

https://dspace.org/

Characterising the combined eects of cytochrome P450 missense variants within the star allele nomenclature

Khoza, Nhlamulo

2024
University of the Witwatersrand, Johannesburg
https://hdl.handle.net/10539/43072
Khoza, Nhlamulo. (2024). Characterising the combined eects of cytochrome P450 missense variants within the star allele nomenclature [Master's dissertation, University of the Witwatersrand, Johannesburg]. WireDSpace.

Downloaded from a DSpace repository. https://dspace.org/

Characterising the combined effects of cytochrome P450 missense variants within the star allele nomenclature

Nhlamulo Khoza

Student number: 1620584

Supervisors:

Doctor Houcemeddine Othman

Professor Scott Hazelhurst



UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

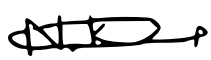
Faculty of Health Sciences

School of Pathology

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

Declaration

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



(Signature of Candidate)

12th day of February 2024 in Johannesburg, South Africa.

Acknowledgement

To my supervisors, Houcemeddine and Scott, thank you very much for the support and guidance. Working with you in this captivating field has been a tremendous opportunity, and I sincerely thank you for it.

To David, Jorge, Blessing, the ADME team and every student at Sydney Brenner Institute for Molecular Bioscience (SBIMB), thank you for your support, encouragement, and thoughtful interactions throughout this journey. Thanks to the NHLS, SBIMB and University of the Witwatersrand for the opportunity to pursue this MSc study.

My deepest gratitude to my family and friends for their unwavering support and encouragement throughout my MSc journey. Thank you for being a pillar of strength, providing a listening ear, and for your constant motivation. Your support has been the driving force behind my accomplishments.

Financial and computational support

Thank you to the funders. This research received external funding from GlaxoSmithKline (GSK) under the Wits-H3A/GSK ADME collaboration and N Khoza was supported by The National Research Foundation (NRF) of South Africa (grant number: 140383). The funders were not responsible for the study design, and the views and opinions expressed in the publications arising from this dissertation are not necessarily those of the funders. I would like to acknowledge the Centre for High-Performance Computing (CHPC), South Africa, for generously providing computational resources for this research project.

Presentations arising from this dissertation

1. International Congress of Human Genetics (2023) poster presentation
   Title: Characterising the combined effects of cytochrome P450 missense variants within star allele definitions

2. Centre for High-Performance Computing conference (2022) oral presentation
   Title: Characterising the combined effects of cytochrome P450 missense variants within star allele definitions

3. MBRT symposium (2022) oral presentation
   Title: Characterising the combined effects of cytochrome P450 missense variants within star allele definitions

4. SASBi-SAGS conference (2022) poster presentation
   Title: Characterising the combined effects of cytochrome P450 missense variants within the star allele nomenclature

5. SASBi-SAGS student symposium (2022) oral presentation
   Title: Characterising the combined effects of cytochrome P450 missense variants within the star allele nomenclature

Publication arising from this dissertation

Khoza, N., Twesigomwe, D., & Othman, H. (2023). Characterizing the combined effects of cytochrome P450 missense variation within star allele definitions. *Pharmacogenomics*, 24(10), 561-578. doi: 10.2217/pgs-2023-0068. PMID: 37503750.

Contents

Abstract viii

1 Introduction 1

1.1 Problem statement . . . 2

1.2 Research aim and objectives . . . 3

1.3 Dissertation outline . . . 3

2 Literature review 5

2.1 The star allele nomenclature . . . 5

2.2 Structure of human Cytochrome P450 proteins . . . 6

2.3 Clinical application . . . 8

2.4 Pharmacogenomics in the context of African populations . . . 9

2.5 Structural bioinformatics techniques . . . 10

2.5.1 Protein modelling . . . 11

2.5.2 Variant effect prediction . . . 13

2.5.3 Normal Mode Analysis . . . 14

2.5.4 Molecular dynamics simulation . . . 16

2.5.5 C-alpha and CYP protein structural landscape . . . 17

3 Characterising the combined effects of cytochrome P450 missense variation within star allele definitions 20

3.1 Summary . . . 20

3.2 Introduction . . . 21

3.3 Materials and Methods . . . 23

3.3.1 Data . . . 23

3.3.2 Normal mode analysis . . . 23

3.3.3 Calculating Gibbs free energy of folding . . . 25
3.3.4 Comparison of NMA with Deep Mutational Scanning and other approaches . . . 26
3.3.5 Molecular dynamics simulation . . . 27
3.4 Results . . . 28
3.4.1 Comparison with computational variant effect predictors . . . 30
3.4.2 Comparison with Deep Mutational Scanning . . . 30
3.4.3 Comparison with star allele CPIC clinical functions . . . 33
3.4.4 Molecular dynamics simulation of *CYP2D6*2* and *CYP2D6*17* . . . 34
3.5 Discussion . . . 37
3.6 Conclusion . . . 42

4 Discussion and conclusion **66**
4.1 Consolidated findings . . . 66
4.2 Emerging themes . . . 68
4.2.1 Utility of structural bioinformatics in variant analysis . . . 69
4.2.2 Mechanisms of variant interaction in CYP star alleles . . . 70
4.3 Contextual relevance . . . 71
4.4 Strengths and limitations . . . 71
4.5 Future work and directions . . . 73
4.6 Conclusion . . . 75

Appendices **91**
Appendix A: Ethics waiver certificate . . . 92
Appendix B: Authors' agreement . . . 93
Appendix C: Plagiarism declaration and similarity index . . . 94

List of Figures

2.1 Examples of *CYP2D6* star alleles. 6
2.2 Structure of a CYP protein. 7
2.3 Illustration of a protein folding energy landscape 12
2.4 Illustration of coarse-graining 15
2.5 Broad range of equilibrium motions accessible under native state conditions. 18

