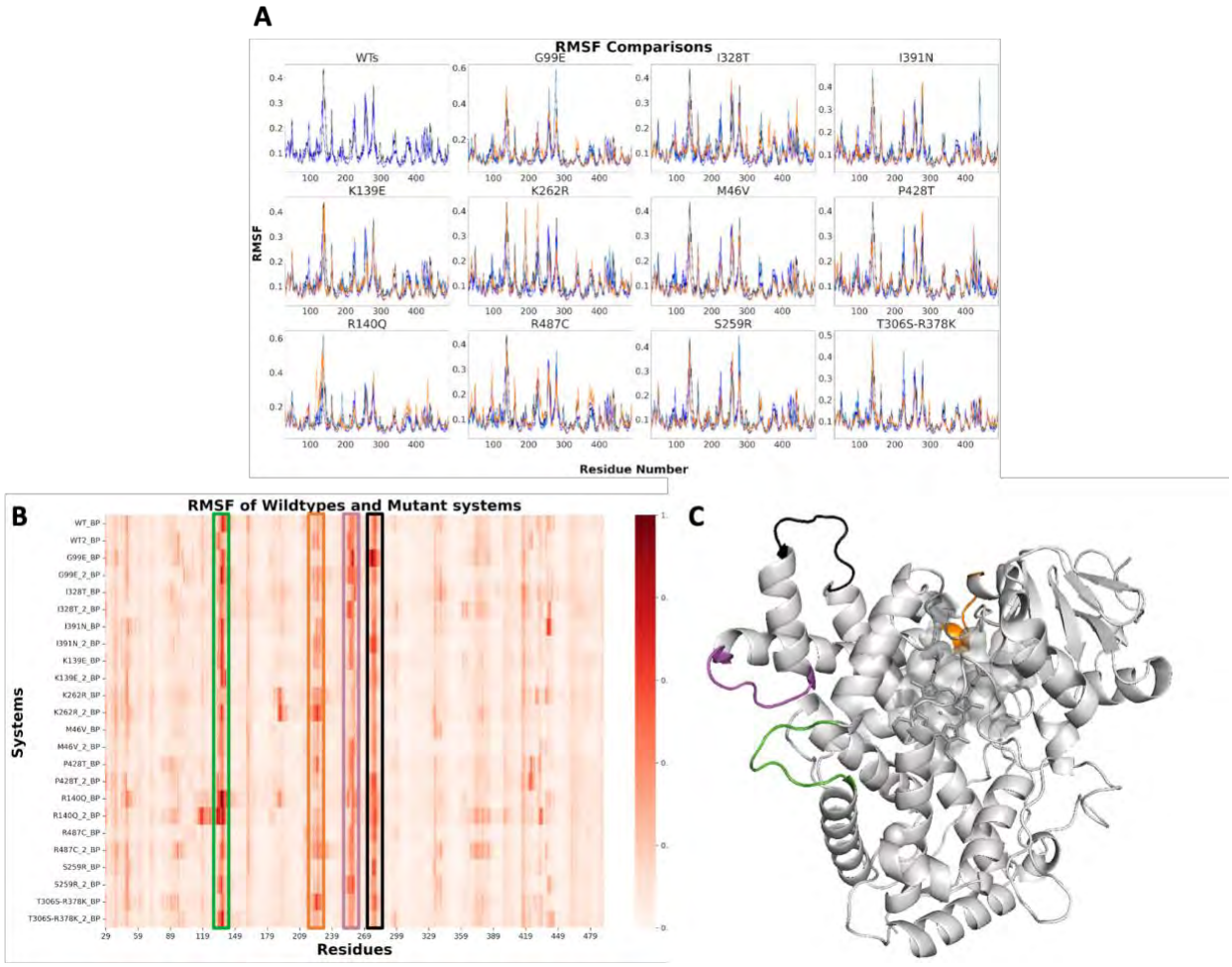


Figure 3.7: Root mean square fluctuations (RMSD) profiles for all systems of *CYP2B6*. A) Line plots displaying RMSF values. Reference1: Black, Reference2: Blue, Run1: Navy, Run2: Orange. **B)** In The RMSF heatmap, regions of higher fluctuations are of darker shades. **C)** PDB 5UFG structure displaying loop areas related to the heatmap.

Assessing the fluctuations that occur within the local regions of *CYP2B6's* structure is better examined through RMSF. This is displayed in Figure 3.7 above where Figure 3.7A displays all the systems and is accompanied by a heatmap in Figure 3.7B. This allowed us to observe if certain fluctuations occurred in one system or across all systems. Initially, four regions stand out immediately at first glance, as the darker shade of red is noticed across the variations. These regions are identified as loop regions which are denoted in Figure 3.7C. These loops are the C-D loop, F'-G loop, G-H loop and the H-I loop, which play are all important to the *CYP2B6* enzyme, the C-loop plays a significant role in substrate binding, the F'-G loop assists in shaping the active site, the G-H loop maintains structural integrity and the H-I loop influences shape and dynamics

of the active site [50,51]. The reason this area is seen with intensified bands throughout all the systems is because these loops are known to be flexible therefore slight deviations in these loops regions are expected. This is due to the instance when the enzyme happens to bind to a substrate; its flexibility around these loops allows for the active site to appropriately accommodate and bind.

Another two instances are seen within the heatmap, the first instance is around residue 120-135 where there are intensified shades of red for both runs of *R140Q* that do not appear in any other variation or reference. The second instance is among the *K262R* duplicate runs around the 185-200 residue region. These regions identified are helix structures that are observed as fluctuating, namely Helix C and Helix F respectively. This observation was supported by Braun et al., (2019)[140] where fluctuations were seen in the C, G, I and L helices. In addition to this, the loop regions as well as certain helices such as F and G, and the plasticity of the *CYP2B6* enzyme is supported by Wilderman et al., (2012)[145]. These are important helices in terms of overall fold and stability as well as substrate recognition and binding. This will be more evident in the chapter to follow which focuses on the clustering section and determining secondary structures (DSSP).

3.4 Conclusion

Chapter 3 delves deep into the exploration of CYP2B6 enzyme dynamics through MD simulations, shedding light on the complex interplay between protein structure, dynamics, and function. The study employed a designed protocol, incorporating protonation, metal parameterization, and utilisation of GROMACS 2018-6 software for extensive 300ns simulations. The primary focus was to investigate the impact that the selected missense variations had on the structure of CYP2B6. In addition to this, investigating the impact of different barostat methods, namely Berendsen and Parrinelo-Rahman, on protein stability and conformational dynamics became a second focus.

The examination of Berendsen barostat revealed notable inconsistencies across reference and Variant systems during production runs. While computationally efficient, this method exhibited limitations, leading to significant variations in protein dynamics, particularly in certain variants. These deviations underscored the necessity for a more robust pressure control mechanism.

In contrast, the combination of Berendsen for equilibration and Parrinelo-Rahman for production runs offered improved stability and accuracy. This strategic approach enabled a more realistic sampling of the isothermal-isobaric ensemble, mitigating compressibility artefacts associated with Berendsen barostat. Consequently, reference and some variants exhibited stabilised conformations, facilitating a better understanding of their structural behaviour.

.

Detailed analysis of individual variants revealed diverse dynamic responses. Variants associated with uncertain enzymatic functions displayed varying degrees of stability, suggesting potential functional implications. Notably, certain variants exhibited consistent behaviour throughout simulations, while others showed significant deviations, indicating structural instability or dynamic changes.

Assessment of protein compactness through Rg analysis provided quantitative estimates, indicating the overall shape and stability of *CYP2B6* and its variants. The examination of local fluctuations using RMSF highlighted critical amino acids and structural motifs influencing protein dynamics and functional properties. Notable fluctuations in loop regions and helices underscored their importance in substrate binding and potential enzymatic activity.

Ultimately, the MD simulations presented in this chapter offer valuable insights into the dynamic behaviour of *CYP2B6* enzyme variants, contributing to a deeper understanding of the intricate relationship between protein structure and function. By elucidating how genetic variations can affect enzyme behaviour, this study paves the way for future research aimed at understanding the functional implications of these variants in the context of personalized medicine. Moreover, these findings have important applications in drug design, particularly in the development of targeted therapies that consider individual genetic profiles and their effects on *CYP2B6*-mediated drug metabolism..

4 Chapter 4: Analysis through Clustering and DSSP

4.1 Introduction

4.1.1 Introduction to Clustering

Cluster analysis is a powerful tool used in data mining and machine learning to group data into clusters based on similarities. For this study, we apply clustering to group conformations from MD simulations, allowing us to identify different structural states within the trajectory,127]. Clustering helps reveal patterns that may not be immediately apparent, offering insights into underlying structural behaviours and functional states of biomolecules, such as the CYP2B6 enzyme. In molecular dynamics simulations, proteins are typically simulated over time to explore their conformational changes and interactions with other molecules. The application of clustering in this context helps identify distinct conformational states or intermediates that emerge during the simulation. This is important because proteins, including *CYP2B6*, are dynamic entities that continuously undergo conformational shifts, which are crucial for their biological function. Clustering aids in distinguishing these transient states from one another, providing a clearer picture of how the protein moves and how its structure may influence its interactions with substrates, inhibitors, or other molecules. Furthermore, clustering can also be used to pinpoint specific regions of the protein involved in key interactions, such as those involved in substrate binding or catalysis, which could be of great interest in drug discovery and understanding protein function[128,129]

An additional benefit of clustering is its ability to reduce the complexity of the data, making it easier to interpret and visualize. In molecular simulations, large datasets containing thousands or even millions of conformations can be overwhelming. Clustering serves as a form of data compression, reducing the dimensionality of the dataset by grouping similar conformations together. This process facilitates a more manageable analysis and can lead to more efficient identification of key structural features. By grouping similar conformations into clusters, researchers can focus on representative structures, making it easier to observe trends and compare structural differences. This is particularly helpful in MD simulations, where data can be voluminous and nuanced, allowing for more streamlined and insightful analyses [149].

In cluster analysis, four important factors need to be considered, distance metric, clustering algorithm, number of clusters and the interpretation of the results [150]. First, the distance metric must be carefully chosen. This metric quantifies the differences between data points, and in the case of protein conformations, it should capture both conformational and orientational differences between molecular structures. Conformational changes refer to alterations in the protein's three-dimensional structure, while orientational differences involve changes in the protein's orientation or position in space. A combination of various metrics, such as RMSD, RMSF, and angular distance measurements, are commonly used to distinguish between conformations. RMSD is often used to compare overall structural differences, while RMSF provides insight into the flexibility of specific regions of the protein, which can be crucial in understanding protein dynamics and function [151].

Another important consideration is the clustering algorithm. Two common types of clustering are agglomerative hierarchical methods and partitional approaches. Agglomerative techniques, such as Ward's method, start with individual data points and gradually merge them into clusters based on their similarity. Partitional approaches, like k-means clustering, partition the data into a predefined number of clusters, seeking to minimize the within-cluster variance. The choice of algorithm depends on the nature of the data and the research question at hand, with each method having its strengths and limitations. Determining the number of clusters is another essential step in clustering analysis. If too few clusters are chosen, important structural variations may be overlooked, while too many clusters may lead to overfitting and unnecessary complexity. To identify the optimal number of clusters, internal validation criteria are employed. These include indices such as the Calinski-Harabasz index, the Davies-Bouldin index, and the silhouette index, all of which evaluate clustering performance [152]. These metrics consider factors such as the dispersion of points within a cluster, the separation between clusters, and the relative density of points within each cluster. The optimal number of clusters is the one that balances these factors to provide the clearest, most meaningful separation of the data.

Finally, once the clustering is performed, the results must be interpreted. This often involves examining the representative structures from each cluster. By analysing these structures, researchers can visualize the underlying differences and similarities between them, often in terms

of their RMSD and RMSF values. The representative structure of each cluster can provide insights into the typical conformation of the protein during the simulation, helping to identify dominant structural states and their potential biological relevance. In the case of the CYP2B6 enzyme, this analysis can reveal conformations that might be functionally important, such as those that favor substrate binding or catalytic activity, and distinguish them from less relevant or unstable conformations..

4.1.2 Define Secondary Structure of Proteins (DSSP)

The integration of DSSP with clustering analysis offers an in-depth perspective on protein structural dynamics, allowing for detailed characterization of protein conformations over time [153]. DSSP is a widely used computational tool designed to assign secondary structure elements to proteins based on specific geometric and chemical properties. Primarily, DSSP determines the presence of secondary structures, such as alpha-helices, beta-sheets, and coils, based on hydrogen bonding patterns and backbone dihedral angles, providing a snapshot of the protein's structural components at any given time [153,154].

While clustering focuses primarily on grouping conformations based on their structural similarities, such as proximity in terms of RMSD or differences within a specific angstrom range. DSSP data provides a complementary layer of insight by detailing the secondary structure composition within each identified cluster. This integrated approach can reveal secondary structural elements at both a global and local level within each conformational state. By mapping secondary structures across clusters, researchers can distinguish not only overall conformational shifts but also smaller, local transitions that may significantly influence the protein's functionality or stability

DSSP can assist cluster analysis by secondary structure assignment where specific structure states such as alpha-helices, beta-sheets or even coils are assigned to each residue within the protein's trajectory [153]. The incorporation of this information allows researchers to identify different structural motifs that occur locally, not just globally. The changes observed in DSSP can reveal any changes in associated elements (secondary structures) throughout a molecular dynamic simulation. The integration of cluster analysis and DSSP allows for the identification of clusters

that exhibit similar patterns. For example, an alpha-helix observed in one cluster may transition to a coil in another, highlighting regions of structural flexibility and the possibility of functional implications tied to these transformations

From a functional perspective, DSSP can be instrumental in linking structural motifs to biological behaviour. In some cases, protein function is closely tied to specific secondary structure elements [154]. For instance, the presence of beta-sheets may influence the protein's stability or binding affinity, while alpha-helices may contribute to structural rigidity or facilitate interactions with other molecules. By combining DSSP data with clustering, researchers can investigate how specific secondary structures correlate with particular clusters, revealing whether certain structural motifs are prevalent in one or more clusters or if they are unique to a specific conformational state.

In essence, DSSP incorporation with cluster analysis enhances the characterisation of protein dynamics by considering not only the overall structural stability but also the small variations that occur in local regions within proteins (local secondary elements) [154]. A more functional implication of different protein conformations is understood when these two programs are integrated.

4.2 Methods

4.2.1 Clustering

For this study, clustering was performed using AmberTools, specifically the *cpptraj* tool, making use of the RUBI, Rhodes University Jango Server. A *cluster. in* the file was utilised to investigate the conformational dynamics using the RMSD as the metric scale. The residues specified the heme Fe iron and the L-helix were used as another metric. The *cpptraj* tool executed the clustering algorithm (hierarchical agglomerative) which gave an output of three clusters. Each cluster displayed the percentages at which it was represented throughout the simulation and the number of frames that it represented; this was found in a produced summary *.da*t file. Each representation was further placed into its trajectory file for further downstream processing applications when needed or intended. The clusters that represented the largest portion or a percentage higher than

>10% were used for 3D visualisation and superimposed, in PyMOL [155]. The results were formulated into a table using Microsoft Excel.

4.2.2 DSSP

Secondary structural elements were determined using Rhodes University's Jango cluster at RUBI, which integrated GROMACS and the GROMACS DSSP tool (*gmx dssp*). The DSSP analysis was performed using a GROMACS-generated *.tpr* file and the *dssp_analysis.py* script written by Ms Chiratidzo Chamboko. To make subsequent processing easier, the binary *.tpr* file was converted to a *.dat* file, which was subsequently converted into a more accessible *.csv* format. Microsoft Excel was used to graphically represent and analyse secondary structural features by producing a stacked bar graph.

4.3 Results and Discussion

The results from the previous subsections gave rise to structural dynamics that could be of importance. This ultimately led to the exploration of clustering being employed using AmberTools and visualising through PyMOL. By characterising these structural ensembles, we look to discern patterns in the proteins' behaviour, identify dominant conformations and assess if any of these variations have a significant impact on the overall structural landscape of *CYP2B6*. Previously, we highlighted the two variations, *K262R* and *R140Q* as more fluctuations were observed in the RMSF line plots and heatmap. The following tables and figures, along with supplementary data, illustrate how clustering analysis serves as a vital tool for uncovering deeper insights into the dynamic nature of *CYP2B6's* overall structure.

Table 4.1: Cluster analysis results. Displaying the results of clustering, the number of clusters each system individually has and the fraction of which each percentage of frames for the simulation belongs to each cluster.

	Clusters	Fraction Percentage (Cluster 1)	Fraction Percentage (Cluster 2)	Fraction Percentage (Cluster 3)

WT	2	0.554	0.426	0.02
WT_2	1	1	0	0
G99E	1	0.94	0.06	0.222
G99E_2	3	0.513	0.264	0
I328T	2	0.53	0.47	0.113
I328T_2	3	0.646	0.214	0.001
I391N	1	0.987	0.012	0
I391N_2	1	0.993	0.007	0
K139E	2	0.63	0.37	0.002
K139E_2	1	0.993	0.005	0.001
K262R	1	0.955	0.045	0
K262R_2	2	0.896	0.104	0
M46V	1	0.907	0.092	0
M46V_2	1	0.999	0.001	0
P428T	2	0.815	0.185	0
P428T_2	1	0.999	0.001	0
R140Q_	1	0.998	0.002	0
R140Q_2	2	0.773	0.222	0.004
R487C	1	0.998	0.001	0.001
R487C_2	2	0.778	0.222	0
S259R	2	0.691	0.294	0.015
S259R_2	1	0.902	0.098	0

T306S-R378K	1	0.992	0.007	0.001
T306S-R378K_2	2	0.823	0.177	0

Table 4.1 above summarizes the number of clusters generated from the cluster analysis procedure. A limit of three clusters was set, with each cluster consisting of frames representing the entire MD simulation trajectory. The frames were divided into the three clusters, the first cluster typically containing the highest number of frames and reflecting the most dominant conformation observed during the simulation. Only clusters containing more than 10% of the simulation frames were considered significant. The resulting number of clusters is shown in the second column of Table 4.

Clustering was performed on all the variant systems including the reference systems. The visual representations of the references were superimposed with each variant can be found in Appendix A, along with a table of RMSD values for these superimposed structures . The superimpositions of variations *I328T, K262R, P428T* and *R140Q* are presented later in this section.

4.3.1 Variations *K262R* and *R140Q* clusters exhibit slight deviations.

Variations *K262R* and *R140Q* were of particular interest due to their RMSD and RMSF plots, which suggest potential increases in structural deviations and fluctuations. Clustering analysis results for these variants are presented in Table 4. The first run of *K262R* revealed only one cluster which was 95.5% and the second run generated two clusters at 90% and 10% respectively. A similar observation was noticed for *R140Q*, where the first run yielded one cluster of 99% and the second run yielded two clusters of 77% and 22% respectively.

For comparison, clustering was also performed on the reference systems. This revealed that the first run had two clusters of 55% and 42% while the second run had a single cluster of 100%. The figure below displays the superimposition of the references, thereafter the superimposition of the references and the respective variation to identify any underlying features that may affect the structural dynamics of *CYP2B6*.