3.1 An overview of the employed methodologies and their interrelationships. 24
3.2 Mapping of single amino acid variants on CYP P450 proteins. 25
3.3 Distribution of Fold-X ($\Delta\Delta G$) and ENCoM-computed ($\Delta\Delta S_{\text{vib}}$) missense variant scores and SWAAT red-flag protein structural feature count. . . . 29
3.4 Assessment of the concordance of CPIC clinical functional classification with NMA predications. 34
3.5 Analysis of backbone protein structural deviations by selected CYP2D6 variants and haplotypes. 36
3.6 Analysis of residue-level fluctuations induced by selected CYP2D6 variants and haplotypes. 37
3.7 Principal Component Analysis from Molecular Dynamics simulations. . . 38
S1 Comparison of ENCoM-computed ($\Delta\Delta S_{\text{vib}}$) and FoldX-computed ($\Delta\Delta G$) star allele and missense variant energy scores. 44

List of Tables

3.1 Comparison of NMA-based CYP2C9 and CYP2C19 missense variant effect predictions with deep mutational scanning results. 30
S1 Missense variant NMA analysis. 45
S2 CYP star allele/haplotype NMA analysis 56
S3 SWAAT analysis of CYP missense variant structural consequence. 62

Abstract

Genetic variations in Cytochrome P450 (CYP) enzymes shape drug responses. However, many CYP haplotypes (star alleles) lack functional annotations, posing a barrier to understanding drug metabolism comprehensively. To address this, our study investigates combined missense variant effects on CYP enzyme structures, analyzing 261 variants across 91 CYP haplotypes. We utilized Normal Mode Analysis (NMA; FoldX and ENCoM) to explore variant impact on protein stability. Subsequently, we conducted Molecular Dynamics (MD) simulations on key alleles, *CYP2D6*2* and *CYP2D6*17*, to reveal star allele impact on protein dynamics. Integrating NMA and MD, we uncover the interactions that collectively shape the conformation and attributes of CYP enzymes. Notably, our investigation highlights significant deviations between wild-type and *CYP2D6*17*-encoded proteins in the F/G loop region, pivotal for CYP functionality. Additionally, we compare NMA results with CYP2C9 and CYP2C19 Deep Mutational Scanning (DMS) results, identifying 65% concordance. Furthermore, our NMA predictions show 80% concordance with commonly used Variant Effect Predictor tools. This alignment underscores our approach's reliability in predicting variant effects. Our study illuminates missense variants' nuanced impact on CYP protein structures, significant for personalized medicine and drug response prediction, as accurate drug response predictions hinge on a comprehensive understanding of CYP missense variants. Moreover, our study highlights multi-scale modelling potential for interpreting CYP missense variants, especially in star alleles. The synergy of NMA, MD simulation, and assays like DMS is invaluable, each with distinct strengths and limitations. This research deepens our understanding of the complexity of CYP metabolism profiles, providing insights into the functional assessment of CYP star alleles and missense variants with unknown functional classifications.

Chapter 1

Introduction

The "post-genomic era" has brought an extensive understanding of the mechanisms underlying the interindividual genetic variation (Collins et al. 2003, Meystre & Gouripeddi 2019). One of the many researched genomic aspects is how proteins can have varied functions resulting from variations in the gene that encodes them (Smith et al. 2021). Proteins regulate many biological mechanisms such as immune response, pH regulation, molecular storage and transportation, and drug response. Proteins that are involved in drug response mechanisms are classified as ADME proteins. They are involved in at least one of the drug-related mechanisms: Absorption, Distribution, Metabolism, or Excretion (ADME) of drugs (Fujikura et al. 2015). Pharmacogenomics is a branch of genetics that is primarily concerned with how genetic variation affects protein function (among other features) in the context of drug response.

Genetic variation may be in the form of single nucleotide polymorphisms (SNPs), short insertions/deletions (indels), and various structural variants. Generally, SNPs and indels tend to be inherited within various combinations/blocks referred to as haplotypes. In pharmacogenomics, haplotypes and structural variants are commonly represented via a star allele nomenclature. For example, *CYP2D6*17* is defined by a combination of missense variants T107I, S486T, and R296C. Investigating the interaction of variants in a haplotype may provide insight into another level of the drug metabolism mechanism. Understanding the protein phenotypes produced by various combinations of missense variation has great potential to enhance our understanding of interindividual variable drug response and pave the path for precision medicine applications.

The cytochrome P450 (CYP) protein superfamily is one of the most widely studied

families in pharmacogenomics as it is responsible for metabolic detoxification of over 75% of exogenous and endogenous compounds (Lynch & Neff 2007, Anzenbacher & Anzenbacherova 2001). The human genome encodes 57 functional *CYP* genes. *CYP* genes are very diverse and can be grouped into multiple families and different star alleles (haplotypes). Different *CYP* genes encode characteristically different CYP proteins each with a particular function. This study is focused on *CYP* star allele variation and its impact on protein stability and conformational behaviour which in turn influences drug response.

Advances in sequencing and genotyping technology have led to a wealth of CYP star allele data, some of which remains understudied. However, the availability of bioinformatics tools such as variant effect prediction tools and structural analysis tools provide a means to understand the impact of genetic variation on biological systems. Investigating the thermodynamic properties of proteins could inform the development of bioinformatics approaches to assess the impact of genetic variation on drug metabolism. This will in turn complement the ever-expanding scientific literature on precision medicine and contribute to establishing individualised treatments more rapidly. Furthermore, the computational modelling of biological systems has enabled researchers to investigate the potential impact of rare and/or complex haplotypes. The extent of the cumulative effect of CYP genetic variants (as represented by star alleles) can be seen as another level of variability that affects drug-metabolising characteristics and, as such, should be prioritised in pharmacogenomic investigations. Studying CYP star alleles may potentially reveal the relevance of select core variants or perhaps unmask any underlying mechanisms of interaction of variants (and their contributions to the overall impact on protein structural properties). Therefore, the exploration of star allele functional impact is another step towards understanding the precision medicine paradigm.

1.1 Problem statement

The ever-growing body of genomic data has created a backlog of uncharacterised variants and star alleles (Gaedigk et al. 2020). The precise mechanisms underlying how star alleles influence protein function and subsequently contribute to diverse drug responses via properties like stability and dynamics remain largely unexplored. Furthermore, studies focusing on African-specific genetic variations constitute a minority within the global

landscape of genetic research (Hamdi et al. 2021). Computational techniques have shown promise in assess a large amount of genetic data for their functional impact in rapid rates (Othman et al. 2022).

1.2 Research aim and objectives

This study aims to assess the potential impact of a combination of missense variants within CYP star alleles on protein stability and conformational dynamics. The thesis analyses CYP star alleles found in the PharmVar database to understand their effects on CYP protein structure and the functional relevance of core variants in pharmacogenomics. In addition, the thesis intends to provide a proof-of-concept concerning the utility of structural bioinformatics/multiscale modelling in interpreting CYP missense variants and star alleles.

The objectives of this study are:

1. To collect data regarding known CYP missense variants and star alleles catalogued in PharmVar.
2. To assess the effects of missense variants (single and known combinations) within CYP star alleles on protein stability and conformational dynamics via energy-based approaches.
3. To compare our *in silico* derived results with other commonly used variant interpretation tools and results from experimental procedures.
4. To perform multi-scale molecular dynamics simulation on selected CYP star alleles to further analyse their impact on protein structural properties.

1.3 Dissertation outline

This dissertation adheres to the 'publication' format endorsed by the Wits Faculty of Health Sciences. It consists of a literature review, a published article, and an overall discussion of the research conducted in relation to the aim of this study. The results

chapter (i.e., published article) is meant to be self-standing and, as such, there may be some intended repetition.

Chapter 2 – Literature review

This literature review chapter explains the star allele nomenclature and highlights its importance in the CYP pharmacogenomics community. Thereafter, the dissertation chapter explores the important structural properties of human CYP proteins and the consequences of missense mutations on CYP proteins. A brief contextual landscape of pharmacogenomics implementation and relevance is provided. This chapter concludes with an overview of the utility of structural bioinformatics and its potential in pharmacogenomics while also highlighting some of the tools used in this study.

Chapter 3 – Published article

Chapter 3 presents a published study detailing the findings of this MSc research. This includes findings on the use of structural bioinformatics approaches such as multi-scale modelling to study variant and star allele impact on protein energy and physical properties.

Chapter 4 – Discussion and conclusion

The final chapter highlights the major findings from the study, and emerging themes, provides additional contextual relevance to our study, explores the study's strengths and limitations, and provides an overall conclusion of the MSc study.

Chapter 2

Literature review

2.1 The star allele nomenclature

As sequencing technologies advance, the accumulation of sequence data creates a need to standardise the way genetic variation is recorded and stored (Robarge et al. 2007). *CYP* genes tend to harbour numerous haplotypes; therefore, the star allele nomenclature was introduced to represent SNP combinations/haplotypes, in a consistent way (Gaedigk et al. 2019, Twesigomwe et al. 2020) (Figure 2.1). Therefore, star alleles designate haplotypes with a collection of single or small nucleotide variations (SNVs). The establishment of a standardised star allele nomenclature provides a coherent way of communicating and sharing information pertaining to genomic variation amongst the scientific community (Gaedigk et al. 2018, Robarge et al. 2007, Kalman et al. 2016).

Star alleles haplotypes are denoted by a star number following the gene symbol. For instance, *CYP2D6*1* denotes the reference haplotype of the CYP2D6 gene – **1* serves as a reference to compare CYP2D6 polymorphic haplotypes. Non-reference haplotypes are given star numbers greater than one (subsequently) after submissions from the pharmacogenomics community have been curated and reviewed by PharmVar expert panels (Gaedigk et al. 2019, Robarge et al. 2007) (Figure 2.1). While **1* (reference) represents the 'haplotype' that produces a protein with normal function, **2*, **10*, **17* and so on produce proteins that may have altered functionality and possibly lead to variable drug response (in the context of CYP proteins) (Robarge et al. 2007). Non-reference star alleles have at least one core variant (variant characterised by sequence differences that alter protein function by producing an amino acid change, affecting expression levels, or

interfering with splicing) and at times supplementary variants (Botton et al. 2021). Core variations are critical for identifying star alleles.