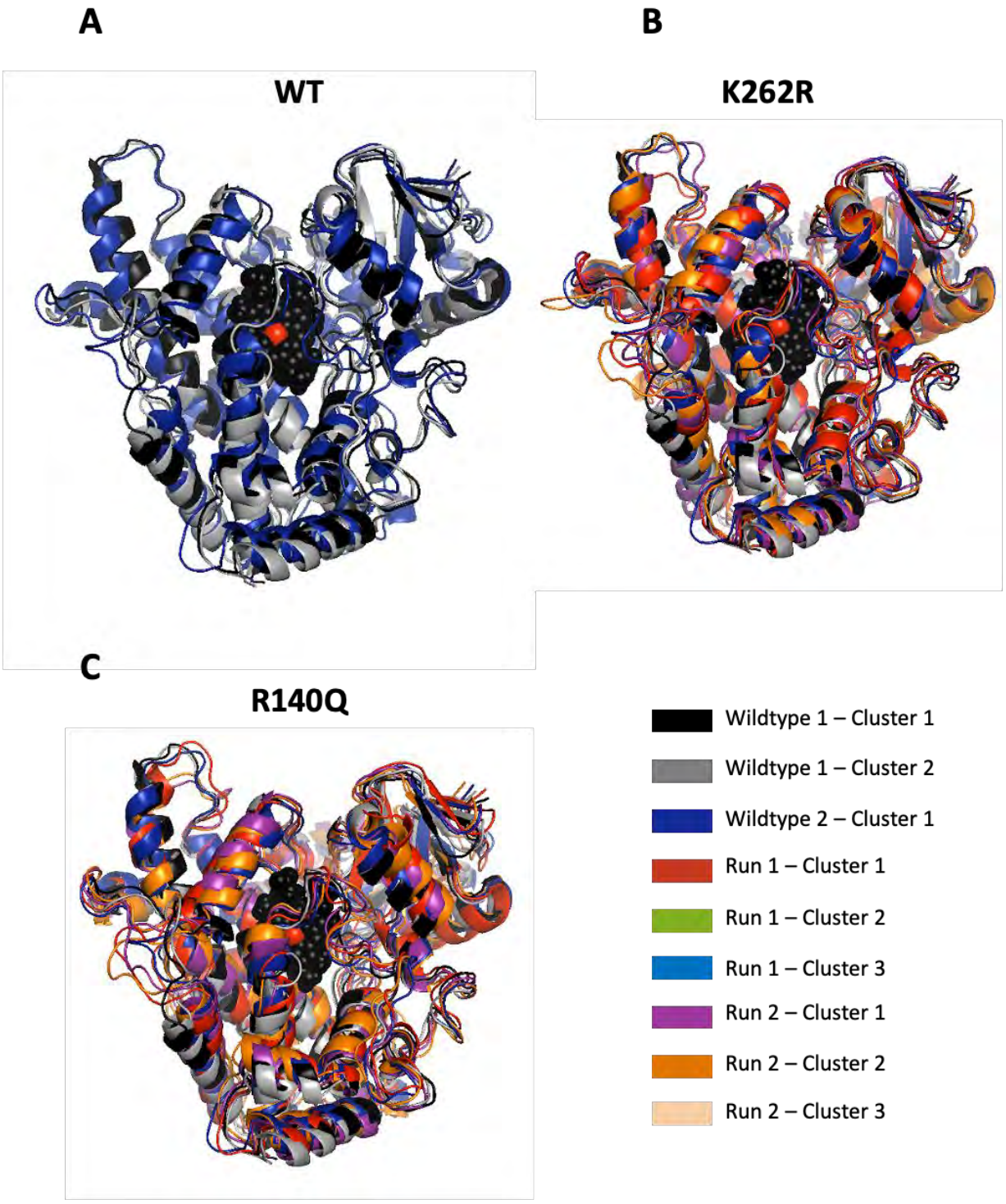


Figure 4.1: Clustering on the references and duplicate runs. A) represents the reference clustering **B)** represents the variation K262R and **C)** represents the variation R140Q. The heme is represented in black spheres with the iron represented by a red sphere that can be identified in the middle of the heme.

Figure 4.1A reveals the reference sytems superimposed. Following the figure legend provided, we noticed that the second reference run deviated from the first reference run. However, this was expected as their post-MD calculations differed slightly. As there were two clusters for the first run, the second run was superimposed on the first run cluster to evaluate the RMSD values. The RMSD values for each superimposition are provided in the Supplementary Material, with variations between 1.2 and 2.3 Å across the reference run. According to Kim et al.,[147] and Reva et al.,[152] RMSD values between 2 and 3 Å in proteins and biomolecules are generally considered structurally similar. These values indicate slight deviations from the reference structure while maintaining the overall orientation, which may result in minor structural shifts, as shown in Figure 4.1A It is when the RMSD is larger than 3 (>3), we would identify the structures as being completely different [156]. Following this we then superimposed the reference clusters with that of each of the variant structures (clusters) (Figure S4).

Figure 4.1B and 4.1C illustrate the K262R and R140Q variants, respectively. While minor changes are apparent, the RMSD differences fall within the range considered structurally similar. Given the indications from post-MD calculations, specific areas of interest were examined further. The figure below provides a closer view of these regions, as identified through RMSF line plots and the heatmap

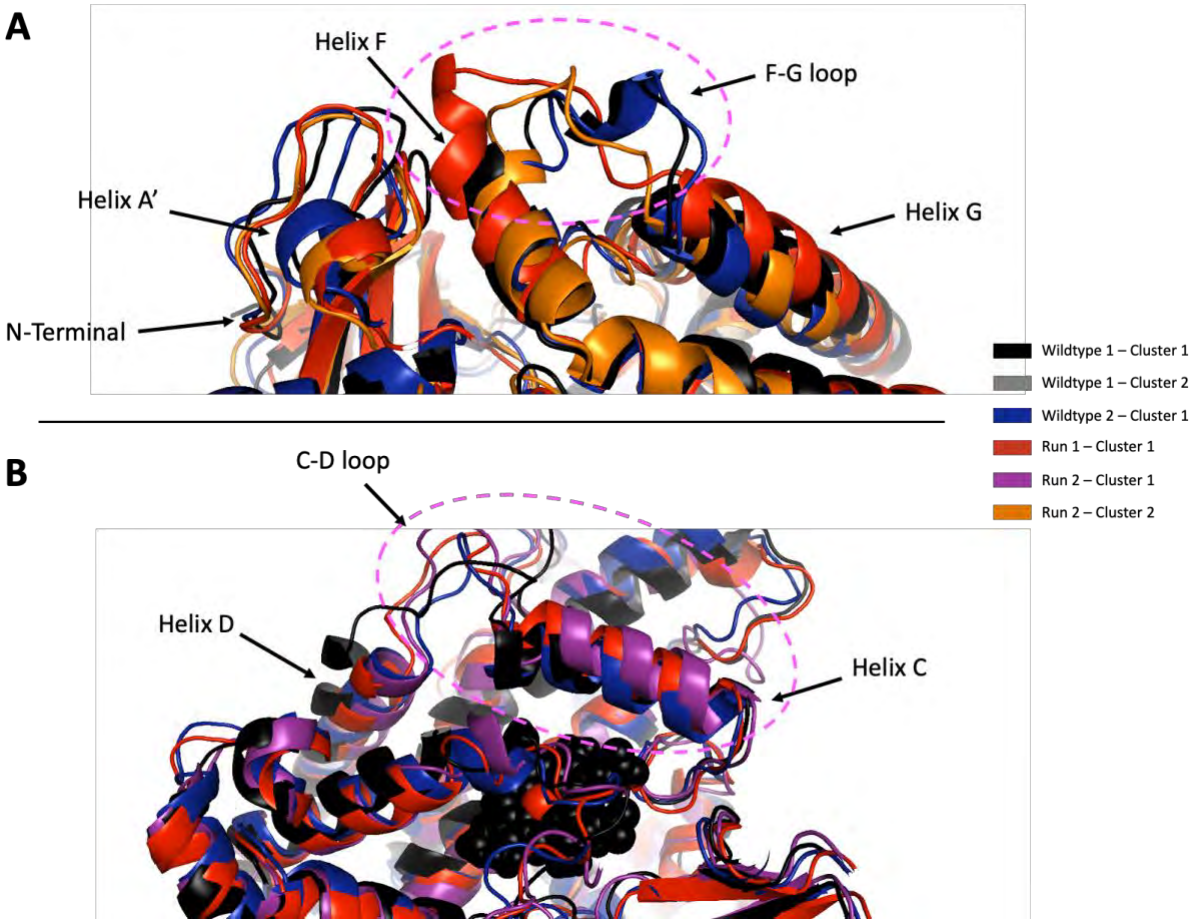


Figure 4.2: Zoomed portions of clustering. A) variation *K262R*, displaying the change in F' helix and F-G loop deviations as well as helix G protrusion, **B)** variation *R140Q*, displaying the extension and extrusion of helix C leading to changes in the C-D loop and deviations in helix D.

For both variations in Figures 4.2A and B, the clusters with the highest fraction were considered. Focusing on *K262R* in Figure 4.2A, the difference in RMSD values were between 1.7-2.3Å. Notable changes include an extension of the F' helix and a slight extrusion of the G helix (shown in red). Although these changes are subtle, they are significant because they do not appear in any other variations. Since these helices play a crucial role in substrate binding and recognition, their alteration may affect the active site pocket[50,51]. Further study involving substrate binding would need to occur to investigate if this observation is sound. In addition to this, the deviations seen in the F'-G loop are to be expected as it is considered a flexible region [147,152], However, the loss of the helix may affect the loop's flexibility, potentially influencing the structural deviations seen in the adjacent helices.

When looking at *R140Q* in Figure 4.2B, the region of particular interest observed in RMSF line plots and Heatmap pointed in the direction of the C Helix. Observing the figure, we noticed that the purple and red clusters are similar and deviate from the reference. The RMSD difference for *R140Q* ranges from 1.9 to 2.3 Å. The structural difference observed is the protrusion of the C helix towards the exterior, while helix D shifts inward. Helices C and D are critical for maintaining the protein's overall fold and contributing to the formation of the active site pocket. Alterations to these helices may impact the C-D loop, which is important for substrate specificity and catalytic activity in *CYP2B6*. These findings suggest that *R140Q* could have a more substantial structural impact on *CYP2B6*, potentially altering enzyme or catalytic function [147,152].

Slight deviations are also observed in the B helix (Figure 4.2A), which may indicate a change in the flap region responsible for substrate entry, potentially altering substrate binding and specificity. While these findings are not definitive, they suggest potential effects that warrant further investigation.

Although structural differences between the systems were anticipated based on post-MD calculations, no significant structural changes, such as the loss of secondary structures, were observed. Given that this study focuses solely on the variations' effects without assessing the variations in conjunction with a bound substrate, evaluating enzyme function at this stage is not feasible.

4.3.2 Variations *I328T* and *P428T* clusters display differences.

Clustering of all the systems revealed that variations *I328T* and *P482T* were particularly interesting due to the higher number of clusters observed, indicating a greater diversity of conformations throughout the MD simulation. The figure below illustrates the clustering of these variations, superimposed with their respective clusters and the reference structure for comparison.

.

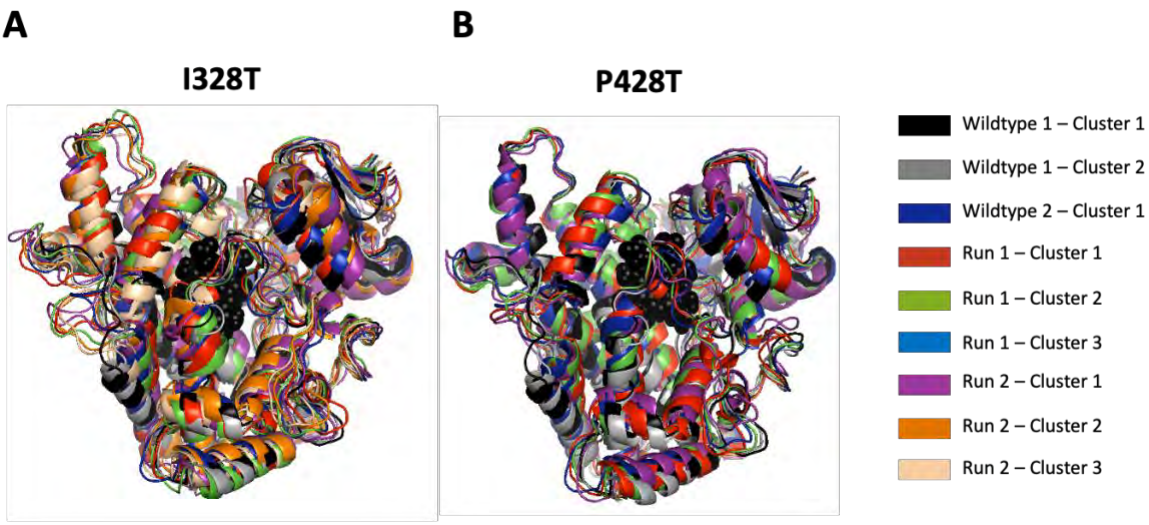


Figure 4.3: Clustering on the references and duplicate runs. A) represents the variation I328T and **B)** represents the variation P428T The heme is represented in black spheres with the iron represented by a red sphere that can be identified in the middle of the heme.

The above figure represents the clustering performed for variations *I328T* and *P428T*, in Figure 4.3A and B, respectively. Following the figure legend provided, it is observed that there are more clusters present within these variations. When focusing primarily on *I328T* we notice that there are large amounts of areas where deviations occur. The difference in RMSD values ranges between 2-3Å. In particular, the B helix appears more open compared to the reference, indicating a potential alteration in the flap region that may impact the substrate entrance channel and recognition site. Similar findings are seen in the C helix, where the helix is both shortened in one cluster and lengthened in another. This supports the suggestion of the flap region being altered however the deviation seen in the B helix has now become of new knowledge as it was not discovered in the studies done by Kim et al.,[147] and Reva et al., [152], which assessed the structural differences and only found the C helix to be deviating from the reference structure.

Similarly, differences in the B and C helix are observed in the variation *P428T*. The extent to which these differences are not identical however they share close resemblance. The difference in RMSD values ranges between 1.1 - 2.8 Å. This suggests that there are structural differences between the reference and the variant structure. These observations are supported in Figures 4.4 and 4.5 respectively.

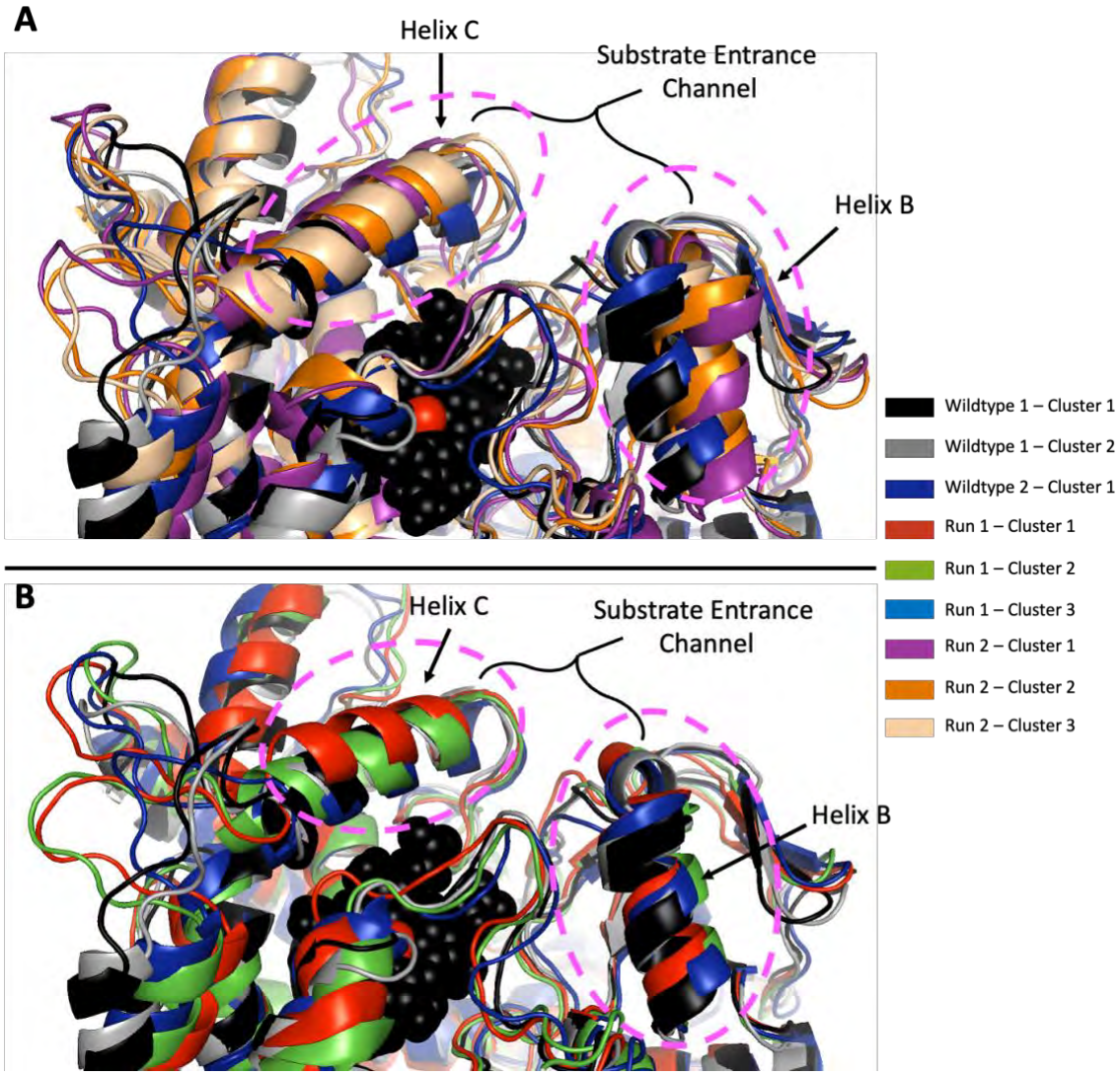


Figure 4.4: Zoomed sections of CYP2B6, displaying the B and C helix, as well as the substrate entrance channel. This represents the I328T variation clusters. **A)** Represents the first runs clusters. **B)** Represents the second run clusters of the variation I328T. The pink dashed circles show the location of the B and C helix, accompanied by black arrows.

The figure above assists in visually displaying the differences found in this variation. The whole trajectory RMSD difference for the duplicate runs of *I328T* was 2.436 Å. This was an indication that structural differences were most likely occurring and that it may be affecting the stability of the protein. We support the suggestion mentioned earlier, that the substrate entrance channel is altered, in this case, it was widened, as the B and C helix is opening more than that of the reference.

Figure 4.4B displays the clusters from the second run, which show similar trends as observed in the first run, particularly in the C helix. Notably, the C helix appears to widen towards its bottom, suggesting that the active site pocket may be enlarged. However, the two runs differ in the behavior of the substrate entrance channel. In Run 1, the substrate entrance channel appears to be widening, while in Run 2, it seems to narrow. These differences highlight the variability in conformational changes between the two runs and their potential implications for substrate binding and recognition.

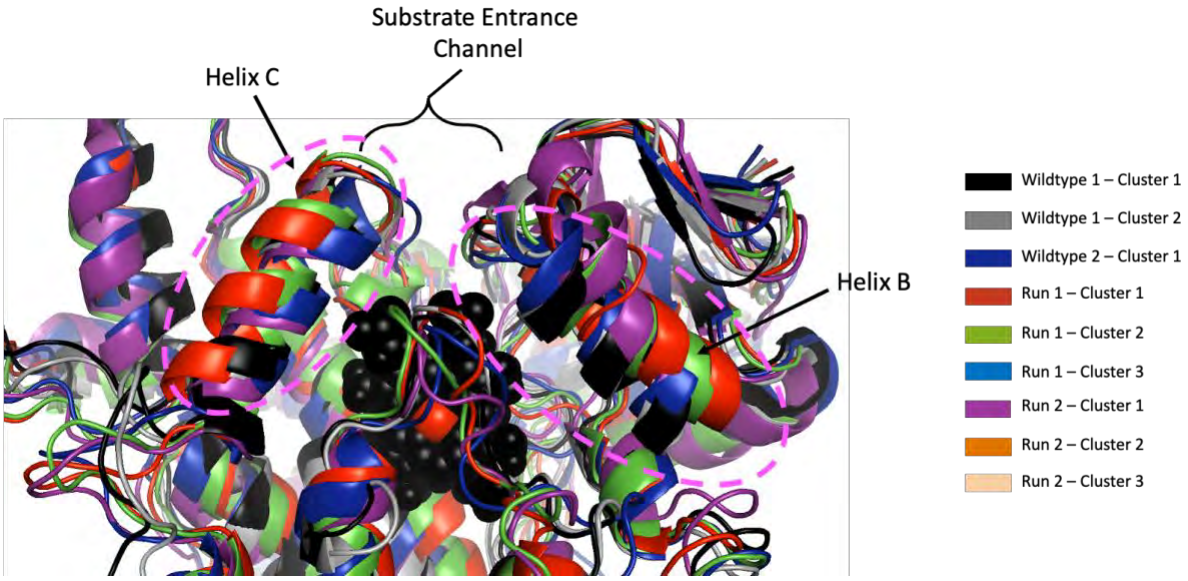


Figure 4.5: Zoomed sections of *CYP2B6*, displaying the B and C helix, as well as the substrate entrance channel. This represents the *P428T* variation clusters. The pink dashed circles show the location of the B and C helix, accompanied by black arrows.

The above suggests that deviations in the B and C helix may affect the substrate binding channel which may result in a list of several other factors being affected, such as the active binding site, the substrate recognition site and the loop region that assist in both substrate recognition and specificity. In Figure 4.5 above, the visual representation displays both the B and C helix deviating from the reference. The bottom of the B Helix points to an outward direction while the top of the C Helix also points to the outward direction resulting in a more closed structure.. The structure deviates and offers insights into future work with substrates to assess if the structural change affects enzymatic activity.

Interestingly, the variations presented in Figure 4.3—specifically *I328T* (associated with no enzymatic function) and *P428T* (associated with uncertain enzymatic function)—raise an important question: Are the observed structural changes dormant with respect to enzymatic function? Further investigation is required to determine if these structural alterations indeed affect enzyme activity, as the results shown in Figures 4.3-4.5 suggest that such changes may have functional implications.