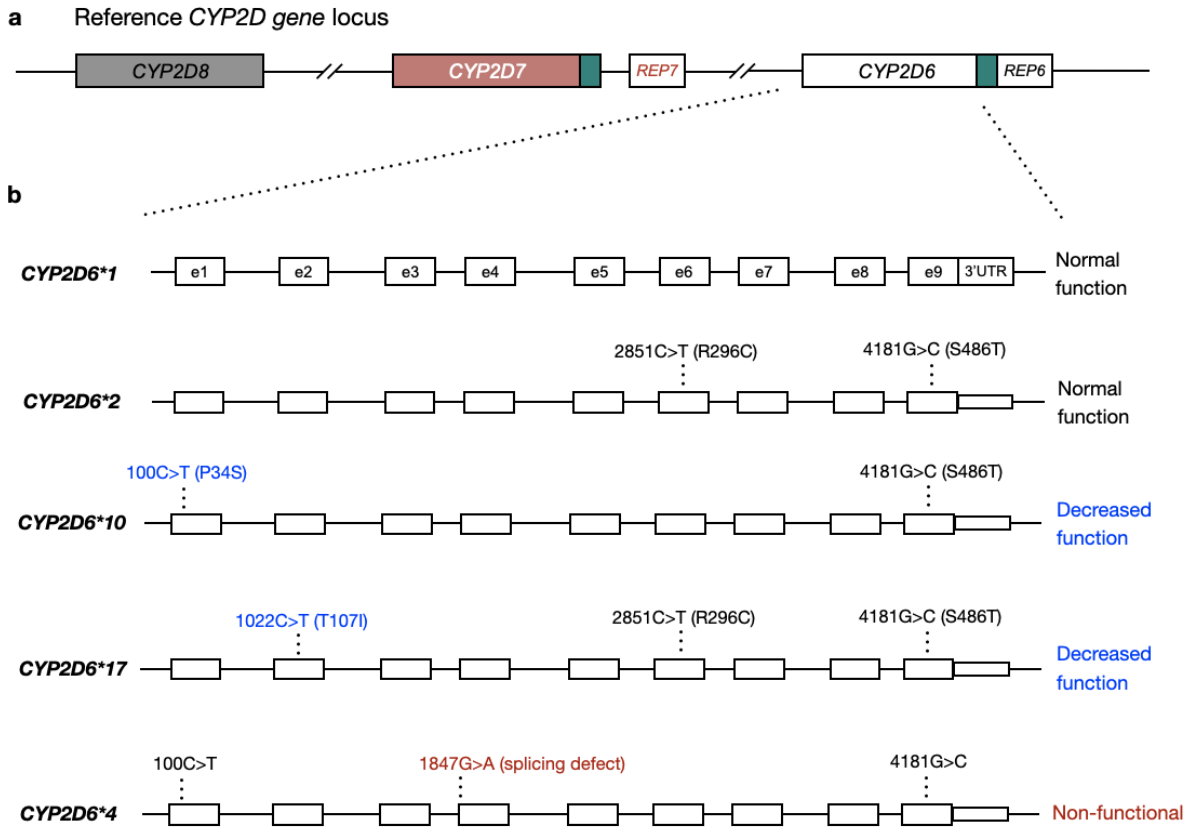


Figure 2.1: **Examples of *CYP2D6* star alleles.** This figure highlights the concept of using a star allele nomenclature to refer to various haplotypes. Consensus star allele functions according to the Clinical Pharmacogenetics Implementation Consortium are provided on the right.

The Pharmacogene Variation Consortium (PharmVar) (Gaedigk et al. 2018) is the primary source of star allele information. PharmVar collaborates transparently with groups of scientists and other consortia to provide standardised allele definition, curate and annotate pharmacogene variation, and disseminate data, all in an interactive and robust database. The PharmVar database (https://www.pharmvar.org) catalogues a wealth of CYP star allele data.

2.2 Structure of human Cytochrome P450 proteins

Although *CYP* gene and protein sequences are diverse, CYP proteins assume a conserved three-dimensional structure (common fold) (Otyepka et al. 2007). All mammalian CYPs

are primarily made up of 12 alpha-helices marked A to L beginning at the N-terminus (Figure 2.2), as well as a few beta-sheets (Poulos et al. 1987). The F- and G-helices, the F/G loop, the B/C loop, and the I-helix are structural areas of the CYP protein that play important roles in its function (Poulos et al. 1987). These areas are largely important in substrate catalysis and are closely related to the proteins' active sites – the F- and G- helices, the F/G loop, and the B/C loop promote substrate entry into the buried active site. Disruption of these mechanisms may cause altered protein functionality and potentially lead to variable drug responses. These disruptions may be due to genetic variants that manifest in a change in amino acid which results in a change in the conformation of the protein region and/or a change in the region's chemical properties (Šrejber et al. 2018). Ultimately, perturbations of the protein structure may affect the binding and recognition of protein substrates (drugs) and co-factor (Haem), and the protein's electron transport chain; all of which are crucial for the normal functioning of the protein (Šrejber et al. 2018).