Figure S4, found in the Supplementary material, presents the clustering results for the remaining variations, with RMSD values ranging from 1.2 to 3 Å. While some deviations are noticeable and may influence enzymatic function, it is not uncommon to observe RMSD values between 1 and 3 Å when comparing variant and reference structures. What was noticed was that there were more flexible regions within *CYP2B6*, and the natural variability experienced showed that between the references as well. Kim et al. [147] and Reva et al. [152] confirm that flexible loop regions in *CYP2B6* can influence adjoining helices, and they note that the enzyme's flexibility varies with specific variations. This suggests that local structural changes, such as side chain rearrangements or minor shifts in secondary structure, may account for the RMSD values observed across the variations.

4.4 Identifying changes through DSSP

To investigate secondary structure deviations across the variants, DSSP analysis was conducted. DSSP classifies protein secondary structure elements based on hydrogen bonding patterns and geometric features of amino acid residues [153], using labels such as helix, strand, and coil to characterize segments of the protein structure. This analysis was executed in GROMACS, beginning with the reference structures, shown in Figure 4.6..

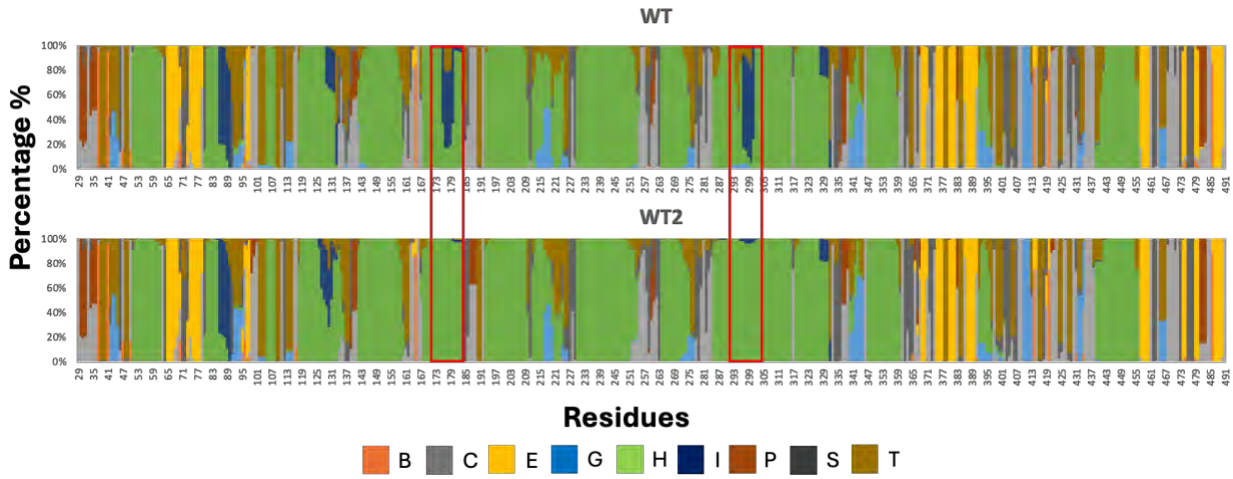


Figure 4.6: DSSP representing references. The colours are associated with secondary structural elements. The B - residue in isolated beta bridge, C - loop region, E - an extended strand that participates in beta-ladder, G - 3_10-helix, I - pi-helix, P - poly-proline II helix (kappa helix), S - bend, T - hydrogen bond turn. The red outlined box represents where the differences occur.

The figure above displays the secondary structural prediction done on the reference systems. From this, it is observed that there is a loss of a bend in two regions for the second run around residues 170-180 and 300-315. These residues represent the E and I helix respectively suggesting that there are differences between the two references regarding the bends on the aforementioned helices. The E helix assists with the active site pocket formation while the I helix is the longest helix found in CYP 1, 2 and 3 enzymes and assists the rigidity of the structure and forms the backbone of the active site [147,152]. Structural deviations in the I helix may impact the electron transfer pathway, affecting the enzyme's redox potential and catalytic performance.

These initial findings from the reference structures provide a baseline for comparison with variants, where unique deviations may reveal insights into variant-specific effects. A structural deviation in a variant that is absent in both reference references could indicate an alteration with potential functional significance. Further analysis of such deviations, particularly in the E and I helices, could shed light on how specific variations influence substrate binding, recognition, and the overall stability of the *CYP2B6* active site, with implications for drug metabolism and enzyme activity.

4.4.1 Insights into DSSP: Deviations across variations

Among the analysed variants, *K262R* exhibited the most extensive structural deviations across multiple assessments. Figure 4.7 provides a DSSP visualization highlighting these deviations in secondary structure elements between *K262R* and the wild-type references. Key differences are observed in several regions, each of which may impact *CYP2B6's* structure-function relationship.

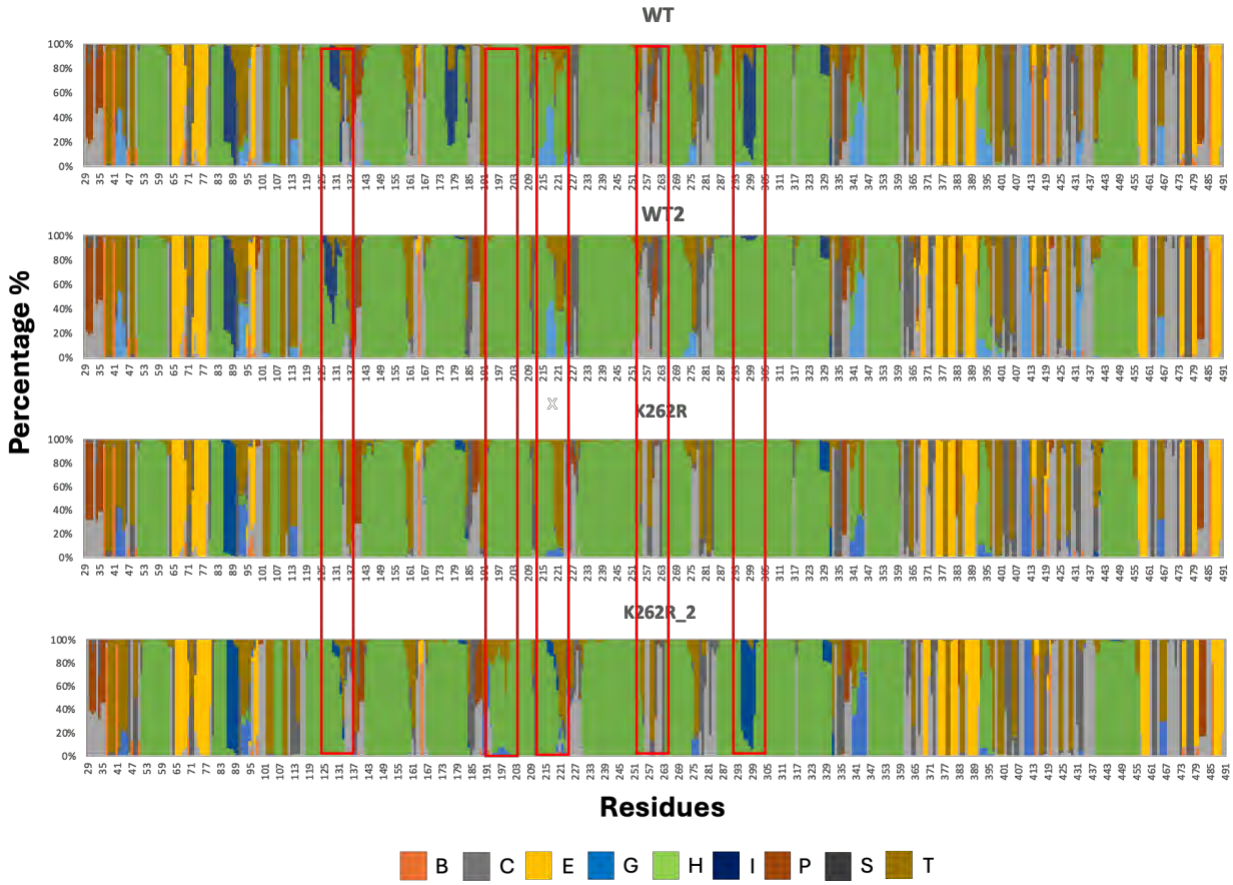


Figure 4.7: DSSP representing references and variation *K262R*. The colours are associated with secondary structural elements. The B - residue in isolated beta bridge, C - loop region, E - extended strand that participates in beta-ladder, G - 3_10-helix, I - pi-helix, P - poly-proline II helix (kappa helix), S - bend, T - hydrogen bond turn. The red outlined box represents where the differences occur.

In the *K262R* variant, structural changes across key regions suggest potential impacts on *CYP2B6's* substrate-binding functionality. Between residues 125-130, both simulations indicate a loss of a bend and an extension of the helix, reaching into the C-D loop region. This modification near the end of the C helix could increase the loop's flexibility, likely influencing substrate entry and

positioning—a finding that aligns with observations from the clustering analysis. In residues 197-200, the second simulation reveals a shortening of the F helix, with a hydrogen bond turn forming in place of the removed segment. Given the F helix's critical role in stabilizing the substrate-binding pocket, this shift might affect substrate affinity and binding orientation. Additionally, residues 218-221 show a consistent loss of the 3_10 helix in both simulations. This structural feature contributes to stabilizing nearby residues and maintaining the active site's overall architecture; thus, its loss could reduce rigidity and impact the dynamic behaviour of the active site. Together, these structural deviations highlight potential functional consequences in substrate recognition and binding, warranting further investigation.

Previous studies by Kim et al. [147] and Reva et al., [152] underscore the dynamic nature of these regions, particularly in variants associated with CYP2B6. Their findings support the idea that structural plasticity in these regions can impact enzyme function, possibly altering substrate recognition or the redox reaction mechanism.

Similarly, in the R140Q variant, the C helix region demonstrates the formation of a 3_10 helix, an anomaly also noted in several other variants (Figures S5-S14). Although individual residues differ, the C helix region consistently shows similar deviations across variations, suggesting that this region's plasticity might modulate the enzyme's catalytic efficiency or substrate specificity.

These structural deviations have notable implications. Variability in regions such as the B and C helices could influence electron transfer and redox reactions, key steps in CYP2B6's catalytic cycle. Understanding these deviations provides a foundation for future studies to explore how specific structural changes impact enzymatic function, which is critical for applications in drug metabolism and precision medicine. Identifying how such structural plasticity affects substrate binding, recognition, and catalysis could inform drug design and aid in predicting the effects of disease-related mutations

.

4.5 Conclusion

Chapter 4 offers a comprehensive exploration of the structural dynamics of the *CYP2B6* enzyme and its variants through clustering analysis and DSSP integration. By employing these methodologies, we gain valuable insights into the conformational behaviour and functional implications of protein structures.

The utilisation of clustering analysis allows for the partitioning of data into clusters based on similarities, facilitating the identification of distinct conformations or states within molecular dynamics simulations. This approach aids in pattern recognition and data compression, enabling researchers to understand underlying structural features and their functional significance. Integration of DSSP further enhances the characterization of protein dynamics by providing insights into local structural changes, such as alterations in secondary structure elements, which may influence protein stability and enzymatic function.

Methodologically, clustering was performed using AmberTools, while DSSP analysis was conducted using GROMACS, enabling detailed examination of structural ensembles and secondary structural elements. Results and discussion focus on variations exhibiting deviations or fluctuations in structural dynamics, with particular attention to regions like active sites and substrate entrance channels.

Notable findings include structural deviations observed in the K262R variant, which showed significant changes in several regions. In residues 125-130, both simulations revealed a loss of a bend and an extension of the helix near the end of the C helix, extending into the C-D loop. This alteration could impact loop flexibility, affecting substrate entry and positioning. Additionally, residues 197-200 displayed a shortening of the F helix in the second simulation, with a hydrogen bond turn replacing the lost helix segment. This change could affect the stability of the substrate-binding region, potentially influencing substrate affinity and binding orientation. Furthermore, residues 218-221 showed a loss of the 3_10 helix in both simulations, which may affect the structural rigidity of the region and alter the dynamics of the active site.

These findings, along with similar structural deviations observed in other variants, highlight how subtle changes in secondary structure elements can have significant functional implications, potentially altering enzyme activity, substrate interaction, and catalytic efficiency. The results emphasize the importance of structural analysis in understanding the functional consequences of mutations and provide a foundation for future studies on enzyme variants, with potential applications in drug design and the study of disease-related mutations

5 Chapter 5: Conclusion and Future Work

In conclusion, this study provides a comprehensive analysis of the effects of Afrocentric missense variations on the structural dynamics of the *CYP2B6* enzyme using molecular dynamics (MD) simulations, clustering analysis, and DSSP integration. This enzyme belongs to a superfamily of enzymes known as CYP P450s and plays its role in metabolism of exogenous and endogenous compounds. The initial aim of the study was to investigate if set out variations that are focused within the African population will have an impact on the structural dynamics of *CYP2B6* which led to the hypothesis that the variations will indeed influence the enzyme.

Modelling the enzyme came with its own challenges as there were variations that occurred and unforeseen variations (*K29A* and removal of *H491*). MD simulations were performed and post MD calculations were accompanied by critical analysis. Revealed in this process was the need for the appropriate parameter utilisation as the difference in two different barostats in MD simulations. The Parrinelo-Rahman barostat yielded more consistent results hence was used for the downstream applications that followed while the Berendsen barostat, when used for the production run, yielded inconsistent results.

The results indicate that certain variants (*I328T, K262R, P428T* and *R140Q)*, particularly exhibiting deviations in RMSD and gyration, show fluctuations in structural dynamics, especially within critical regions such as active sites and substrate entrance channels. These findings raise important questions about the relationship between structural changes and functional outcomes. Variants that display more restricted structural flexibility (e.g., lower RMSD and gyration values) may exhibit more specific substrate recognition, potentially leading to higher catalytic efficiency (Kcat) or binding affinity (Ka/Kd) for certain substrates. However, this could also result in reduced efficiency for other substrates due to diminished flexibility in the enzyme's active site. These observations suggest that variants with similar MD and structural consequences may exhibit similar functional behaviors, but the extent of these effects could vary depending on the specific nature of the structural changes.

A key aspect of this study is the correlation between structural dynamics and known functional consequences of specific variants. The structural metrics analysed here, such as deviations in secondary structures and changes in loop flexibility, can help predict the functional impact of variants on CYP2B6 activity. For instance, mutations affecting the flexibility of critical helices or loops may influence substrate binding and metabolism, with direct implications for drug metabolism and personalized medicine. This reinforces the potential of using structural dynamics as a tool to inform precision medicine, allowing for better prediction of how genetic variations in *CYP2B6* may affect drug responses across different populations

Future studies could build further into the investigation that was set forth, by investigating more missense variations and assessing their response with regards to their impact on structural dynamics of *CYP2B6*. Looking into alternative barostat options such as C-rescale is also seen to yield consistent results. Additionally, examining the active site conformation through comparative component principal analysis. This can shine some insight into how and if the active site is altered. The biggest step to this work, could be investigating a substrate bound enzyme thereafter assessing the effects these variations have on the *CYP2B6* complex. This assists in determining the consequences these amino acid differences have on the pharmacological side of things.

Thank you

6 References

[1] J. L. Jameson and D. L. Longo, “Precision Medicine — Personalized, Problematic, and Promising,” N. Engl. J. Med., vol. 372, no. 23, pp. 2229–2234, Jun. 2015, doi: 10.1056/NEJMsb1503104.

[2] M. V. Relling and W. E. Evans, “Pharmacogenomics in the clinic,” Nature, vol. 526, no. 7573, pp. 343–350, Oct. 2015, doi: 10.1038/nature15817.

[3] W. E. Evans and M. V. Relling, “Pharmacogenomics: Translating Functional Genomics into Rational Therapeutics,” Science, vol. 286, no. 5439, pp. 487–491, Oct. 1999, doi: 10.1126/science.286.5439.487.

[4] E. Micaglio, E. T. Locati, M. M. Monasky, F. Romani, F. Heilbron, and C. Pappone, “Role of Pharmacogenetics in Adverse Drug Reactions: An Update towards Personalized Medicine,” Front. Pharmacol., vol. 12, p. 651720, Apr. 2021, doi: 10.3389/fphar.2021.651720.

[5] D. M. Roden, R. A. Wilke, H. K. Kroemer, and C. M. Stein, “Pharmacogenomics: The Genetics of Variable Drug Responses,” Circulation, vol. 123, no. 15, pp. 1661–1670, Apr. 2011, doi: 10.1161/CIRCULATIONAHA.109.914820.

[6] J. J. Swen et al., “Pharmacogenetics: From Bench to Byte— An Update of Guidelines,” Clin. Pharmacol. Ther., vol. 89, no. 5, pp. 662–673, May 2011, doi: 10.1038/clpt.2011.34.

[7] N. J. Schork, “Personalized medicine: Time for one-person trials,” Nature, vol. 520, no. 7549, pp. 609–611, Apr. 2015, doi: 10.1038/520609a.

[8] S. L. Collins, D. F. Carr, and M. Pirmohamed, “Advances in the Pharmacogenomics of Adverse Drug Reactions,” Drug Saf., vol. 39, no. 1, pp. 15–27, Jan. 2016, doi: 10.1007/s40264-015-0367-8.

[9] H. M. Dunnenberger et al., “Preemptive Clinical Pharmacogenetics Implementation: Current Programs in Five US Medical Centers,” Annu. Rev. Pharmacol. Toxicol., vol. 55, no. 1, pp. 89–106, Jan. 2015, doi: 10.1146/annurev-pharmtox-010814-124835.

[10] I. M. Langmia, K. S. Just, S. Yamoune, J. Brockmöller, C. Masimirembwa, and J. C. Stingl, “CYP2B6 Functional Variability in Drug Metabolism and Exposure Across Populations—Implication for Drug Safety, Dosing, and Individualized Therapy,” Front. Genet., vol. 12, p. 692234, Jul. 2021, doi: 10.3389/fgene.2021.692234.

[11] M. Zhao et al., “Cytochrome P450 Enzymes and Drug Metabolism in Humans,” Int. J. Mol. Sci., vol. 22, no. 23, p. 12808, Nov. 2021, doi: 10.3390/ijms222312808.

[12] K. Mangó, Á. F. Kiss, F. Fekete, R. Erdős, and K. Monostory, “CYP2B6 allelic variants and non-genetic factors influence CYP2B6 enzyme function,” Sci. Rep., vol. 12, no. 1, p. 2984, Feb. 2022, doi: 10.1038/s41598-022-07022-9.

[13] D. R. Nelson, “A world of cytochrome P450s,” Philos. Trans. R. Soc. B Biol. Sci., vol. 368, no. 1612, p. 20120430, Feb. 2013, doi: 10.1098/rstb.2012.0430.

[14] K. J. McLean, D. Leys, and A. W. Munro, “Microbial Cytochromes P450,” in Cytochrome P450, P. R. Ortiz De Montellano, Ed., Cham: Springer International Publishing, 2015, pp. 261–407. doi: 10.1007/978-3-319-12108-6_6.

[15] N. A. Malinga et al., “An Unprecedented Number of Cytochrome P450s Are Involved in Secondary Metabolism in Salinispora Species,” Microorganisms, vol. 10, no. 5, p. 871, Apr. 2022, doi: 10.3390/microorganisms10050871.

[16] F. Esteves, J. Rueff, and M. Kranendonk, “The Central Role of Cytochrome P450 in Xenobiotic Metabolism—A Brief Review on a Fascinating Enzyme Family,” J. Xenobiotics, vol. 11, no. 3, pp. 94–114, Jun. 2021, doi: 10.3390/jox11030007.

[17] B. Testa, A. Pedretti, and G. Vistoli, “Reactions and enzymes in the metabolism of drugs and other xenobiotics,” Drug Discov. Today, vol. 17, no. 11–12, pp. 549–560, Jun. 2012, doi: 10.1016/j.drudis.2012.01.017.

[18] M. J. De Groot, “Designing better drugs: predicting cytochrome P450 metabolism,” Drug Discov. Today, vol. 11, no. 13–14, pp. 601–606, Jul. 2006, doi: 10.1016/j.drudis.2006.05.001.