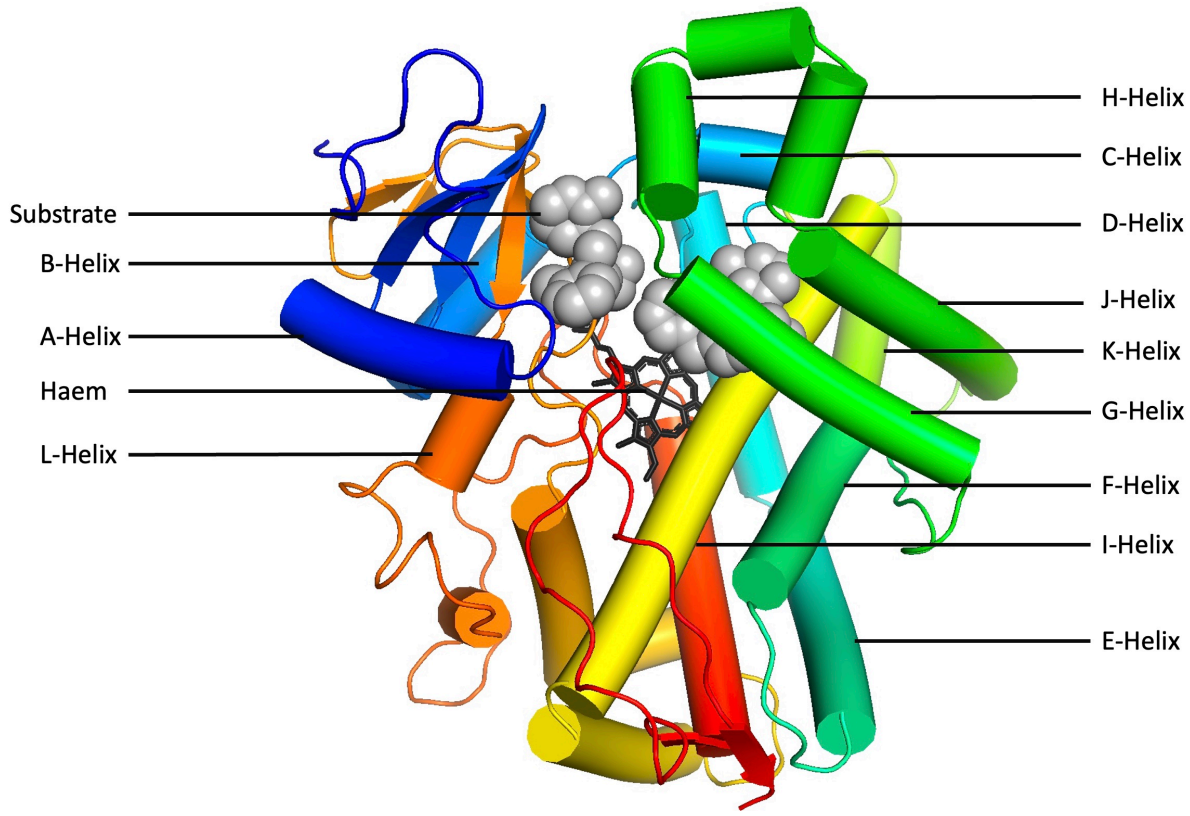


Figure 2.2: **Structure of a CYP protein adapted from Wang et al. (2015).** This figure visualises the major helices and the CYP protein family together with two bound substrates and the haem cofactor which enables the enzymatic oxidation of substrates via an electron transport chain.

2.3 Clinical application

In humans, CYP proteins are primarily found lodged in the membrane of the endoplasmic reticulum of kidney and liver cells, among others, and are in part responsible for metabolic pathways. Here, they are involved in the oxidisation, detoxification, and clearance of various compounds (Fanni et al. 2021, Anzenbacher & Anzenbacherova 2001). Typically, CYP proteins participate in the monooxygenase reaction, utilising a haem cofactor housed deep in the active site cavity (Šrejber et al. 2018). Different CYP proteins are known to contribute differently to the overall 75-90% of drug metabolism-related events reported for CYP proteins (Lynch & Neff 2007, Anzenbacher & Anzenbacherova 2001). For instance, CYP2D6, CYP2B6, and CYP2C9 are reported to metabolise approximately 25%, 10-12%, and 15% of clinical drugs, respectively (Ingelman-Sundberg et al. 2007, Zanger & Klein 2013, Isvoran et al. 2017).

Although not fully realised, knowledge pertaining to pharmacogenomics is currently being used in clinical practice, limited to a few clinical settings globally due to resource limitations and insufficient evidence/evaluation on cost-effectiveness. Some drugs on the market contain pharmacogenetic guidelines – some pertaining to CYP proteins – on their labels (Theken et al. 2020, Bousman et al. 2023). Guidelines for pharmacogenomic implementation, drug prescription, and dosing have been made available by groups and consortia such as the Dutch Pharmacogenetics Working Group (DPWG) (Lunenburg et al. 2020), Clinical Pharmacogenetics Implementation Consortium (CPIC) (Relling & Klein 2011), EU-PIC (European Pharmacogenetics Implementation Consortium) (https://www.eu-pic.net, last accessed: 15-05-2022), SEAPharm (Southeast Asian Pharmacogenomics Research Network) (Chumnumwat et al. 2019), and others. However, under-represented populations such as African populations may have distinct pharmacogenetic variation patterns that can lead to distinct adverse drug reactions (Mhalu 2005, Mulder et al. 2018).

There is limited knowledge regarding the mechanistic impact of star alleles (defined by combinations of core variants) beyond their overall functional designation. Some core variants are shared with other well-characterised star alleles, and some have been studied in isolation. Borba et al. (2016) demonstrated the synergistic impacts of *CYP2D6*14* and *CYP2D6*7* missense variants on protein function through *in silico* analysis of protein

energy properties, aligning these findings with existing clinical knowledge. It is important to build on such studies and conduct more comprehensive *in silico* functional analyses for CYP star alleles as it may draw attention to the functional relevance of core variants.

2.4 Pharmacogenomics in the context of African populations

The realisation of the large genetic diversity observed in the African population, as well as the difference in genetic structure with regard to haplotypes and shorter haploblocks (Campbell & Tishkoff 2008, Mulder et al. 2021) has led to an increased appreciation of African genomics. Therefore, the genetic diversity of the African population presents a challenge in the implementation and replication of studies conducted in other populations for benefits in pharmacogenetics (Wright et al. 2018). Despite African populations having distinct adverse drug reaction (ADR) patterns relative to Western and Asian countries (Mhalu 2005, Mpye et al. 2017), it remains, however, that Africa is lagging in genomic research. Hamdi et al. (2021) state that African studies comprise approximately 4% of the commonly used PharmGKB database, meant to cover the entire diaspora.

The underrepresentation of African populations in pharmacogenetic databases (Cavallari et al. 2017, Da Rocha et al. 2021) is due to (among many other factors) inefficient pharmacovigilance (Ampadu et al. 2016) and underrepresentation in global Genome-Wide Association Studies (GWAS) (Buniello et al. 2019) – partly reflecting the scarcity of resources crucial in genomic research. This scarcity of resources in the African continent may be attributed to generally lower investment into research by governments in the middle to low-income countries (comprising a majority of African countries) (WHO 2021). Moreover, medical scientist skills remain scarce in Africa as reported by the South African government Department of Higher Education and Training (2014). Although many African-specific high-frequency variants (minor allele frequency > 0.05) have been reported, there remain variants with varying inter-ethnic frequencies that are yet to be uncovered (B Tata et al. 2020, Maranville & Cox 2016). Therefore, the pharmacogenetic profiles of other populations may not necessarily be the same as those of African populations. Furthermore, of the currently known genetic variation in Africa, only a few can be analysed for benefits in healthcare also due to resource limitations (WHO

2021). GWAS, for instance, are better suited for evaluating common variation within populations. This is in contrast to rare genetic variants which may present in underrepresented populations, thus not capturing broader genetic landscapes. Remarkably, there has been a surge of high-throughput genome analysis methods that may be employed in resource-limited settings. Structural bioinformatics, for instance, may be coupled with whole-exome or whole-genome sequencing data (which enable the identification of rare genetic variants) to assess the impact of genetic variants in a high-throughput manner.