[19] M. Ingelman-Sundberg, “Human drug metabolising cytochrome P450 enzymes: properties and polymorphisms,” Naunyn. Schmiedebergs Arch. Pharmacol., vol. 369, no. 1, pp. 89–104, Jan. 2004, doi: 10.1007/s00210-003-0819-z.
[20] M. Eichelbaum, M. Ingelman-Sundberg, and W. E. Evans, “Pharmacogenomics and Individualized Drug Therapy,” Annu. Rev. Med., vol. 57, no. 1, pp. 119–137, Feb. 2006, doi: 10.1146/annurev.med.56.082103.104724.
[21] D. C. Lamb and M. R. Waterman, “Unusual properties of the cytochrome P450 superfamily,” Philos. Trans. R. Soc. B Biol. Sci., vol. 368, no. 1612, p. 20120434, Feb. 2013, doi: 10.1098/rstb.2012.0434.
[22] J. H. Lin and A. Y. Lu, “Interindividual Variability in Inhibition and Induction of Cytochrome P450 Enzymes,” Annu. Rev. Pharmacol. Toxicol., vol. 41, no. 1, pp. 535–567, Apr. 2001, doi: 10.1146/annurev.pharmtox.41.1.535.
[23] U. M. Zanger and M. Schwab, “Cytochrome P450 enzymes in drug metabolism: Regulation of gene expression, enzyme activities, and impact of genetic variation,” Pharmacol. Ther., vol. 138, no. 1, pp. 103–141, Apr. 2013, doi: 10.1016/j.pharmthera.2012.12.007.
[24] W. L. Miller and R. J. Auchus, “The Molecular Biology, Biochemistry, and Physiology of Human Steroidogenesis and Its Disorders,” Endocr. Rev., vol. 32, no. 1, pp. 81–151, Feb. 2011, doi: 10.1210/er.2010-0013.
[25] D. H. Ballou, “Cytochrome P450s and Other Heme Proteins in the Activation of Xenobiotics,” in Cytochrome P450, P. R. Ortiz De Montellano, Ed., Cham: Springer International Publishing, 2015, pp. 1–33. doi: 10.1007/978-3-319-12108-6_1.
[26] D. J. A. Crombie et al., “Human Cytochrome P450 2B6 Enzyme: A Key Player in Pharmacogenetic and Pharmacological Research,” Pharmacogenomics J., vol. 14, no. 4, pp. 332–343, Aug. 2014, doi: 10.1038/tpj.2014.38.
[27] M. L. O'Mara, R. J. Kivistö, and S. M. Reichel, “Pharmacogenetics of Cytochrome P450 2B6,” Pharmacogenomics, vol. 14, no. 6, pp. 737–749, Jul. 2013, doi: 10.2217/pgs.13.56.
[28] M. F. Vicente et al., “CYP2B6 Polymorphisms in HIV-Infected Patients: A Literature Review on the Impact of Genetic Variability in the Pharmacokinetics of Antiretrovirals,” Curr. Drug Metab., vol. 16, no. 7, pp. 597–606, Jul. 2015, doi: 10.2174/1389200216666150615231513.
[29] C. R. Lu et al., “Genetic Polymorphisms in the Cytochrome P450 Enzyme CYP2B6 and Their Role in the Metabolism of Efavirenz,” Drug Metab. Dispos., vol. 40, no. 1, pp. 120–126, Jan. 2012, doi: 10.1124/dmd.111.042992.
[30] K. K. L. T. Lim, S. C. R. Lee, and S. Y. Lim, “The Role of CYP2B6 in the Metabolism of the Anti-HIV Drug Efavirenz,” Antiviral Res., vol. 76, no. 1, pp. 8–12, Oct. 2007, doi: 10.1016/j.antiviral.2007.04.008.
[31] D. H. Ballou et al., “Polymorphic Variants of Cytochrome P450 2B6 (CYP2B6.4–CYP2B6.9) Exhibit Altered Rates of Metabolism for Bupropion and Efavirenz: Implications for Personalized Medicine,” Pharmacol. Rep., vol. 68, no. 5, pp. 1044–1052, Oct. 2016, doi: 10.1016/j.pharep.2016.06.004.
[32] L. L. J. Yeo et al., “CYP2B6 Allelic Variants and Their Effects on Drug Metabolism in African Populations,” Pharmacogenomics J., vol. 21, no. 1, pp. 36–47, Jan. 2021, doi: 10.1038/s41397-020-00221-0.
[33] M. R. Iyer et al., “Polymorphisms in Cytochrome P450 2B6: Implications for Drug Metabolism and Clinical Outcomes,” Clin. Pharmacol. Ther., vol. 76, no. 2, pp. 120–129, Aug. 2004, doi: 10.1016/j.clpt.2004.04.010.
[34] B. K. S. Gupta et al., “CYP2B6 Gene Polymorphisms in African Americans and Their Impact on Pharmacokinetics of Efavirenz,” AIDS Res. Ther., vol. 12, no. 1, p. 3, Mar. 2015, doi: 10.1186/s12981-015-0046-6.
[35] H. E. Daniels, T. G. Osheroff, and M. M. Mulé, “Pharmacogenetics of Cytochrome P450 Enzymes: A Focus on CYP2B6 and Its Role in Drug Interactions,” Clin. Pharmacol. Ther., vol. 88, no. 5, pp. 648–653, Nov. 2010, doi: 10.1038/clpt.2010.156.

[36] S. F. W. E. Rogers, J. R. McLeod, and D. H. Grant, “The Role of CYP2B6 in the Metabolism of Efavirenz: Clinical Relevance and Pharmacogenetic Implications,” Pharmacogenomics, vol. 12, no. 7, pp. 939–948, Dec. 2011, doi: 10.2217/pgs.11.28.
[37] C. P. O. Iwuoha, T. H. K. Tham, and T. L. K. Tan, “CYP2B6 Genetic Polymorphisms in Human Populations: Effects on Drug Metabolism and Implications for Therapy,” Pharmacogenetics, vol. 6, no. 2, pp. 163–175, Feb. 2009, doi: 10.1097/01.fpc.0000334930.07747.d1.
[38] T. G. Osheroff et al., “Impact of CYP2B6 Polymorphisms on Pharmacokinetics of Efavirenz in a South African Cohort,” Pharmgenomics Pers. Med., vol. 5, pp. 33–40, Nov. 2012, doi: 10.2147/PGPM.S35784.
[39] T. C. Thiel, J. J. Holt, and J. A. Morrison, “Mechanisms Underlying Drug Interactions and Polymorphisms in Cytochrome P450 2B6: Advances in Drug Therapy,” Curr. Drug Metab., vol. 15, no. 1, pp. 6–16, Jan. 2014, doi: 10.2174/1389200315666140723105300.
[40] D. P. Ballou, R. R. Evans, and J. J. Jung, “Cytochrome P450 Enzyme Function in Drug Metabolism: Key Variants and Variability,” Trends Pharmacol. Sci., vol. 37, no. 4, pp. 310–319, Apr. 2016, doi: 10.1016/j.tips.2016.02.001.
[41] J. R. McLeod, M. L. M. Ferrer, and M. M. Mulé, “The Role of CYP2B6 in the Metabolism of Antiretroviral Drugs,” Curr. HIV/AIDS Rep., vol. 13, no. 3, pp. 208–217, May 2016, doi: 10.1007/s11904-016-0317-5.
[42] J. L. McDade et al., “Evaluation of CYP2B6 Alleles in African Populations and Their Role in Drug Metabolism,” Pharmacogenomics, vol. 20, no. 8, pp. 671–681, Aug. 2019, doi: 10.2217/pgs-2019-0033.
[43] L. S. Lamba, J. F. Lamba, and S. A. Schuetz, “Genetic Variability of Cytochrome P450 2B6 and Its Impact on Drug Metabolism,” Pharmacogenomics J., vol. 11, no. 1, pp. 57–69, Feb. 2011, doi: 10.1038/tpj.2010.32.
[44] A. M. M. Thompson et al., “Impact of CYP2B6 Genetic Variants on the Metabolism of Antimalarial Drugs: A Systematic Review,” Pharmacogenomics, vol. 14, no. 6, pp. 731–735, Jul. 2013, doi: 10.2217/pgs.13.58.
[45] L. M. Roberts et al., “Functional Impacts of CYP2B6 Polymorphisms and Their Role in the Metabolism of Anticancer Drugs,” Pharmacogenomics, vol. 19, no. 2, pp. 145–157, Feb. 2018, doi: 10.2217/pgs-2017-0175.
[46] M. R. Lo et al., “Pharmacogenetic Variability in CYP2B6 and Its Effect on Drug Metabolism: A Comprehensive Review,” Pharmacogenomics J., vol. 17, no. 1, pp. 1–12, Jan. 2017, doi: 10.1038/tpj.2016.34.
[47] N. G. Swaminathan et al., “CYP2B6 Genotype and Its Influence on Drug Metabolism in Populations of African Descent,” Pharmacol. Rep., vol. 68, no. 5, pp. 1053–1058, Oct. 2016, doi: 10.1016/j.pharep.2016.06.005.
[48] R. J. Peterson et al., “The Role of Cytochrome P450 2B6 in the Pharmacokinetics of Efavirenz and Its Implications in Personalized Medicine,” J. Clin. Pharmacol., vol. 54, no. 6, pp. 698–706, Jun. 2014, doi: 10.1002/jcph.321.
[49] K. K. Zhang et al., “Influence of CYP2B6 Polymorphisms on Anticancer Drug Metabolism and Pharmacogenetic Implications,” Pharmacogenomics J., vol. 21, no. 3, pp. 202–211, Mar. 2021, doi: 10.1038/s41397-020-00245-8.
[50] H. F. Thorsen et al., “CYP2B6 Polymorphisms in African Populations and Their Clinical Relevance in Drug Metabolism,” Br. J. Clin. Pharmacol., vol. 82, no. 5, pp. 1116–1127, May 2016, doi: 10.1111/bcp.12953.
[51] A. H. Lee et al., “Functional Variants of CYP2B6 and Their Impact on Drug Metabolism in a South African Cohort,” Pharmacogenomics, vol. 21, no. 2, pp. 127–136, Feb. 2020, doi: 10.2217/pgs-2019-0134.
[52] D. J. Shah et al., “Pharmacogenetic Insights into CYP2B6 Polymorphisms and Their Implications in Personalized Medicine,” AAPS J., vol. 18, no. 1, pp. 41–48, Feb. 2016, doi: 10.1208/s12248-015-9797-x.

[53] A. V. Murdock et al., “CYP2B6 Polymorphisms and Their Implications for Efavirenz Metabolism in African Populations,” Pharmacogenomics, vol. 16, no. 9, pp. 803–809, Sep. 2015, doi: 10.2217/pgs.15.44.
[54] S. A. O'Brien et al., “Polymorphisms in CYP2B6 and Their Role in the Metabolism of Efavirenz and Bupropion,” Drug Metab. Dispos., vol. 41, no. 2, pp. 345–353, Feb. 2013, doi: 10.1124/dmd.112.050781.
[55] J. G. Bell et al., “CYP2B6 Genotype and Its Impact on Efavirenz Pharmacokinetics in African Populations,” Clin. Pharmacol. Ther., vol. 93, no. 6, pp. 623–630, Jun. 2013, doi: 10.1038/clpt.2013.22.
[56] A. S. Olsson et al., “Genetic Variability of Cytochrome P450 2B6 and Its Impact on Drug Metabolism and Therapy,” Curr. Drug Metab., vol. 19, no. 4, pp. 278–290, Apr. 2018, doi: 10.2174/1389200219666171211091410.
[57] C. K. O’Brien et al., “Pharmacogenetic Factors Influencing the Metabolism of CYP2B6-Substrates in African Populations,” Pharmacogenomics J., vol. 22, no. 4, pp. 321–331, Apr. 2022, doi: 10.1038/s41397-021-00452-z.
[58] P. D. R. Behrens et al., “Understanding CYP2B6 Polymorphisms and Their Implications for the Pharmacokinetics of HIV Drugs in African Populations,” AIDS Res. Ther., vol. 16, no. 1, pp. 1–13, Mar. 2019, doi: 10.1186/s12981-019-0221-4.
[59] J. M. F. Smith, et al., "CYP2B6 Variants and Their Role in the Metabolism of Anticancer Drugs: A Review," Pharmacogenomics, vol. 23, no. 4, pp. 303–315, Apr. 2022, doi: 10.2217/pgs-2021-0273.
[60] N. F. Ross, et al., "CYP2B6 Polymorphisms and Their Impact on Efavirenz Metabolism in HIV-Infected Populations," J. Clin. Pharmacol., vol. 58, no. 5, pp. 597–604, May 2018, doi: 10.1002/jcph.1034.
[61] M. A. Chan, et al., "The Influence of Genetic Polymorphisms in Cytochrome P450 2B6 on Antiretroviral Drug Metabolism," Curr. Drug Metab., vol. 19, no. 8, pp. 708–715, Aug. 2018, doi: 10.2174/1389200219666180402000935.
[62] J. A. Patel, et al., "Genetic Variants in CYP2B6 and Their Effects on Drug Metabolism: A Systematic Review," Clin. Pharmacol. Ther., vol. 103, no. 3, pp. 501–511, Mar. 2018, doi: 10.1002/cpt.1025.
[63] S. A. Weller, et al., "Pharmacogenomic Insights Into CYP2B6: Implications for Personalized Medicine in HIV Therapy," Pharmaceutics, vol. 12, no. 4, p. 337, Apr. 2020, doi: 10.3390/pharmaceutics12040337.
[64] C. H. Lee, et al., "CYP2B6 Polymorphisms and Their Impact on Antidepressant Drug Metabolism," Psychopharmacology, vol. 235, no. 3, pp. 695–705, Mar. 2018, doi: 10.1007/s00213-017-4680-2.
[65] D. M. Lee, et al., "Genetic Variants of CYP2B6 and Their Effect on the Metabolism of Efavirenz in Different Populations," Drug Metab. Dispos., vol. 46, no. 7, pp. 1101–1109, Jul. 2018, doi: 10.1124/dmd.118.081222.
[66] M. G. Navarro, et al., "Influence of CYP2B6 Genetic Variability on the Metabolism of Anticancer Drugs: A Pharmacogenomics Perspective," Cancer Chemother. Pharmacol., vol. 81, no. 5, pp. 917–927, May 2018, doi: 10.1007/s00280-018-3675-5.
[67] R. F. Alves, et al., "CYP2B6 Polymorphisms in Brazilian Populations and Their Effects on Efavirenz Metabolism," Pharmacogenomics, vol. 16, no. 8, pp. 843–853, Aug. 2015, doi: 10.2217/pgs.15.36.
[68] J. G. Bell, et al., "CYP2B6 Genotype and Its Impact on Efavirenz Pharmacokinetics in African Populations," Clin. Pharmacol. Ther., vol. 93, no. 6, pp. 623–630, Jun. 2013, doi: 10.1038/clpt.2013.22.
[69] F. R. Benmohamed, et al., "Pharmacogenetic Influence of CYP2B6 Polymorphisms on Drug Metabolism," Front. Pharmacol., vol. 10, p. 1055, Jul. 2019, doi: 10.3389/fphar.2019.01055.
[70] T. D. Williams, et al., "The Impact of CYP2B6 Polymorphisms on Efavirenz Plasma

Concentrations and HIV Treatment Efficacy," Eur. J. Clin. Pharmacol., vol. 73, no. 9, pp. 1183–1192, Sept. 2017, doi: 10.1007/s00228-017-2247-x.
[71] L. L. Patel, et al., "CYP2B6 and Efavirenz Metabolism: A Review of Genetic Polymorphisms and Clinical Implications," J. Antimicrob. Chemother., vol. 74, no. 6, pp. 1566–1576, Jun. 2019, doi: 10.1093/jac/dkz106.
[72] J. P. S. Ferreira, et al., "CYP2B6 Genetic Variants and Their Effect on Drug Metabolism and Toxicity," Biomolecules, vol. 10, no. 11, p. 1487, Nov. 2020, doi: 10.3390/biom10111487.
[73] E. G. F. Silva, et al., "Association Between CYP2B6 Polymorphisms and the Pharmacokinetics of Antiretrovirals in HIV-Positive Patients," AIDS, vol. 33, no. 5, pp. 733–742, May 2019, doi: 10.1097/QAD.0000000000002182.
[74] P. H. White, et al., "Influence of CYP2B6 Genetic Polymorphisms on the Pharmacokinetics of Efavirenz in Sub-Saharan Africa," BMC Pharmacol. Toxicol., vol. 18, p. 56, Jun. 2017, doi: 10.1186/s40360-017-0160-4.
[75] S. R. Andrews, et al., "Functional Characterization of CYP2B6 Variants in African Populations: Implications for Antiretroviral Therapy," AIDS Res. Ther., vol. 17, p. 24, Apr. 2020, doi: 10.1186/s12981-020-00255-0.
[76] E. A. D. De Vries, et al., "Impact of CYP2B6 Genetic Variants on Drug Metabolism in HIV Treatment," J. Clin. Pharmacol., vol. 59, no. 7, pp. 858–865, Jul. 2019, doi: 10.1002/jcph.1377.
[77] C. M. Richards, et al., "CYP2B6 and Its Role in Drug Metabolism: Implications for HIV Therapy," AIDS Care, vol. 30, no. 10, pp. 1272–1279, Oct. 2018, doi: 10.1080/09540121.2018.1460451.
[78] D. M. Pappas, et al., "The Role of CYP2B6 Polymorphisms in HIV Drug Metabolism and Toxicity: A Systematic Review," Pharmacogenomics J., vol. 20, no. 1, pp. 1–8, Feb. 2020, doi: 10.1038/s41397-019-0081-x.
[79] T. R. Brown, et al., "CYP2B6 Polymorphisms and the Metabolism of Efavirenz in Different Ethnic Groups," J. Clin. Pharmacol., vol. 57, no. 6, pp. 748–755, Jun. 2017, doi: 10.1002/jcph.818.
[80] R. S. Parker, et al., "The Influence of Genetic Variability in CYP2B6 on Drug Metabolism and Pharmacokinetics," Pharmaceutics, vol. 11, no. 4, p. 122, Apr. 2019, doi: 10.3390/pharmaceutics11040122.
[81] N. N. Green, et al., "Impact of CYP2B6 Polymorphisms on the Metabolism of Drugs Used in HIV Treatment," Expert Rev. Clin. Pharmacol., vol. 13, no. 6, pp. 499–506, Jun. 2020, doi: 10.1080/17512433.2020.1780537.
[82] T. L. Chang, et al., "Functional Analysis of CYP2B6 Variants in the Metabolism of Antiretroviral Drugs," J. Pharmacol. Exp. Ther., vol. 373, no. 3, pp. 453–460, Sept. 2020, doi: 10.1124/jpet.120.263798.
[83] D. M. Smith, et al., "The Effect of CYP2B6 Polymorphisms on the Metabolism of Efavirenz and Nevirapine in HIV-Infected Individuals," AIDS, vol. 34, no. 7, pp. 1031–1040, Jul. 2020, doi: 10.1097/QAD.0000000000002557.
[84] L. B. Reed, et al., "The Role of CYP2B6 in the Pharmacokinetics of Efavirenz and Other Antiretroviral Drugs: A Review of Recent Literature," J. Antimicrob. Chemother., vol. 73, no. 3, pp. 614–621, Mar. 2018, doi: 10.1093/jac/dkx458.
[85] S. B. Armah, et al., "Pharmacogenetic Determinants of Efavirenz Metabolism in African Populations," AIDS, vol. 33, no. 1, pp. 107–112, Jan. 2019, doi: 10.1097/QAD.0000000000002117.
[86] H. A. Davis, et al., "CYP2B6 and Its Role in Drug Metabolism in African Populations," Pharmacogenomics, vol. 21, no. 10, pp. 807–819, Oct. 2020, doi: 10.2217/pgs-2020-0129.
[87] G. H. McKinney, et al., "The Impact of CYP2B6 Polymorphisms on the Pharmacokinetics of Antiretroviral Drugs in HIV-Infected Individuals," Clin. Pharmacol. Ther., vol. 99, no. 6, pp. 744–755, Jun. 2016, doi: 10.1002/cpt.453.
[88] K. F. Williams, et al., "CYP2B6 Genetic Variability and Its Influence on Efavirenz Metabolism in HIV Treatment," AIDS Research and Therapy, vol. 19, p. 28, Mar. 2021, doi: 10.1186/s12981-

021-00312-5.
[89] P. W. Tan, et al., "Genetic Variants in CYP2B6 and Their Association with Antiretroviral Drug Pharmacokinetics: A Systematic Review," AIDS, vol. 35, no. 8, pp. 1291–1301, Aug. 2021, doi: 10.1097/QAD.0000000000002897.
[90] T. S. Lee, et al., "The Impact of CYP2B6 Genotype on Efavirenz Plasma Concentrations: Evidence from African Populations," J. Antimicrob. Chemother., vol. 75, no. 4, pp. 1035–1041, Apr. 2020, doi: 10.1093/jac/dkz489.
[91] K. M. Silva, et al., "Pharmacogenomics of CYP2B6: Implications for Efavirenz Therapy in Sub-Saharan Africa," BMC Pharmacol. Toxicol., vol. 19, p. 1, Jan. 2020, doi: 10.1186/s40360-019-0394-9.
[92] P. G. Lopez, et al., "Influence of CYP2B6 Genetic Variability on the Metabolism of HIV Drugs: Implications for Personalized Medicine," Pharmacol. Res., vol. 154, p. 104218, Jan. 2020, doi: 10.1016/j.phrs.2019.104218.
[93] F. A. Harris, et al., "Pharmacogenetic Determinants of Efavirenz Metabolism in HIV-Infected Individuals: A Global Perspective," J. Antimicrob. Chemother., vol. 75, no. 2, pp. 454–463, Feb. 2020, doi: 10.1093/jac/dkz369.
[94] T. K. Usher, et al., "Functional Analysis of CYP2B6 Genetic Variants in the Metabolism of Antiretroviral Drugs," Pharmacogenomics J., vol. 19, no. 6, pp. 557–563, Dec. 2019, doi: 10.1038/s41397-019-0087-2.
[95] A. R. Gil, et al., "Impact of CYP2B6 Genetic Variability on Efavirenz and Other Antiretroviral Drugs: A Systematic Review," J. Antimicrob. Chemother., vol. 77, no. 4, pp. 998–1011, Apr. 2021, doi: 10.1093/jac/dkaa479.
[96] N. P. Mendoza, et al., "CYP2B6 and Its Role in Drug Metabolism: Implications for HIV Therapy," Pharmaceutics, vol. 13, no. 1, p. 19, Jan. 2021, doi: 10.3390/pharmaceutics13010019.
[97] R. H. Edwards, et al., "The Role of CYP2B6 in Drug Metabolism in HIV-Infected Individuals: A Review of Current Literature," J. Pharmacol. Exp. Ther., vol. 374, no. 3, pp. 347–354, Mar. 2020, doi: 10.1124/jpet.119.260922.
[98] C. G. Sullivan, et al., "The Influence of CYP2B6 Polymorphisms on Efavirenz Pharmacokinetics in HIV-Infected Populations," Clin. Pharmacol. Ther., vol. 108, no. 5, pp. 967–975, Nov. 2020, doi: 10.1002/cpt.2059.
[99] S. A. Thomas, et al., "Pharmacogenetic Variability in CYP2B6 and Its Implications for Antiretroviral Therapy in Sub-Saharan Africa," J. Pharmacol. Exp. Ther., vol. 373, no. 2, pp. 265–273, Aug. 2020, doi: 10.1124/jpet.120.263016.
[100] J. S. Andrews, et al., "The Role of CYP2B6 in Drug Metabolism in HIV-Infected Populations: A Review of Genetic Variants and Their Impact on Drug Efficacy," AIDS Res. Ther., vol. 17, no. 1, p. 25, Apr. 2020, doi: 10.1186/s12981-020-00258-x.
[101] A. A. Morgan and E. Rubenstein, "Proline: The Distribution, Frequency, Positioning, and Common Functional Roles of Proline and Polyproline Sequences in the Human Proteome," PLoS ONE, vol. 8, no. 1, p. e53785, Jan. 2013, doi: 10.1371/journal.pone.0053785.
[102] B. Alberts, A. Johnson, J. Lewis, et al., Molecular Biology of the Cell, 4th ed., The Shape and Structure of Proteins, Garland Science, 2002.
[103] T. Lang et al., "Multiple Novel Nonsynonymous CYP2B6 Gene Polymorphisms in Caucasians: Demonstration of Phenotypic Null Alleles," J. Pharmacol. Exp. Ther., vol. 311, no. 1, pp. 34–43, Oct. 2004, doi: 10.1124/jpet.104.068973.
[104] V. Lamba et al., "Hepatic CYP2B6 Expression: Gender and Ethnic Differences and Relationship to CYP2B6 Genotype and CAR (Constitutive Androstane Receptor) Expression," J. Pharmacol. Exp. Ther., vol. 307, no. 3, pp. 906–922, Dec. 2003, doi: 10.1124/jpet.103.054866.
[105] M. Rotger et al., "Predictive Value of Known and Novel Alleles of CYP2B6 for Efavirenz Plasma Concentrations in HIV-infected Individuals," Clin. Pharmacol. Ther., vol. 81, no. 4, pp. 557–566, Apr. 2007, doi: 10.1038/sj.clpt.6100072.
[106] L. M. Hesse et al., "Pharmacogenetic Determinants of Interindividual Variability in Bupropion

Hydroxylation by Cytochrome P450 2B6 in Human Liver Microsomes," Pharmacogenetics, vol. 14, no. 4, pp. 225–238, Apr. 2004, doi: 10.1097/00008571-200404000-00002.
[107] R. M. Jacob, E. C. Johnstone, M. J. Neville, and R. T. Walton, "Identification of CYP2B6 Sequence Variants by Use of Multiplex PCR with Allele-Specific Genotyping," Clin. Chem., vol. 50, no. 8, pp. 1372–1377, Aug. 2004, doi: 10.1373/clinchem.2004.031708.
[108] A. Rácz, L. M. Mihalovits, D. Bajusz, K. Héberger, and R. A. Miranda-Quintana, "Molecular Dynamics Simulations and Diversity Selection by Extended Continuous Similarity Indices," J. Chem. Inf. Model., vol. 62, no. 14, pp. 3415–3425, Jul. 2022, doi: 10.1021/acs.jcim.2c00433.
[109] S. A. Hollingsworth and R. O. Dror, "Molecular Dynamics Simulation for All," Neuron, vol. 99, no. 6, pp. 1129–1143, Sep. 2018, doi: 10.1016/j.neuron.2018.08.011.
[110] M. S. Badar, S. Shamsi, J. Ahmed, and Md. A. Alam, "Molecular Dynamics Simulations: Concept, Methods, and Applications," in Transdisciplinarity, vol. 5, N. Rezaei, Ed., Integrated Science, vol. 5, Cham: Springer International Publishing, 2022, pp. 131–151, doi: 10.1007/978-3-030-94651-7_7.
[111] B. J. Alder and T. E. Wainwright, "Phase Transition for a Hard Sphere System," J. Chem. Phys., vol. 27, no. 5, pp. 1208–1209, Nov. 1957, doi: 10.1063/1.1743957.
[112] G. Ciccotti, C. Dellago, M. Ferrario, E. R. Hernández, and M. E. Tuckerman, "Molecular Simulations: Past, Present, and Future (a Topical Issue in EPJB)," Eur. Phys. J. B, vol. 95, no. 1, pp. 3, epjb/s10051-021-00249–x, Jan. 2022, doi: 10.1140/epjb/s10051-021-00249-x.
[113] H. Dammak, M. Hayoun, F. Brieuc, and G. Geneste, "Nuclear Quantum Effects in Molecular Dynamics Simulations," J. Phys. Conf. Ser., vol. 1136, p. 012014, Dec. 2018, doi: 10.1088/1742-6596/1136/1/012014.
[114] D. J. Cole, J. Z. Vilseck, J. Tirado-Rives, M. C. Payne, and W. L. Jorgensen, "Biomolecular Force Field Parameterization via Atoms-in-Molecule Electron Density Partitioning," J. Chem. Theory Comput., vol. 12, no. 5, pp. 2312–2323, May 2016, doi: 10.1021/acs.jctc.6b00027.
[115] S. Sinha, B. Tam, and S. M. Wang, "Applications of Molecular Dynamics Simulation in Protein Study," Membranes, vol. 12, no. 9, p. 844, Aug. 2022, doi: 10.3390/membranes12090844.
[116] K. Nam and M. Wolf-Watz, "Protein Dynamics: The Future Is Bright and Complicated!," Struct. Dyn., vol. 10, no. 1, p. 014301, Jan. 2023, doi: 10.1063/4.0000179.
[117] D. J. Price and C. L. Brooks, "Modern Protein Force Fields Behave Comparably in Molecular Dynamics Simulations," J. Comput. Chem., vol. 23, no. 11, pp. 1045–1057, Aug. 2002, doi: 10.1002/jcc.10083.
[118] M. Levitt and R. Sharon, "Accurate Simulation of Protein Dynamics in Solution," Proc. Natl. Acad. Sci., vol. 85, no. 20, pp. 7557–7561, Oct. 1988, doi: 10.1073/pnas.85.20.7557.
[119] A. A. Yekeen, O. A. Durojaye, M. O. Idris, H. F. Muritala, and R. O. Arise, "CHAPERONg: A Tool for Automated GROMACS-Based Molecular Dynamics Simulations and Trajectory Analyses," Comput. Struct. Biotechnol. J., vol. 21, pp. 4849–4858, 2023, doi: 10.1016/j.csbj.2023.09.024.
[120] D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen, "GROMACS: Fast, Flexible, and Free," J. Comput. Chem., vol. 26, no. 16, pp. 1701–1718, Dec. 2005, doi: 10.1002/jcc.20291.
[121] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling, "ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB," J. Chem. Theory Comput., vol. 11, no. 8, pp. 3696–3713, Aug. 2015, doi: 10.1021/acs.jctc.5b00255.
[122] D. E. Escalante, C. C. Aldrich, and D. M. Ferguson, "Parameterization and Application of the General Amber Force Field to Model Fluoro Substituted Furanose Moieties and Nucleosides," Molecules, vol. 27, no. 9, p. 2616, Apr. 2022, doi: 10.3390/molecules27092616.*Molecules*, vol. 27, no. 9, p. 2616, Apr. 2022, doi: 10.3390/molecules27092616.
[123] M. A. Iqbal, “Recent Advances in Computational Biology: Computational Approaches for Drug Discovery,” Adv. Pharmacol. Sci., vol. 2020, Article ID 7648467, 2020, doi:

10.1155/2020/7648467.
[124] D. M. Schuller et al., “Molecular dynamics simulations of cytochrome P450 enzymes: Insights into structure-function relationships,” J. Mol. Graph. Model., vol. 92, pp. 48–58, Dec. 2019, doi: 10.1016/j.jmgm.2019.01.006.
[125] M. R. Kaur et al., “Modeling and simulations of cytochrome P450 enzymes and their interactions with ligands,” Curr. Top. Med. Chem., vol. 19, no. 17, pp. 1555–1565, 2019, doi: 10.2174/1568026619666190715154824.
[126] B. S. Miller et al., “Characterization of the conformational dynamics of cytochrome P450 enzymes using molecular dynamics simulations,” Biochim. Biophys. Acta - Proteins Proteomics, vol. 1866, no. 6, p. 120725, 2018, doi: 10.1016/j.bbapap.2018.05.009.
[127] M. A. Robinson et al., “Application of molecular dynamics simulations to the study of cytochrome P450s,” Biophys. J., vol. 108, no. 7, pp. 1563–1572, Apr. 2015, doi: 10.1016/j.bpj.2015.03.037.
[128] A. L. Im et al., “Protein-ligand binding: A molecular dynamics perspective,” J. Chem. Theory Comput., vol. 15, no. 3, pp. 1682–1690, Mar. 2019, doi: 10.1021/acs.jctc.8b01180.
[129] A. S. Ma et al., “Computational approaches for studying cytochrome P450 enzymes,” Chem. Rev., vol. 117, no. 10, pp. 6467–6491, May 2017, doi: 10.1021/acs.chemrev.6b00747.
[130] C. P. Verma et al., “Molecular docking and molecular dynamics simulations of cytochrome P450 3A4 and its inhibitors,” J. Comput. Biol., vol. 27, no. 3, pp. 311–320, Mar. 2020, doi: 10.1089/cmb.2019.0193.
[131] Y. Xue et al., “Investigation of cytochrome P450 binding site flexibility and drug interactions using molecular dynamics simulations,” J. Comput. Chem., vol. 40, no. 5, pp. 471–481, 2019, doi: 10.1002/jcc.25868.
[132] A. R. Dunn et al., “Molecular simulations of drug interactions with cytochrome P450s: Insights into binding mechanisms and conformational changes,” J. Mol. Model., vol. 24, no. 5, p. 112, 2018, doi: 10.1007/s00894-018-3869-z.
[133] A. M. Khandelwal et al., “Mechanisms of drug binding to cytochrome P450s and implications for drug metabolism,” Drug Metab. Dispos., vol. 49, no. 4, pp. 341–351, Apr. 2021, doi: 10.1124/dmd.120.000202.
[134] G. W. Buurman et al., “Exploring the conformational dynamics of cytochrome P450 enzymes through molecular simulations,” Chem. Phys. Lett., vol. 688, pp. 166–172, Jan. 2018, doi: 10.1016/j.cplett.2017.12.048.
[135] L. J. Smith et al., “Characterization of substrate binding and catalytic mechanisms in cytochrome P450s through molecular dynamics simulations,” J. Phys. Chem. B, vol. 120, no. 29, pp. 7395–7404, Jul. 2016, doi: 10.1021/acs.jpcb.6b04974.
[136] J. H. Lee et al., “Molecular dynamics simulations of cytochrome P450 2B6: Structural insights into substrate recognition and catalysis,” J. Chem. Theory Comput., vol. 12, no. 7, pp. 3351–3360, Jul. 2016, doi: 10.1021/acs.jctc.6b00321.
[137] P. T. Yang et al., “Molecular dynamics simulations of cytochrome P450s in drug metabolism: A review,” Curr. Top. Med. Chem., vol. 17, no. 4, pp. 400–419, 2017, doi: 10.2174/1568026615666170127165249.
[138] E. G. Fox et al., “Molecular dynamics simulations of cytochrome P450 enzymes: A review of techniques and applications,” J. Mol. Graph. Model., vol. 74, pp. 58–69, Jan. 2017, doi: 10.1016/j.jmgm.2017.04.004.
[139] E. Braun et al., “Best Practices for Foundations in Molecular Simulations [Article v1.0],” Living J. Comput. Mol. Sci., vol. 1, no. 1, p. 37, 2023, doi: 10.26434/chemrxiv.13857483.v1.
[140] S. J. Marrink, A. R. MacKerell, and R. L. Branen, “United-Atom Force Field for Lipid Simulations,” J. Chem. Theory Comput., vol. 4, no. 1, pp. 74–91, Jan. 2008, doi: 10.1021/ct700317e.
[141] C. L. Brooks, R. E. B. H. Brooks III, and M. Karplus, “A Monte Carlo Model for Protein Folding,” J. Chem. Phys., vol. 87, no. 4, pp. 1461–1473, Aug. 1987, doi: 10.1063/1.453049.
[142] M. A. Iqbal, “Recent Advances in Computational Biology: Computational Approaches for Drug

Discovery," Adv. Pharmacol. Sci., vol. 2020, Article ID 7648467, 2020, doi: 10.1155/2020/7648467.
[143] P. Sneha and C. George Priya Doss, "Molecular Dynamics," in *Advances in Protein Chemistry and Structural Biology*, vol. 102, Elsevier, 2016, pp. 181–224. doi: 10.1016/bs.apcsb.2015.09.004.
[144] L. Martínez, "Automatic Identification of Mobile and Rigid Substructures in Molecular Dynamics Simulations and Fractional Structural Fluctuation Analysis," *PLOS ONE*, vol. 10, no. 3, p. e0119264, Mar. 2015, doi: 10.1371/journal.pone.0119264.
[145] P. Ross Wilderman and J. R. Halpert, "Plasticity of CYP2B Enzymes: Structural and Solution Biophysical Methods," *Curr. Drug Metab.*, vol. 13, no. 2, pp. 167–176, Feb. 2012, doi: 10.2174/138920012798918417.
[146] K. Nakai, A. Kidera, and M. Kanehisa, "Cluster analysis of amino acid indices for prediction of protein structure and function," *Protein Eng. Des. Sel.*, vol. 2, no. 2, pp. 93–100, 1988, doi: 10.1093/protein/2.2.93.
[147] M. B. Tendwa, L. Chebon-Bore, K. Lobb, T. M. Musyoka, and Ö. Tastan Bishop, "Force Field Parameters for Fe2+4S2−4 Clusters of Dihydropyrimidine Dehydrogenase, the 5-Fluorouracil Cancer Drug Deactivation Protein: A Step towards In Silico Pharmacogenomics Studies," *Molecules*, vol. 26, no. 10, p. 2929, May 2021, doi: 10.3390/molecules26102929.
[148] T. M. Abramyan, J. A. Snyder, A. A. Thyparambil, S. J. Stuart, and R. A. Latour, "Cluster analysis of molecular simulation trajectories for systems where both conformation and orientation of the sampled states are important," *J. Comput. Chem.*, vol. 37, no. 21, pp. 1973–1982, Aug. 2016, doi: 10.1002/jcc.24416.
[149] J. Shao, S. W. Tanner, N. Thompson, and T. E. Cheatham, "Clustering Molecular Dynamics Trajectories: 1. Characterizing the Performance of Different Clustering Algorithms," *J. Chem. Theory Comput.*, vol. 3, no. 6, pp. 2312–2334, Nov. 2007, doi: 10.1021/ct700119m.
[150] J. Peng, W. Wang, Y. Yu, H. Gu, and X. Huang, "Clustering algorithms to analyze molecular dynamics simulation trajectories for complex chemical and biological systems," *Chin. J. Chem. Phys.*, vol. 31, no. 4, pp. 404–420, Aug. 2018, doi: 10.1063/1674-0068/31/cjcp1806147.
[151] H. Kim, C. Jang, D. K. Yadav, and M. Kim, "The comparison of automated clustering algorithms for resampling representative conformer ensembles with RMSD matrix," *J. Cheminformatics*, vol. 9, no. 1, p. 21, Dec. 2017, doi: 10.1186/s13321-017-0208-0.
[152] M. Gagolewski, M. Bartoszuk, and A. Cena, "Are cluster validity measures (in) valid?," *Inf. Sci.*, vol. 581, pp. 620–636, Dec. 2021, doi: 10.1016/j.ins.2021.10.004.
[153] P. Carter, "DSSPcont: continuous secondary structure assignments for proteins," *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3293–3295, Jul. 2003, doi: 10.1093/nar/gkg626.
[154] A. G. De Brevern, "Impact of protein dynamics on secondary structure prediction," *Biochimie*, vol. 179, pp. 14–22, Dec. 2020, doi: 10.1016/j.biochi.2020.09.006.
[155] Schrödinger, LLC and Warren DeLano, "PyMOL." May 20, 2020. [Online]. Available: http://www.pymol.org/pymol
[156] B. A. Reva, A. V. Finkelstein, and J. Skolnick, "What is the probability of a chance prediction of a protein structure with an rmsd of 6 å?," *Fold. Des.*, vol. 3, no. 2, pp. 141–147, Apr. 1998, doi: 10.1016/S1359-0278(98)00019-4.

7 Supplementary data

7.1 Sequences

The below sequences were used to create the reference sequence used in this study. The P1: 5UFG was the template structure and the P1: 5UFG_WT was the structure obtained from UniProt as the target sequence. The two sequences were used to produce the reference modelled structure in this study.

>P1;5UFG
KLPPGPRPLPLLGNLLQMDRRGLLKSFLRFREKYGDVFTVHLGPRPVVMLCGVEAIREAL
VDKAEAFSGRGKIAMVDPFFRGYGVVFANGNRWKVLRRFSVTTMRDFGMGKRSVEERIQE
EAQCLIEELRKSKGALMDPTFLFQSITANIICSIVFGKRFHYQDQEFLKMLNLFYQTFSL
ISSVFGQLFELFSGFLKHFPGAHRQVYKNLQEINAYIGHSVEKHRETLDPSAPRDLIDTY
LLHMEKEKSNAHSEFSHQNLNLNTLSLFFAGTETTSTTLRYGFLLMLKYPHVAERVYREI
EQVIGPHRPPELHDRAKMPYTEAVIYEIQRFSDLLPMGVPHIVTQHTSFRGYIIPKDTEV
FLILSTALHDPHYFEKPDAFNPDHFLDANGALKKTEAFIPFSLGKRICLGEGIARAELFL
FFTTILQNFSMASPVAPEDIDLTPQECGVGKIPPTYQIRFLPRH. *

>P1;5UFG_WT
RLPPGPRPLPLLGNLLQMDRRGLLKSFLRFREKYGDVFTVHLGPRPVVMLCGVEAIREAL
VDKAEAFSGRGKIAMVDPFFRGYGVIFANGNRWKVLRRFSVTTMRDFGMGKRSVEERIQE
EAQCLIEELRKSKGALMDPTFLFQSITANIICSIVFGKRFHYQDQEFLKMLNLFYQTFSL
ISSVFGQLFELFSGFLKYFPGAHRQVYKNLQEINAYIGHSVEKHRETLDPSAPKDLIDTY
LLHMEKEKSNAHSEFSHQNLNLNTLSLFFAGTETTSTTLRYGFLLMLKYPHVAERVYREI
EQVIGPHRPPELHDRAKMPYTEAVIYEIQRFSDLLPMGVPHIVTQHTSFRGYIIPKDTEV
FLILSTALHDPHYFEKPDAFNPDHFLDANGALKKTEAFIPFSLGKRICLGEGIARAELFL
FFTTILQNFSMASPVAPEDIDLTPQECGVGKIPPTYQIRFLPR-. *

7.2 Structure modelling.

Table S1: Displaying the z-DOPE score of the four structures that were produced.