2.5 Structural bioinformatics techniques

In vitro and *in vivo* studies for assessing CYP star allele impact on protein activity are often expensive, labour-intensive, and limited to common star alleles. Conversely, *in silico* tools are inexpensive and more accessible given the recent uptake of open-source practices. With regards to assessing variant effects on protein structure, structural bioinformatics techniques such as protein modelling, docking studies and molecular dynamics are well-established (de Almeida Paiva et al. 2022, Holm 2020). However, their application for assessing the impact of pharmacogenetic variants has not been extensively validated. Structural bioinformatics leverages physics, chemistry, and mathematics to address a variety of biological problems at rates that are often quicker than would be possible experimentally (Peter et al. 2019). This has made it possible to investigate biological systems (in atomic detail) that would otherwise be practically impossible to research using conventional wet laboratory techniques, while also reducing the time and financial cost.

Structural bioinformatics techniques play a crucial role in complementing empirical wet laboratory studies within the field of pharmacogenetics. In scenarios where empirical experimental data is limited, protein bioinformatics becomes invaluable by offering essential insights into the pharmacogenetic profiles of individuals through the modelling of 3D protein structures and their variations (Burley et al. 2023, Cheng et al. 2023). Structural bioinformatics is employed to discern variants that have the potential to influence drug response, allowing for their strategic prioritisation in subsequent experimental validation (Cheng et al. 2023, Livesey & Marsh 2022). This process unravels the nuanced changes in the interaction between the protein and the drug, arising from missense variations,

thereby refining our understanding of variable drug efficacy and safety (Cheng et al. 2023).

This collaborative effort not only aids researchers in gaining profound insights into individualised responses to medications but also guides the development of tailored treatment strategies. An intricate interplay exists between empirical protein experiments and structural bioinformatics, where experimental techniques refine the 3D structures of proteins predicted using structural bioinformatics approaches (Burley et al. 2023). Concurrently, structural bioinformatics contributes by providing additional annotations for protein structure and function (Cheng et al. 2023, Livesey & Marsh 2022). This dynamic interaction enhances our capacity to unravel the intricacies of pharmacogenetic mechanisms, facilitating the seamless integration of computational predictions with empirically validated insights. The subsequent subsections delineate specific applications and tools relevant to structural bioinformatics, pertinent to the focus of this study.

Additionally, structural bioinformatics offers a rapid and efficient means to analyse many protein variants (Othman et al. 2022). Empirical techniques may be limited in their capacity to analyse numerous variants due to resource constraints and time-intensive protocols. Structural bioinformatics enables the prioritisation of the most promising variants for downstream empirical studies, thereby maximising the efficiency and effectiveness of experimental efforts in elucidating protein structure-function relationships.

2.5.1 Protein modelling

The structure of a protein is intrinsically related to its function; the structure plays an important role in the types of interactions that take place within and between proteins (Ahmad et al. 2021, Sinai & Kelsic 2020). Experimentally-solved structures contained within Protein Data Bank (PDB) (Burley et al. 2021) cover only 17% of the human proteome (Somody et al. 2017, Porta-Pardo et al. 2022) and this is merely not enough to cover all functionally crucial regions of the proteome. Consequently, the sequence-structure gap is presented which has been a major challenge in the fields of protein bioinformatics and biochemistry, and the subject for many undertakings that seek to predict the structure of proteins based on their sequence and computational molecular physics (physical interactions) and/or homology with other experimentally solved structures (Ittisoponpisan et al. 2019). Through these undertakings, protein modelling has seen rapid advances in

the last decade, even surpassing Moore's Law (Brini et al. 2020).

Briefly, the protein folding problem may be described as the limited knowledge about how sequences of amino acids inform the structures of three-dimensional folded proteins. There are many routes an unfolded protein can take to enter the folding funnel Figure 2.3, dependent on the environment. All folding routes lead to decreased conformational entropy and total energy, facilitated by the hydrophobic effect which increases the number of intramolecular contacts as the protein moves down the folding funnel (Adamcik & Mezzenga 2018, Sinner et al. 2014, Dill & MacCallum 2012). Proteins can get stuck in non-native intermediate states as they move down the funnel – these are states of relatively low conformational entropy and total energy that may be representative of misfolded proteins. These 'misfolded' proteins would have to overcome the transition state (energy barrier) to rejoin the folding funnel and fold to the lowest energy state and conformational energy – the native state (Adamcik & Mezzenga 2018, Dill & MacCallum 2012). The protein folding problem lies in understanding the interactions of distant amino acids to bring about the native state under defined physiological conditions (Dill & MacCallum 2012).

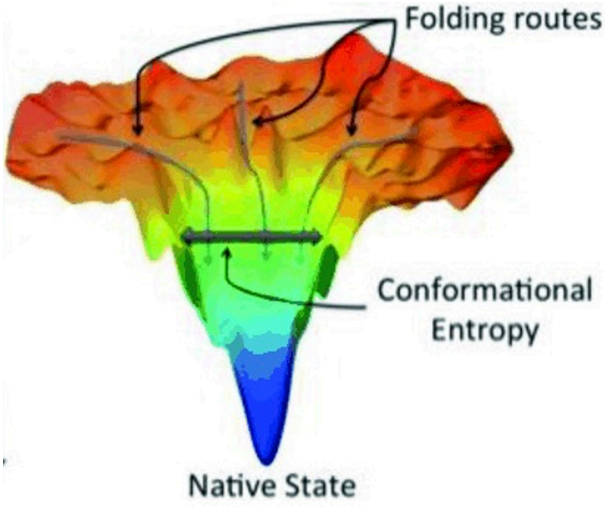


Figure 2.3: **Illustration of a protein folding energy landscape adapted from Sinner et al. (2014).** The top (red) of the folding funnel represents isoenergetic unfolded proteins with high conformational entropy (represented by the width of the folding funnel) and high total energy. The protein has to move down the funnel to reach an energetically minimum state (blue) with the least conformational entropy while avoiding intermediate states (green) that may present an energy barrier and undesirably folded proteins.