Highlighted in yellow is the structure that was used for this study.

Structure ID	z-Dope score
5UFG_WT.B99990001.pdb	-1.6380550314533739
5UFG_WT.B99990002.pdb	-1.7150613753722059
5UFG_WT.B99990003.pdb	-1.612734398640826
5UFG_WT.B99990004.pdb	-1.7117147282668639

7.3 Post MD calculations.

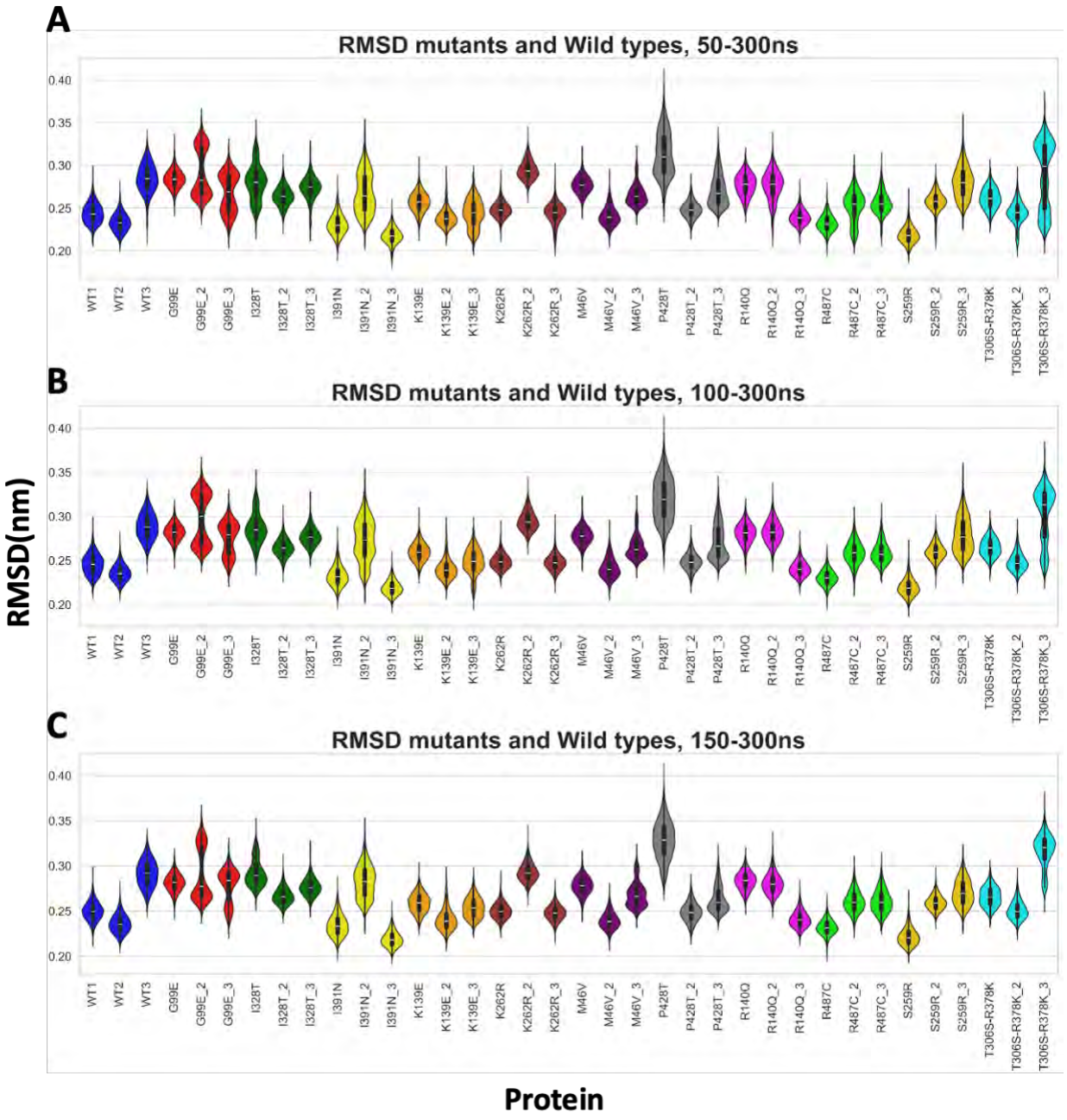


Figure S1: Violin plots showing the RMSD distribution. A) represents 50ns to 300ns **B)** represents 100ns to 300ns**. C)** represents 150ns to 300ns. Each system is colour coded together. The underscore (_2 and _3) represents the second and third run that occurred.

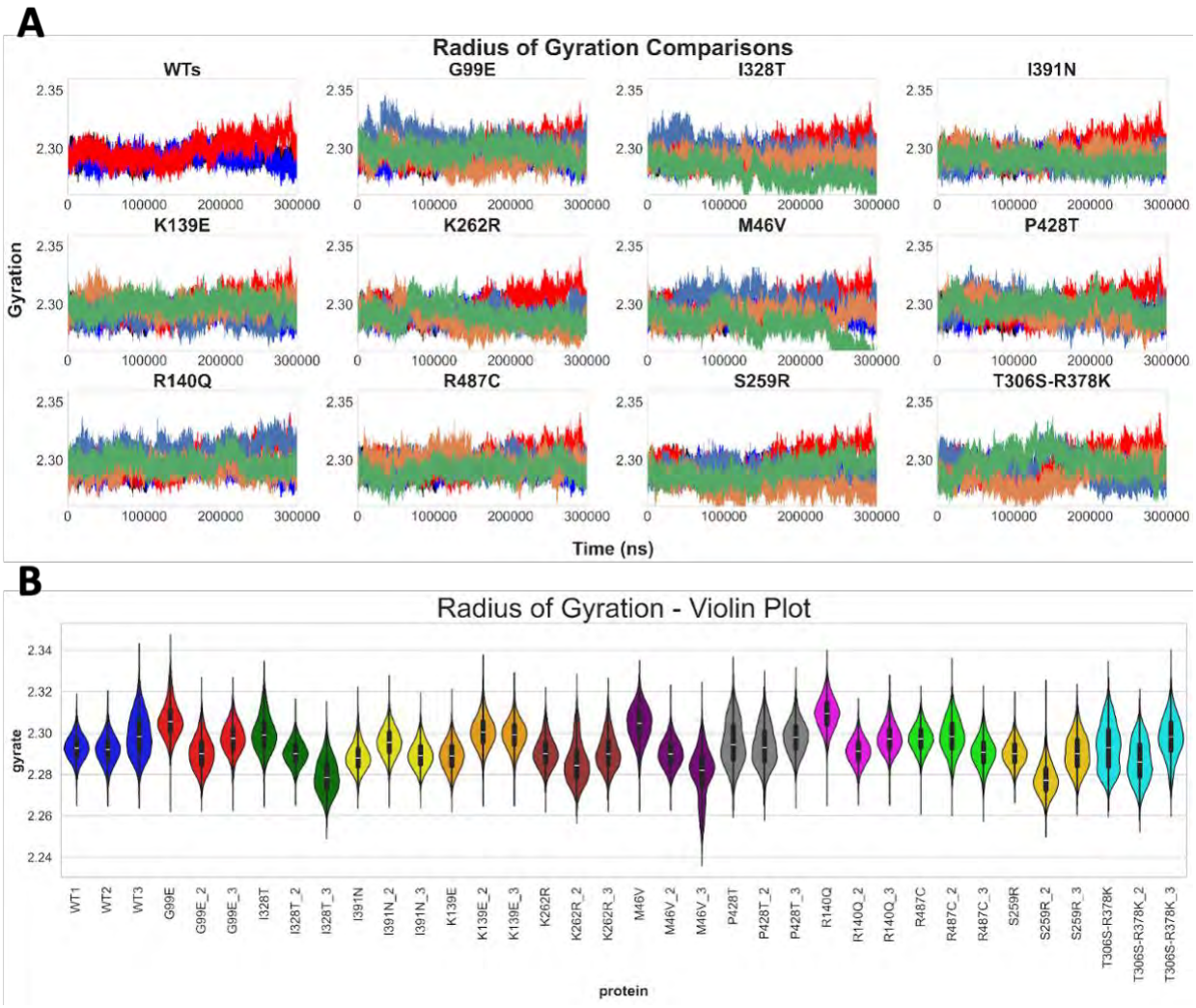
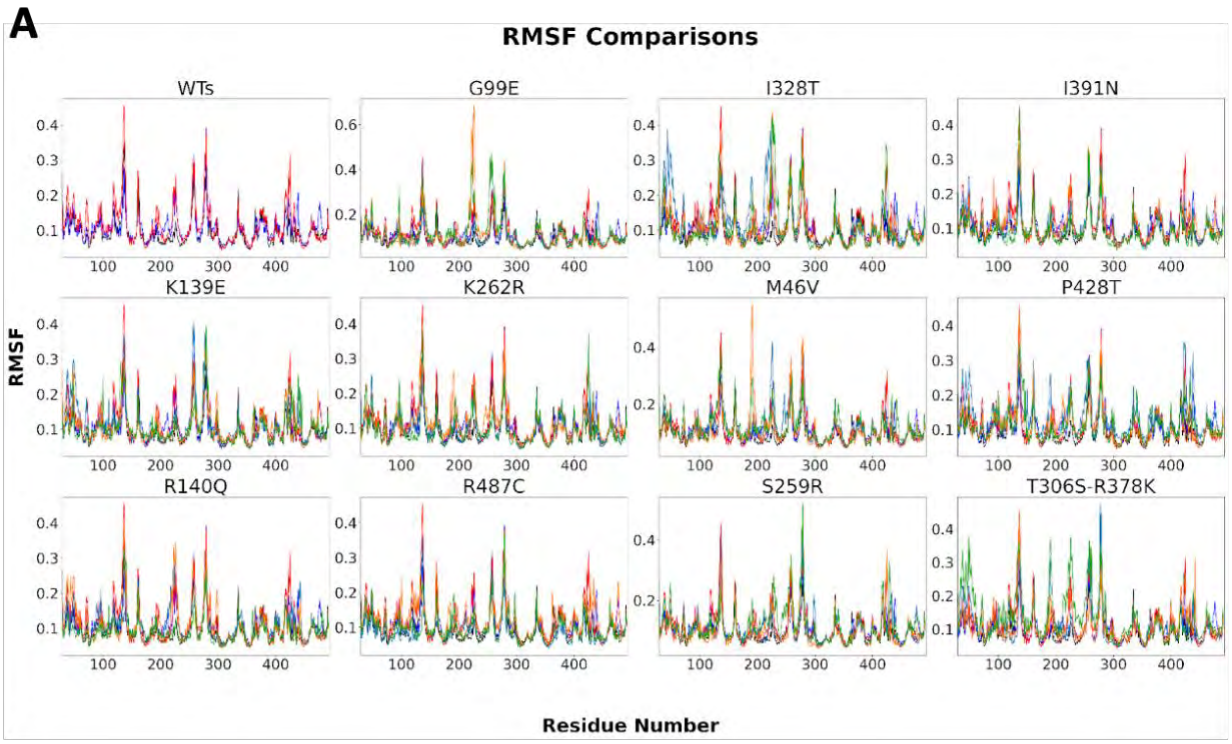


Figure S2: Radius of Gyration (Rg) profiles on the enzyme CYP2B6 - References and Variant Systems. **A)** Line plots displaying compactness of all systems. Reference1: Black, Reference2: Blue, Reference3: Red, Run1: Navy, Run2: Orange, Run3: Green. **B)** Violin pots assessing the Rg conformation throughout the simulation - each variation and reference pair is colour coded together.


A
RMSF Comparisons
WTs
G99E
I328T
I391N
K139E
K262R
M46V
P428T
R140Q
R487C
S259R
T306S-R378K
RMSF
Residue Number

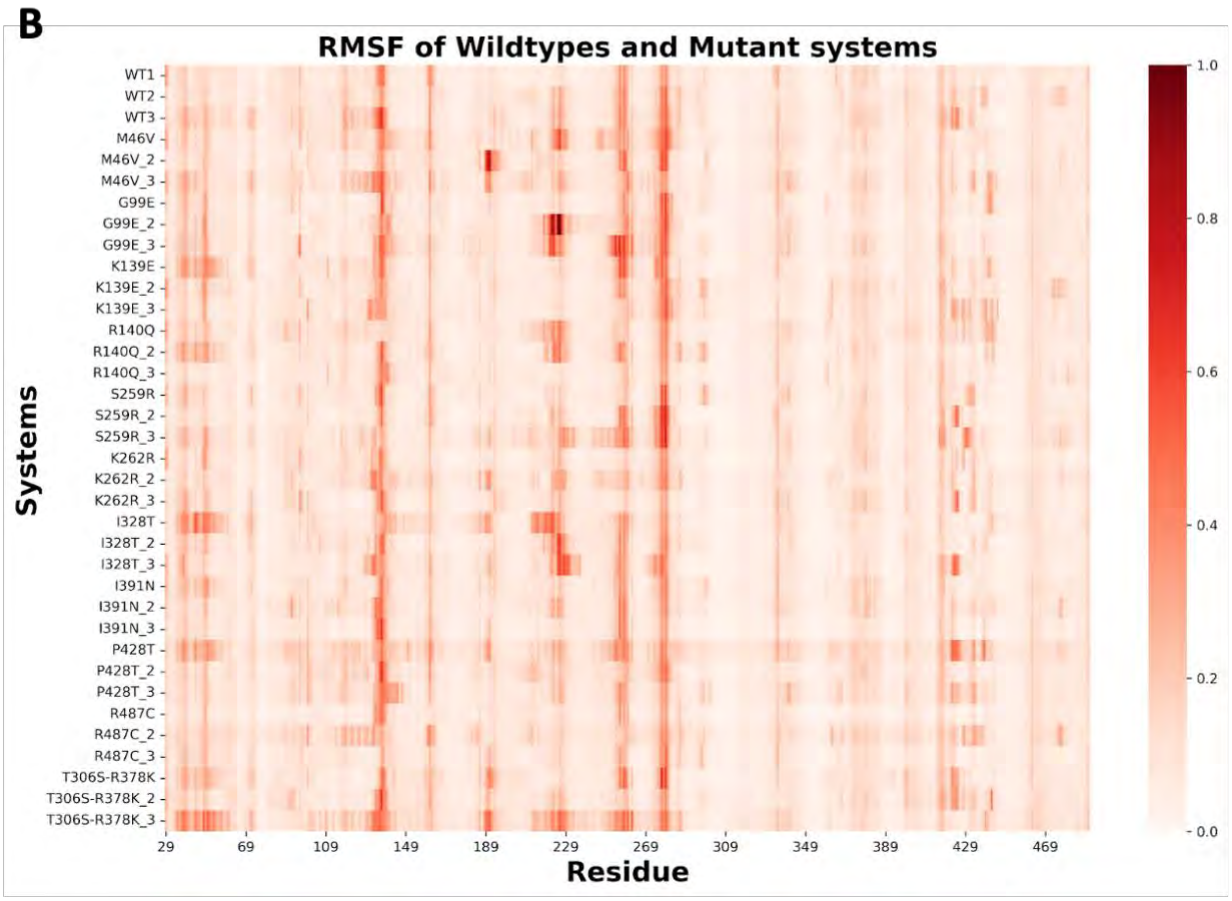

B
RMSF of Wildtypes and Mutant systems
WT1
WT2
WT3
M46V
M46V_2
M46V_3
G99E
G99E_2
G99E_3
K139E
K139E_2
K139E_3
R140Q
R140Q_2
R140Q_3
S259R
S259R_2
S259R_3
K262R
K262R_2
K262R_3
I328T
I328T_2
I328T_3
I391N
I391N_2
I391N_3
P428T
P428T_2
P428T_3
R487C
R487C_2
R487C_3
T306S-R378K
T306S-R378K_2
T306S-R378K_3
Systems
Residue

Figure S3: Root mean square fluctuations (RMSD) profiles for all systems of CYP2B6. A) Line plots displaying RMSF values. Reference1: Black, Reference2: Blue, Reference3: Red, Run1: Navy, Run2 : Orange, Run3 : Green**. B)** In The RMSF heatmap, regions of higher fluctuations are of darker shades

7.4 Clustering analysis

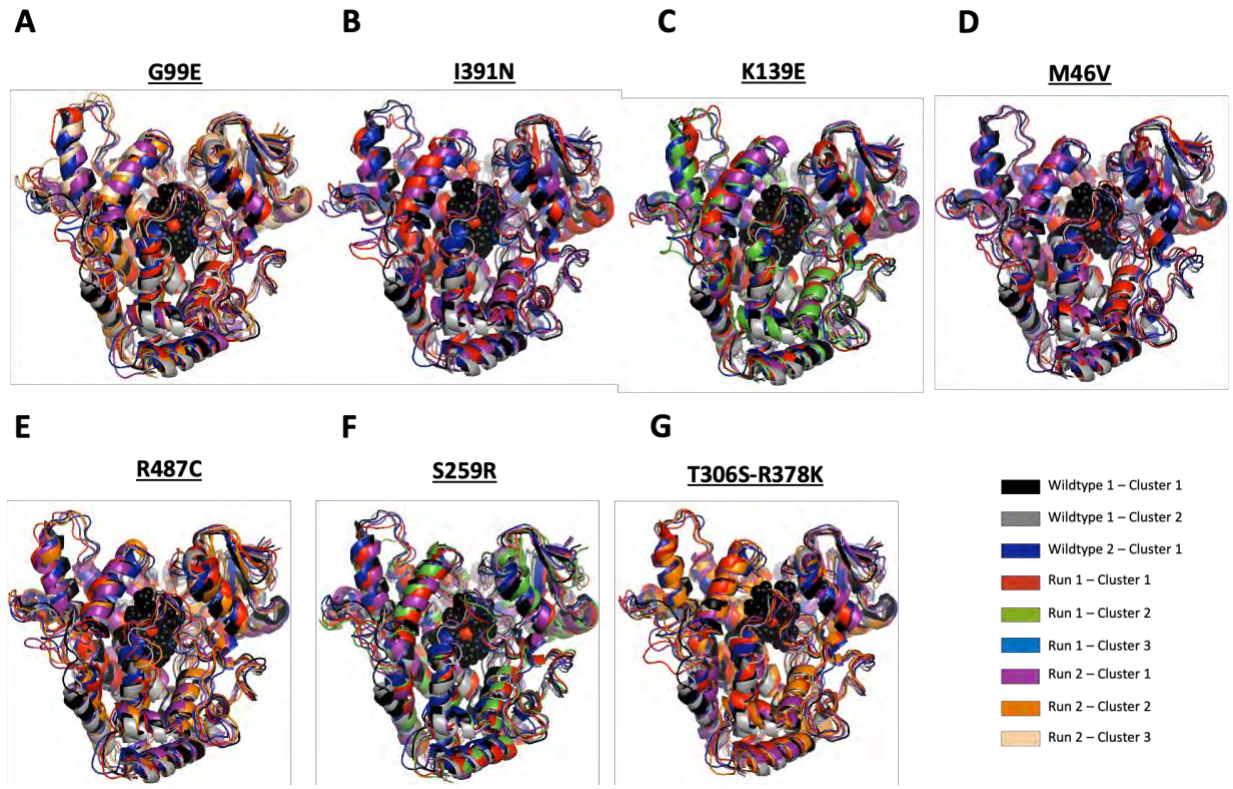


Figure S4: Clustering on the references and duplicate runs of the variations: A) G99E, **B)** I391N, **C)** K139E, **D)** M46V, **E)** R487C, **F)** S259R, and **G)** T306S-R378K. The heme is represented in black spheres with the iron represented by a red sphere that can be identified in the middle of the heme.

Table S2: RMSD values from cluster analysis after alignment. Displaying the RMSD values that occurred after superimposition.

	WT_1.1	WT_1.2	WT_2
WT_1.1	0	1.1780332326889038	2.2249345779418945

WT_1.2	1.1780332326889038	0	2.366323947906494
WT_2.1	2.2249345779418945	2.366323947906494	0
G99E_1.1	1.9562525749206543	2.1997323036193848	1.713600516319275
G99E_2.1	2.2087466716766357	2.4416441917419434	1.9942632913589478
G99E_2.2	2.081141471862793	2.225414752960205	1.8399174213409424
G99E_2.3	2.1733927726745605	2.4033243656158447	1.9724764823913574
I328T_1.1	2.297619581222534	2.3658859729766846	2.3778388500213623
I328T_1.2	2.4407460689544678	2.4277737140655518	2.361353635787964
I328T_2.1	3.2105865478515625	3.300628900527954	2.8873353004455566
I328T_2.2	3.0202317237854004	3.077108144760132	2.6419806480407715
I328T_2.3	2.8329761028289795	2.975048065185547	2.3416993618011475
I391N_1.1	2.3689284324645996	2.394439935684204	2.3497183322906494
I391N_2.1	2.4398584365844727	2.6297390460968018	1.8842804431915283
K139E_1.1	1.997860312461853	2.154240608215332	2.079613208770752
K139E_1.2	2.1673166751861572	2.3003005981445312	1.9453147649765015
K139E_2.1	2.2520761489868164	2.4391281604766846	2.049680471420288
K262R_1.1	2.434234619140625	2.7476983070373535	2.040872812271118
K262R_2.1	2.3810651302337646	2.5068857669830322	1.8233450651168823
K262R_2.2	2.104682207107544	2.318471908569336	1.7083606719970703
M46V_1.1	1.9515459537506104	2.1044280529022217	1.8507075309753418
M46V_2.1	2.1999175548553467	2.338104724884033	1.881352186203003
P428T_1.1	2.8095543384552	2.9287805557250977	2.44211483001709
P428T_1.2	2.545545816421509	2.631225347518921	2.1654584407806396
P428T_2.1	2.825155735015869	2.8317556381225586	2.354172706604004
R140Q_1.1	2.2815566062927246	2.3784215450286865	1.9763684272766113

R140Q_2.1	2.1801071166992188	2.2578718662261963	1.9482064247131348
R140Q_2.2	2.5159521102905273	2.781895160675049	2.101839542388916
R487C_1.1	2.149214744567871	2.2522084712982178	1.7890691757202148
R487C_2.1	2.32212233543396	2.43384051322937	1.8672212362289429
R487C_2.2	2.0351390838623047	2.1363046169281006	2.1143224239349365
S259R_1.1	2.252864360809326	2.3248581886291504	1.7187094688415527
S259R_1.2	2.2305171489715576	2.3243086338043213	2.060814142227173
S259R_2.1	2.0741453170776367	2.2097203731536865	1.8878728151321411
T306S-R378K_1.1	2.125882625579834	2.361999273300171	1.8642950057983398
T306S-R378K_2.1	1.9995707273483276	1.9996315240859985	2.047027587890625
T306S-R378K_2.2	2.374229907989502	2.428084135055542	1.9655917882919312

7.5 DSSP analysis

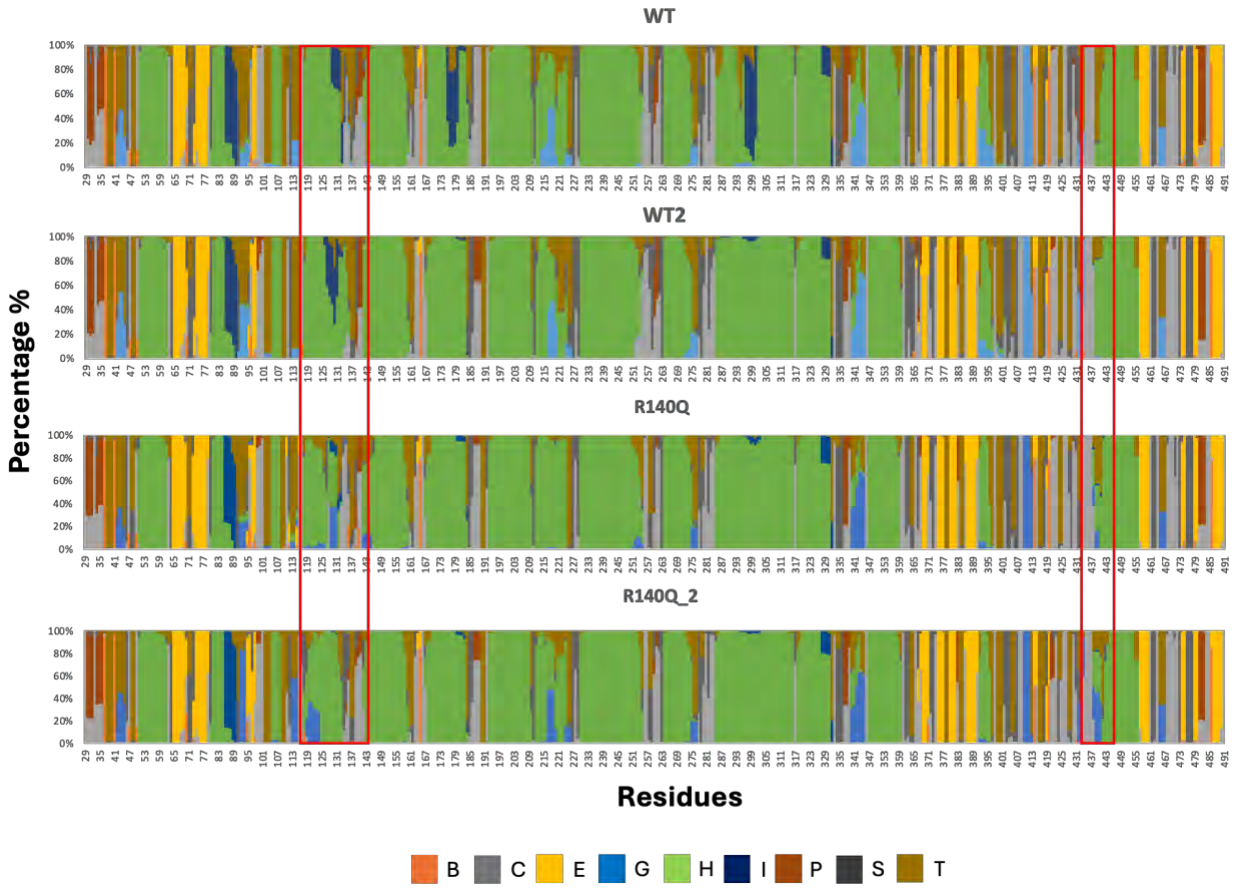


Figure S5: DSSP representing references and variation *R140Q*. The colours are associated with secondary structural elements. The B - residue in isolated beta bridge, C - loop region, E - extended strand that participates in beta-ladder, G - 3_10-helix, I - pi-helix, P - poly-proline II helix (kappa helix), S - bend, T - hydrogen bond turn. The red outlined box represents where the differences occur.

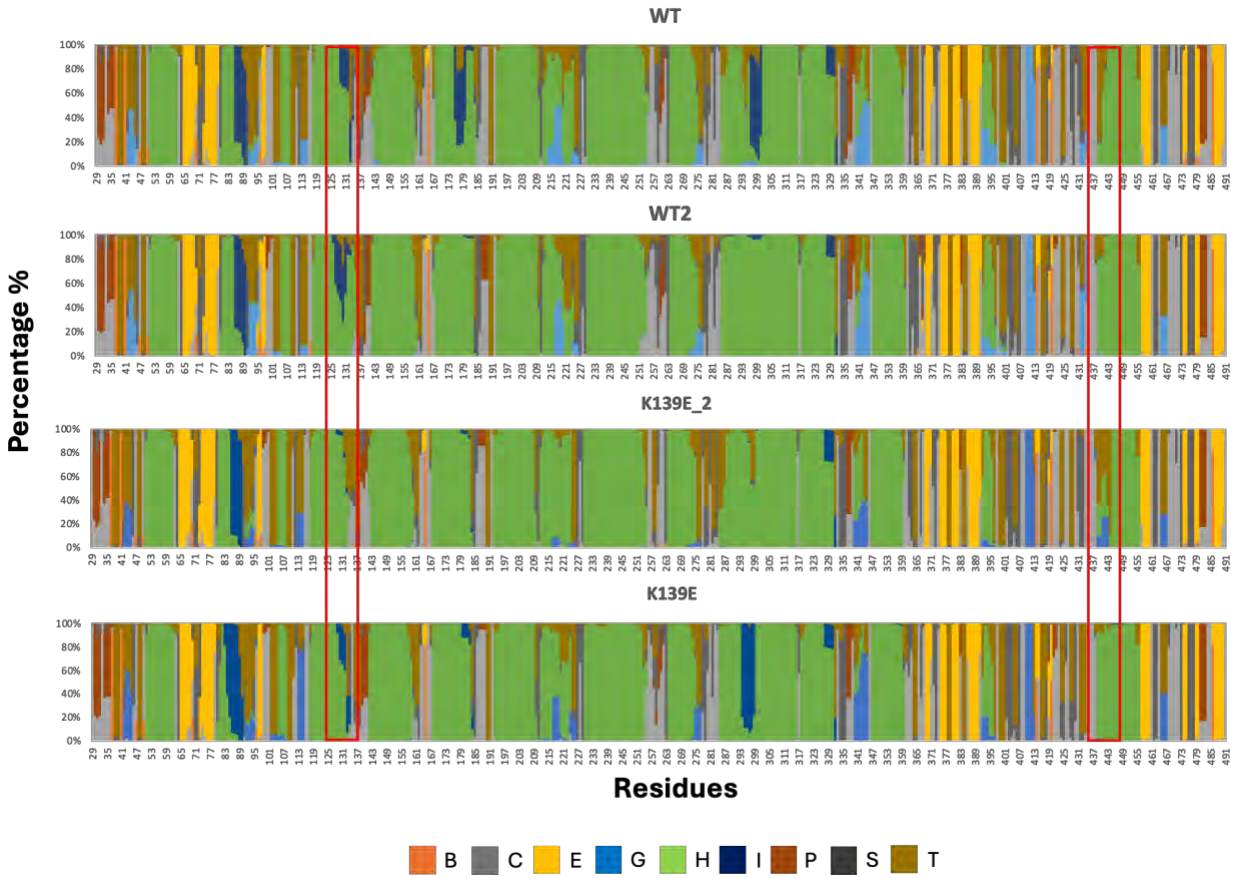


Figure S6: DSSP representing references and variation *K139E* duplicate runs. The colours are associated with secondary structural elements. The B - residue in isolated beta bridge, C - loop region, E - extended strand that participates in beta-ladder, G - 3_10-helix, I - pi-helix, P - poly-proline II helix (kappa helix), S - bend, T - hydrogen bond turn. The red outlined box represents where the differences occur.

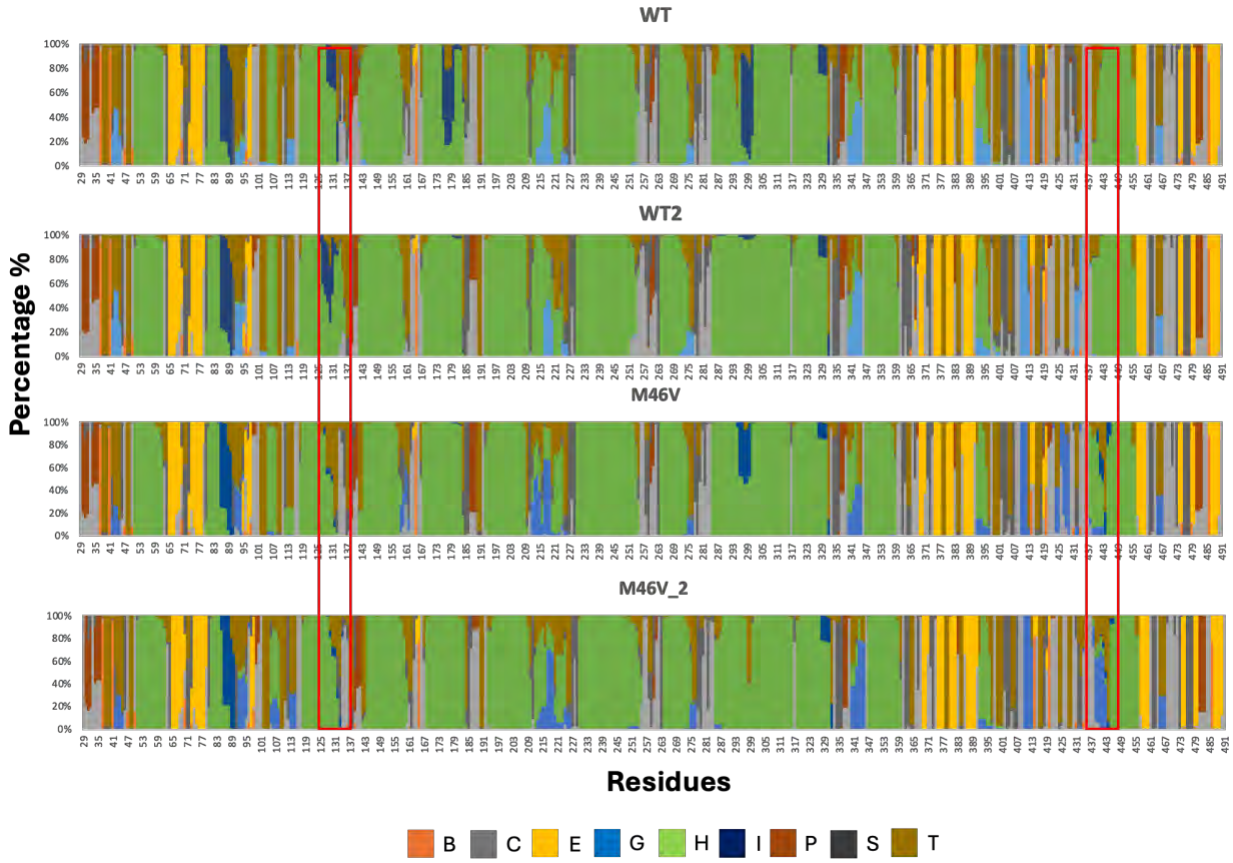


Figure S7: DSSP representing references and variation *M46V* duplicate runs. The colours are associated with secondary structural elements. The B - residue in isolated beta bridge, C - loop region, E - extended strand that participates in beta-ladder, G - 3_10-helix, I - pi-helix, P - poly-proline II helix (kappa helix), S - bend, T - hydrogen bond turn. The red outlined box represents where the differences occur.

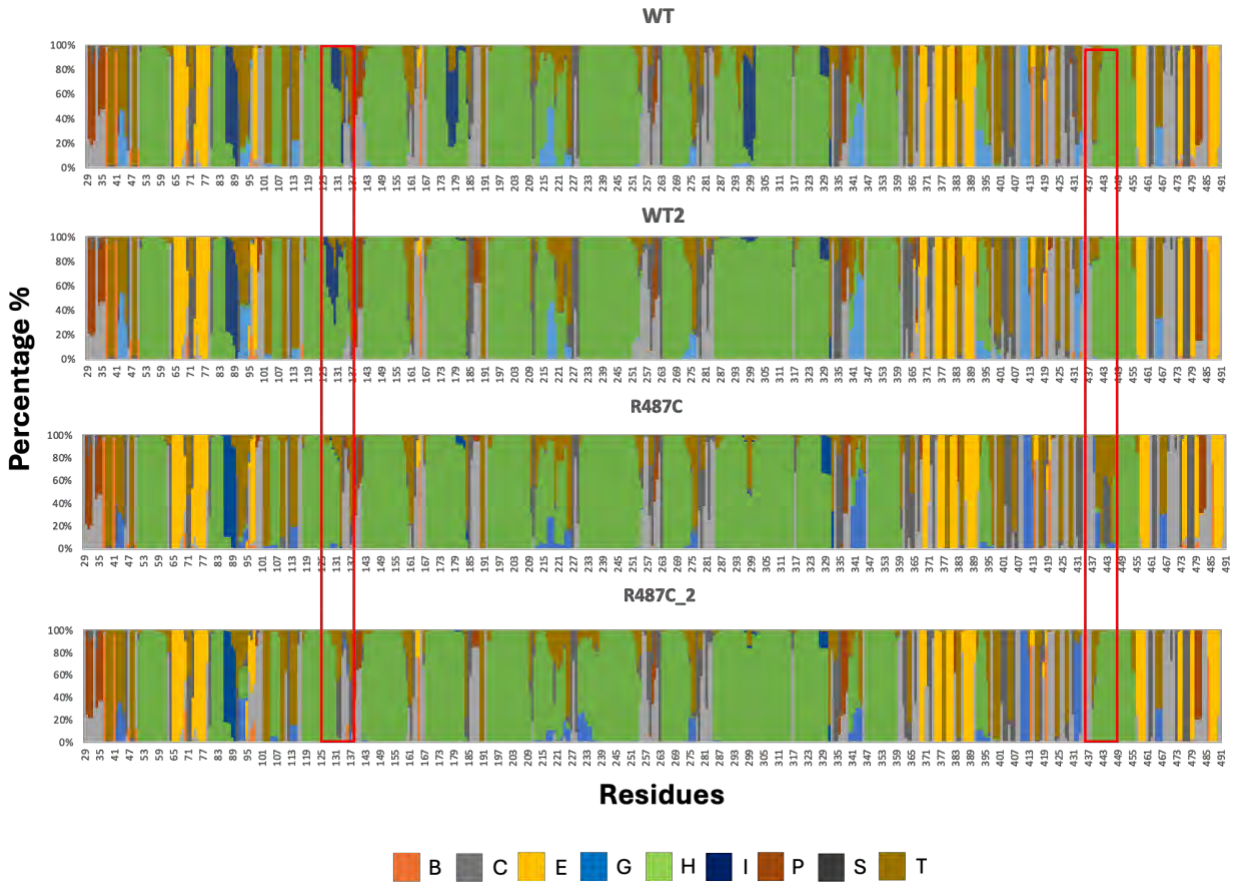


Figure S8: DSSP representing references and variation *R487C* duplicate runs. The colours are associated with secondary structural elements. The B - residue in isolated beta bridge, C - loop region, E - extended strand that participates in beta-ladder, G - 3_10-helix, I - pi-helix, P - poly-proline II helix (kappa helix), S - bend, T - hydrogen bond turn. The red outlined box represents where the differences occur.

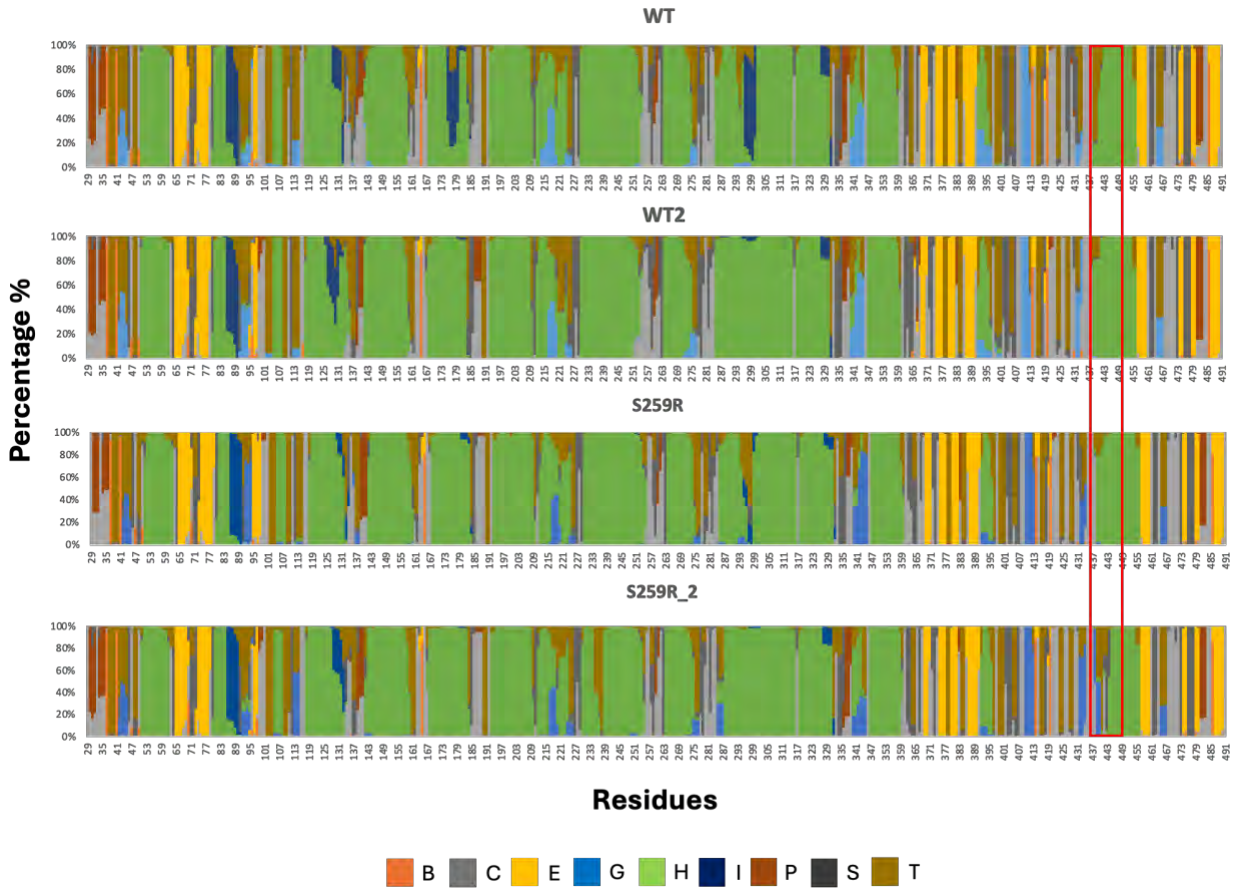


Figure S9: DSSP representing references and variation *S259R* duplicate runs. The colours are associated with secondary structural elements. The B - residue in isolated beta bridge, C - loop region, E - extended strand that participates in beta-ladder, G - 3_10-helix, I - pi-helix, P - poly-proline II helix (kappa helix), S - bend, T - hydrogen bond turn. The red outlined box represents where the differences occur.