Protein modelling methods that rely on physical interactions employ thermodynamic or kinetic simulation – they strive for chemical and thermodynamic correctness. However, they are often computationally demanding and they are not always feasible for large-scale simulations (Jumper et al. 2021, Brini et al. 2020). On the other hand, protein modelling methods that rely on homology models that were solved experimentally are less computationally demanding and, as a result, are apt to be the most accessible and widely used method for protein modelling (Jisna & Jayaraj 2021). Predicting protein structures with experimental accuracy using the methods described is often challenging, especially when there is no close homologue of the target protein (Jumper et al. 2021), so despite rapid advances, there appeared to be more work to be done in the field of protein modelling. More recently, the AI-based AlphaFold (Jumper et al. 2021) tool was developed to address the gap between protein sequence and three-dimensional protein structure and has so far shown one of the most promising results, with astonishing accuracy (Kryshtafovych et al. 2021).

Regarding CYP proteins, some lack experimentally determined 3D structures. Traditional methods used before AI relied on evolutionary approaches to predict these structures. However, AI-based tools now offer enhanced accuracy, boosting confidence in downstream protein analysis outcomes, including the prediction of how protein mutations affect function(Li et al. 2023, Lin & Ngiam 2023).

2.5.2 Variant effect prediction

Variant effect prediction tools have been developed to address the issue of the enormous number of known genetic variations with no functional classification. They seek to prioritise genetic variation that is likely to cause disease and/or aberrant biological function based on features such as sequence phylogenetic conservation, amino acid physicochemical properties, protein structural properties, and others including machine learning tools trained using variants of known clinical consequence (Horne & Shukla 2022, Mahmood et al. 2017). Many prediction tools rely solely on the sequence of a protein to predict the variant effect; however, this may neglect changes in the conformation of the protein (Gyulkhandanyan et al. 2020).

More recently, tools that integrate both sequence-based and structure-based approaches have been developed. The Structural Workflow for Annotating ADME Targets (SWAAT)

(Othman et al. 2022) is an example of a variant effect prediction tool that integrates sequence-based and structure-based approaches. SWAAT seeks to address the lack of structure-based variant effect prediction tools for ADME genes – essentially important for protein conformational analysis – by assessing molecular features relating to protein energetics and large-scale conformational movements. Moreover, SWAAT integrates a predictive machine learning random forest binary model with an accuracy surpassing that commonly used sequence-based variant prediction tools (i.e. MetaSVM PROVEAN, SIFT, PolyPhen-2, FATHMM, and CADD).

Other variant prediction methods utilise deep neural networks to study the effects of variants resulting in missense protein mutations on the conformational dynamics of proteins. These tools may look at the flexibility and stability of the protein to make comments about the expected function (e.g. ENCoM).

2.5.3 Normal Mode Analysis

Normal Mode Analysis (NMA) is a structural bioinformatics tool used to investigate protein conformational behaviour (Frappier & Najmanovich 2014). Based on the concept that a protein harmonically oscillates around an energetically lowest shape, NMA approaches generate a fundamental set of potential protein movements (Cirauqui Diaz et al. 2021, Frappier & Najmanovich 2014). These are the motions that the protein can undergo without severely disrupting its overall structure. The protein's dynamics are defined here as eigenvectors (normal modes) and eigenvalues (frequency of motions), from which the movements can be sampled at various points in conformational space.

Elastic Network Contact Model (ENCoM) is a coarse-grained Normal Mode Analysis NMA method for studying protein conformational behaviour (Frappier & Najmanovich 2014). Therefore, ENCoM is able to calculate variations in the entropy of a protein as a result of variations within the protein (Frappier et al. 2015). Each mode describes the motions of an individual amino acid in a system in the absence of external forces. The most accessible modes are generally characterised as having the lowest frequency (Frappier & Najmanovich 2014, Frappier et al. 2015) and represent global movements of the protein structure. High-frequency motions represent local movements of subset regions of the protein structure.

Coarse-graining in NMA entails the reduction of atomic representation by the group-

ing of atoms into spheres (Figure 2.4), easing the computational burden and enabling faster and simpler system simulation (Kmiecik et al. 2016). This type of modelling does not inform about the specific nature of amino acids and the movement of their side chains and intramolecular interactions, but it can inform researchers about the overall topological movement of the protein which is more often intrinsic to the functioning of the protein. However, ENCoM differs from other coarse-grained NMA approaches in that it incorporates a pairwise atom-type term by defining the nature of the amino acids (Frappier & Najmanovich 2014). The ability to account for the nature of amino acids implies that ENCoM may be used to study the effect of specific missense variants and accurately predict the effect that missense variants would have on the conformational behaviour of proteins.

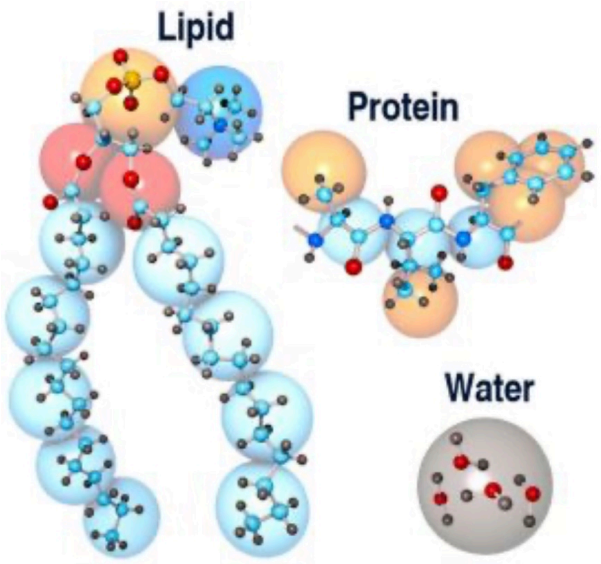


Figure 2.4: **Illustration of coarse-graining adapted from Kmiecik et al. (2016).** The ball-and-stick model represents all-atom representation and the larger sphere encompassing the ball-and-stick models represents coarse-graining – it condenses the molecule by grouping atoms for relatively simpler and quicker simulation of biological molecules.