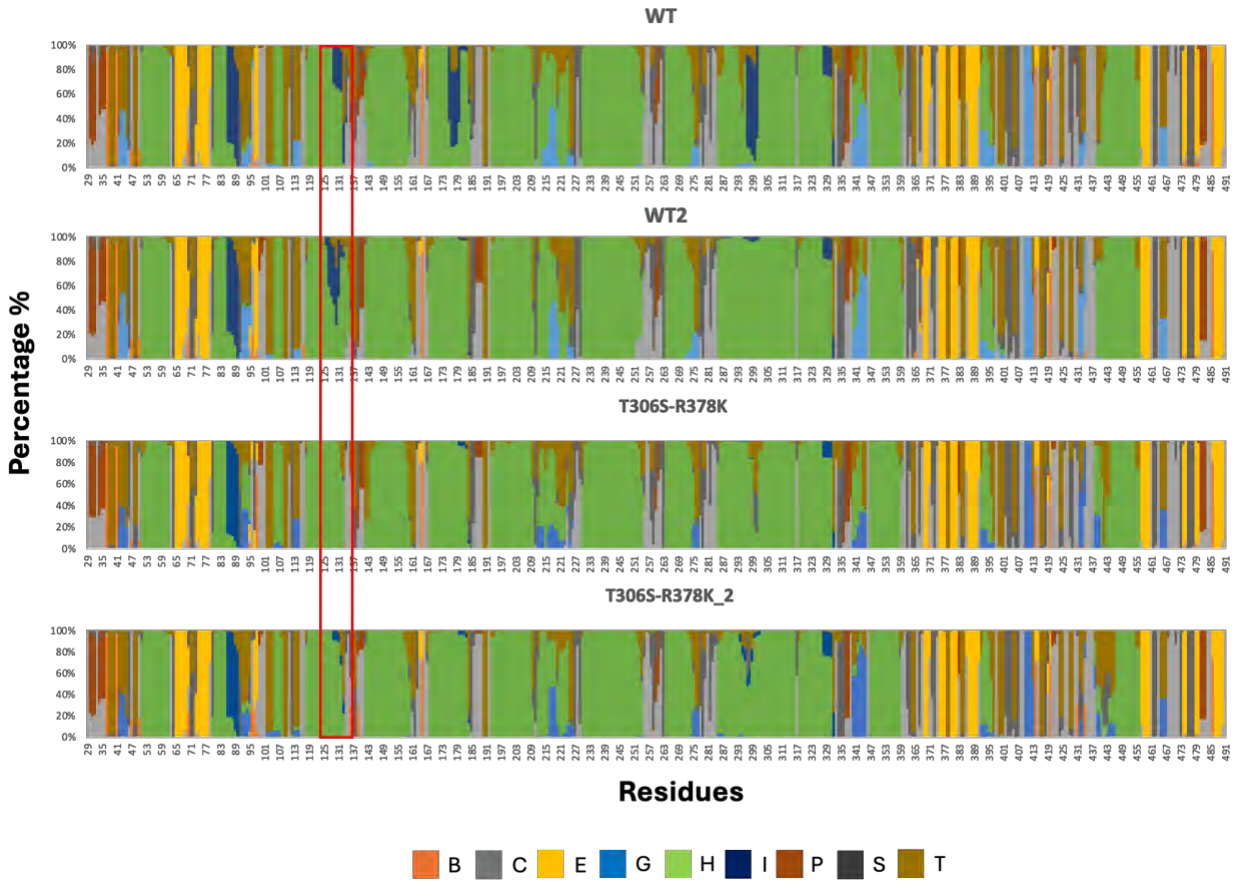


Figure S10: DSSP representing references and variation *T306S-R378K* duplicate runs. The colours are associated with secondary structural elements. The B - residue in isolated beta bridge, C - loop region, E - extended strand that participates in beta-ladder, G - 3_10-helix, I - pi-helix, P - poly-proline II helix (kappa helix), S - bend, T - hydrogen bond turn. The red outlined box represents where the differences occur.

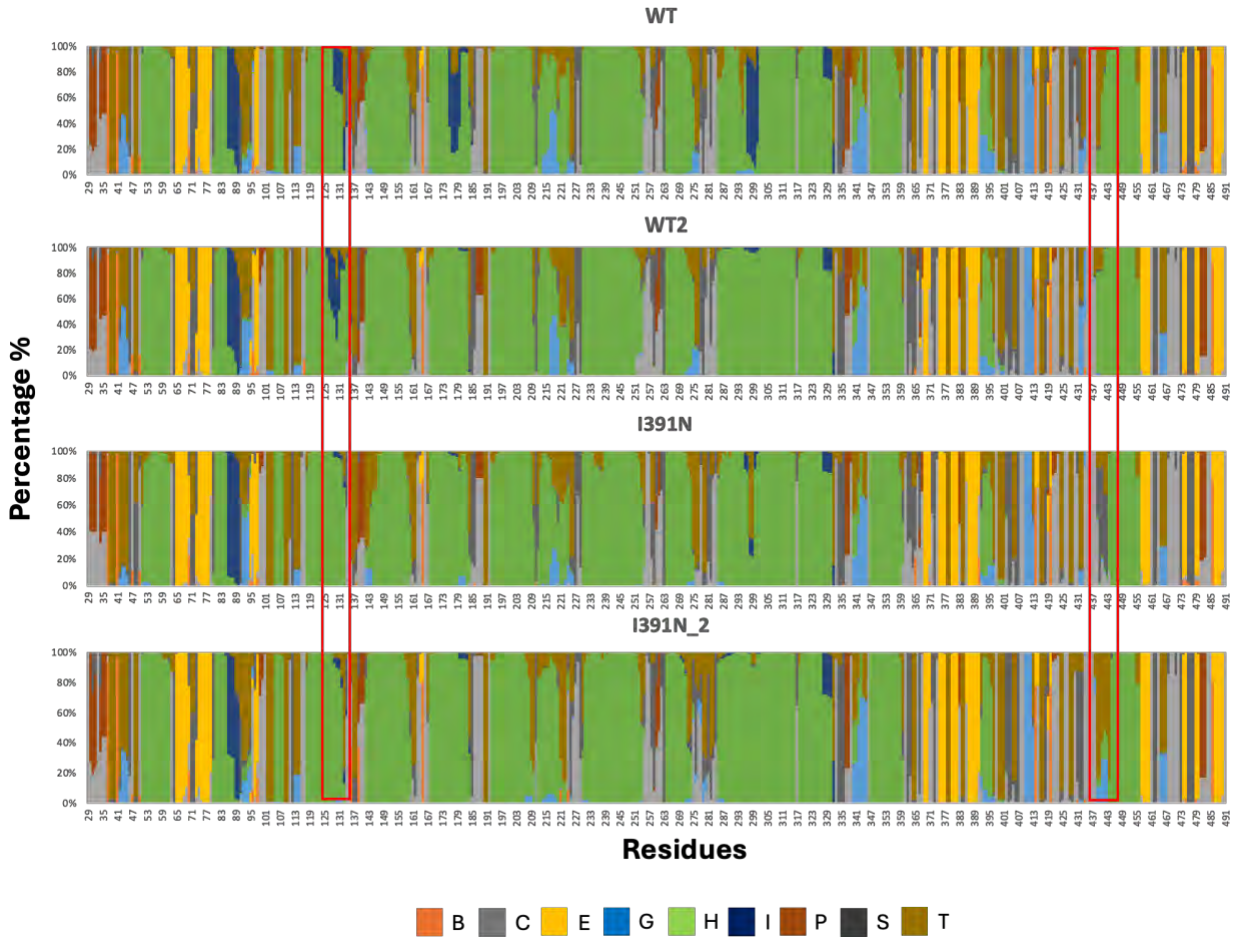


Figure S11: DSSP representing references and variation *I391N* duplicate runs. The colours are associated with secondary structural elements. The B - residue in isolated beta bridge, C - loop region, E - extended strand that participates in beta-ladder, G - 3_10-helix, I - pi-helix, P - poly-proline II helix (kappa helix), S - bend, T - hydrogen bond turn. The red outlined box represents where the differences occur.

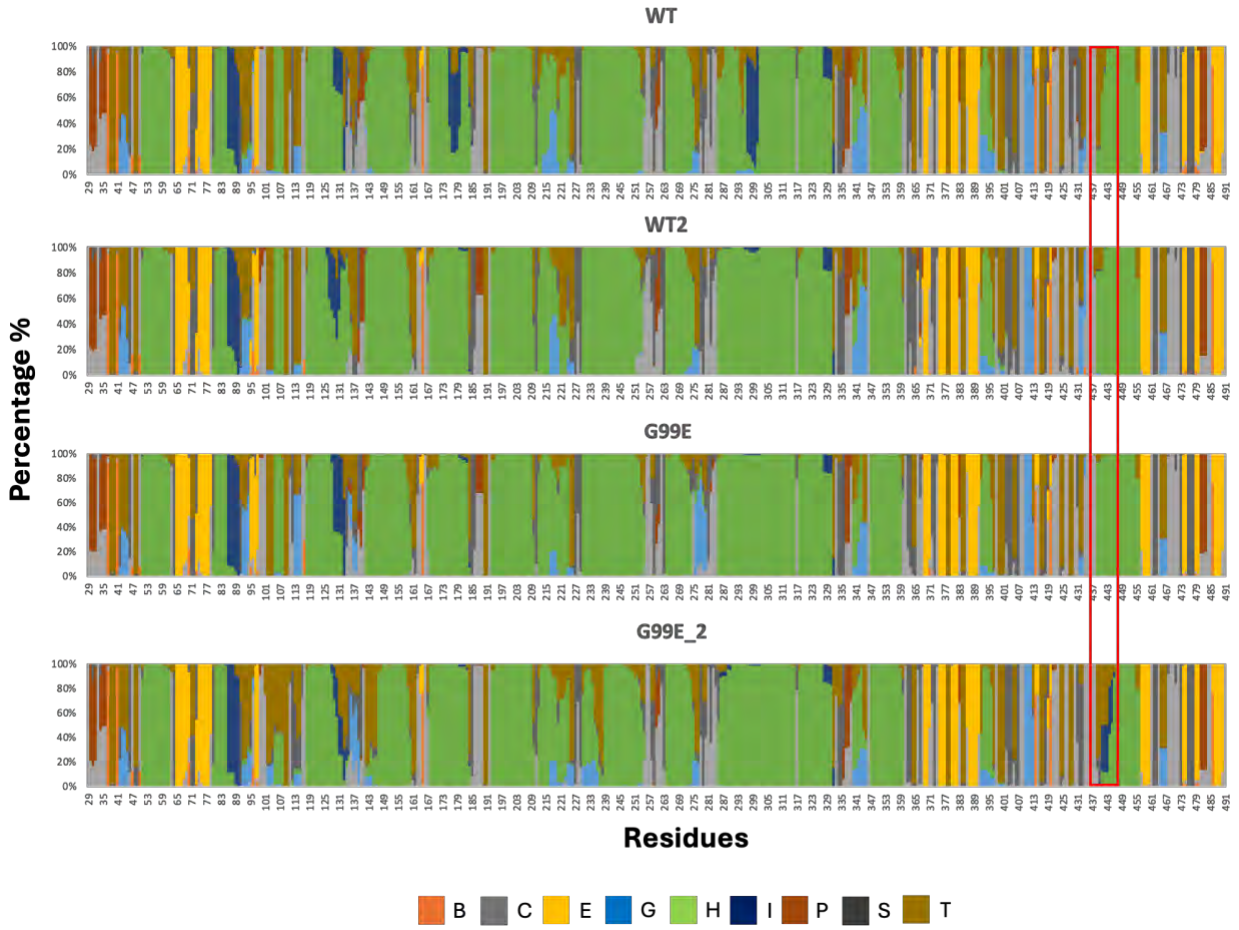


Figure S12: DSSP representing references and variation *G99E* duplicate runs. The colours are associated with secondary structural elements. The B - residue in isolated beta bridge, C - loop region, E - extended strand that participates in beta-ladder, G - 3_10-helix, I - pi-helix, P - poly-proline II helix (kappa helix), S - bend, T - hydrogen bond turn. The red outlined box represents where the differences occur.

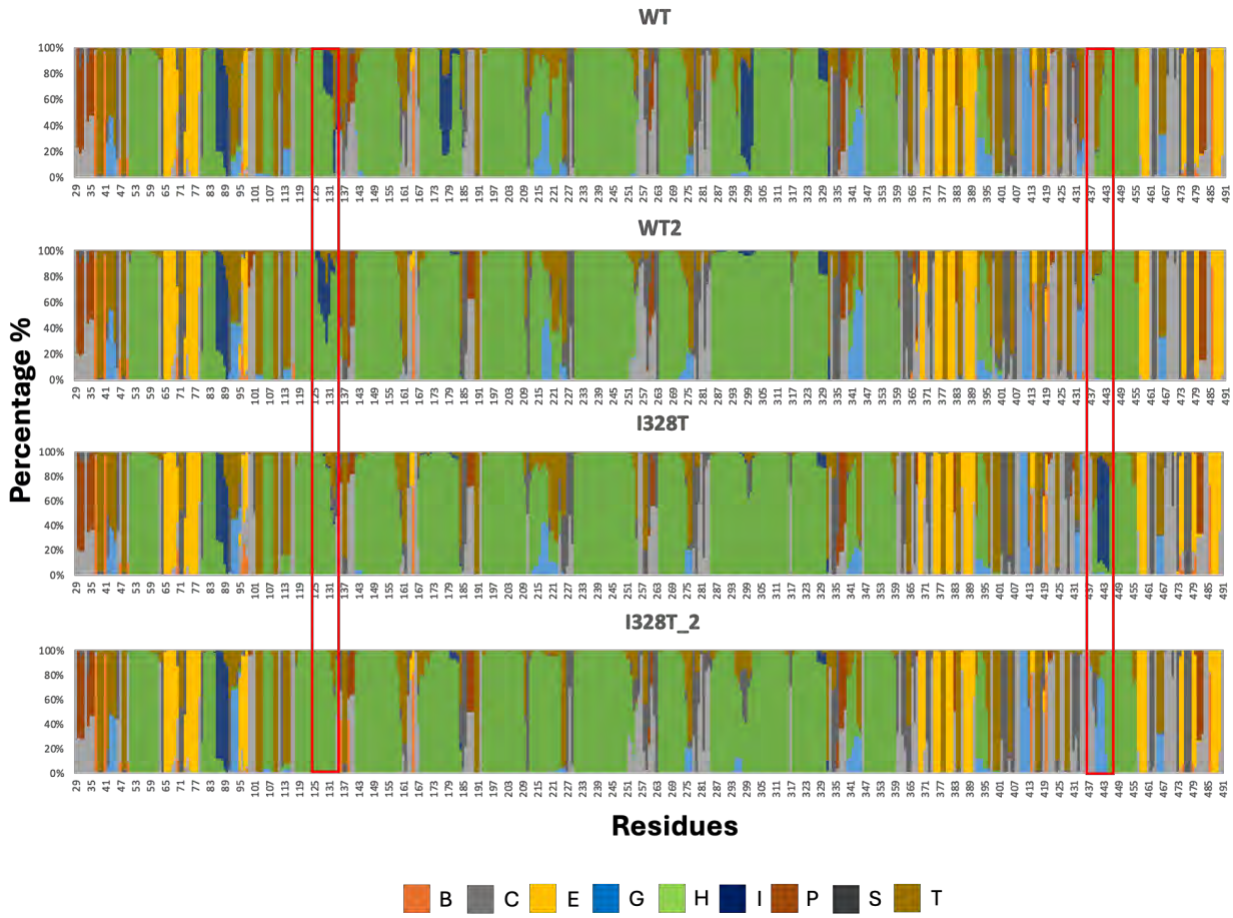


Figure S13: DSSP representing references and variation *I328T*. The colours are associated with secondary structural elements. The B - residue in isolated beta bridge, C - loop region, E - extended strand that participates in beta-ladder, G - 3_10-helix, I - pi-helix, P - poly-proline II helix (kappa helix), S - bend, T - hydrogen bond turn. The red outlined box represents where the differences occur.

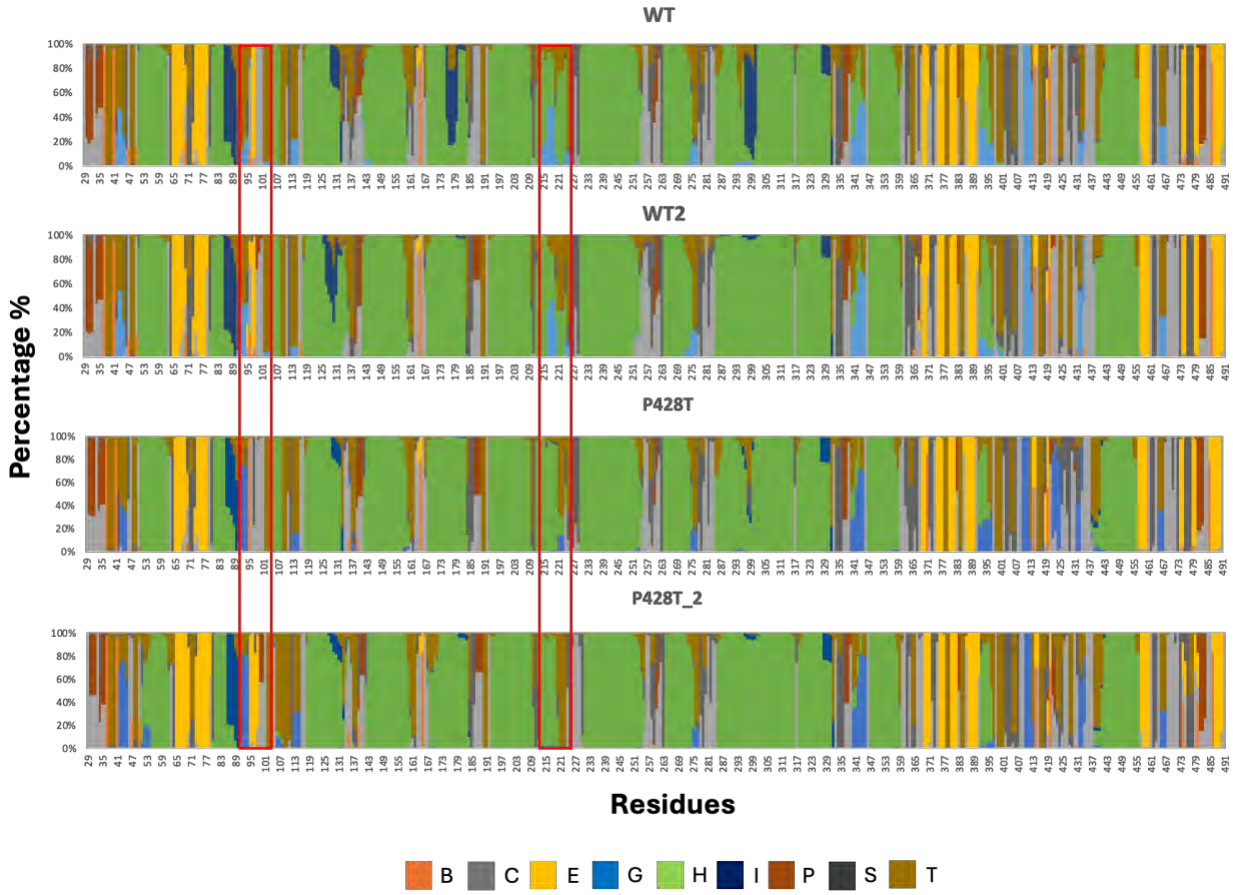


Figure S14: DSSP representing references and variation *P428T*. The colours are associated with secondary structural elements. The B - residue in isolated beta bridge, C - loop region, E - extended strand that participates in beta-ladder, G - 3_10-helix, I - pi-helix, P - poly-proline II helix (kappa helix), S - bend, T - hydrogen bond turn. The red outlined box represents where the differences occur.