FoldX is another structural bioinformatics tool that is used to predict the changes in free energy (ΔG) induced by protein mutations using an empirical force field derived from protein experiments (Schymkowitz et al. 2005). The integration of FoldX and ENCoM yields a synergistic enhancement in the predictive capacity for variant effects, capitalising on the strengths of both methods (Frappier et al. 2015). A force field is a set of equations that describe the energetic properties of a protein in terms of its atomic coordination.

This method analyses the energy contributions from various forms of interactions between atoms in the protein using a comprehensive atomistic description of the protein structure (Schymkowitz et al. 2005). FoldX is capable of predicting how mutations would affect protein stability, folding dynamics, and protein-protein interactions. Equation 2.1 expresses the energy terms defined by FoldX.

$$\begin{aligned}\Delta G = a \cdot \Delta G_{vdw} + b \cdot \Delta G_{solvH} + c \cdot \Delta G_{solvP} + d \cdot \Delta G_{wb} + e \cdot \Delta G_{hbond} \\ + f \cdot \Delta G_{el} + g \cdot \Delta G_{kon} + h \cdot T\Delta S_{mc} + k \cdot T\Delta S_{sc} + l \cdot \Delta G_{clash}\end{aligned} \tag{2.1}$$

Where $a - l$ represents the weights of their coupled energy terms.

2.5.4 Molecular dynamics simulation

Molecular dynamics (MD) simulation is a powerful method used to explore the conformational energy landscape of proteins. MD simulation models biological systems as atoms *in-silico* in a dynamic environment while assuming a given potential energy function (Cheatham III & Case 2013, Case et al. 2005). This method relies on classical mechanics principles, which numerically integrate positions and velocities of particles that make up the protein systems over time (Cheatham III & Case 2013). The dynamics of atoms and molecules are measured via MD simulation, and the given potential energy function is utilised to determine the force experienced by a given atom given the position of the other atoms. Molecular dynamics simulations can reveal details about the mechanics of protein function and the effects of mutations on these mechanisms. MD simulation measures the dynamics of atoms and molecules and the given potential energy function is used to calculate the force experienced by a given atom, given the position of the other atoms. Molecular dynamics simulations can provide insights into the mechanics of protein function and the impact of mutations on these mechanisms.

The Assisted Model Building with Energy Refinement (AMBER) (Case et al. 2005) program is one of the few tools that are used to perform MD simulation. AMBER is unique to other MD simulation programs because it's a complete package that does not require external software to build non-standard residues onto the simulation system. Additionally, the AMBER package has inbuilt features for performing a myriad of calculations of the system including Molecular Mechanics Poisson-Boltzmann Surface Area (MMPBSA), quantum mechanics/molecular mechanics (QM/MM), and steered MD

(Case et al. 2005).

Computational approaches are proving to be useful for characterising disease-causing variants (Gerasimavicius et al. 2020, 2023); however, their utility for characterising pharmacogenetic variants is currently unclear. This points to the gap(s) in the knowledge that your study is going to address with regard to using structural bioinformatics/multiscale modelling in pharmacogenomics and providing predictions of the effects of currently uncharacterised missense variants and their combinations in CYP star alleles.

2.5.5 C-alpha and CYP protein structural landscape

Protein dynamics manifest through the collective vibrations of various atoms, each contributing distinctively to the overall motion. The vibrational patterns exhibited by side chains play a significant role in shaping localised movements, impacting the precise configuration of active sites and molecular recognition regions (Harmalkar & Gray 2021). They undergo relatively higher fluctuations characteristic of high frequency modes Figure 2.5. The inherent diversity in the chemical composition of side chains introduces a range of vibrational modes, thereby influencing nearby conformational changes (Harmalkar & Gray 2021, McCammon 1984). Hydrogen atoms, due to their lighter masses, partake in higher frequency modes, enabling nuanced adjustments in bond angles and distances that subtly fine-tune the stability of secondary structures such as alpha-helices and beta-sheets (Engler et al. 2003). Figure 2.5 illustrates the variable fluctuations observed among the constituent atoms of proteins. These fluctuations manifest a discernible pattern, with their magnitude progressively escalating from the central region of the protein towards the terminal groups.

While acknowledging the significance of side chain movements in understanding local interactions and fine-tuning, it is imperative to underscore that focusing on the backbone offers a global perspective. This global perspective is indispensable for comprehending the overarching structural dynamics of proteins, providing valuable insights into their functional behaviour. C-alpha ($C\alpha$) atoms are crucial for understanding the overall structure and conformational dynamics of proteins. The $C\alpha$ atom links the amino group ($-NH_2$), the carboxyl group ($-COOH$), and a hydrogen atom to form the backbone chain of proteins. This dynamic linkage provides structural integrity and dictates the overall folding and stability of the protein. The $C\alpha$ atoms, positioned in the backbone,

act as molecular scaffolds, determining the three-dimensional arrangement of secondary structures such as alpha-helices and beta-sheets (Drake & Pettitt 2020). Consequently, alterations in the $C\alpha$ positions can ripple through the entire protein, influencing its stability and function (Drake & Pettitt 2020).

The $C\alpha$ exhibit low-frequency normal modes associated with global motions of the protein (Bahar et al. 2010, Catucci et al. 2022, Kaledin et al. 2005). These modes dictate the overarching structural dynamics, influencing the protein's ability to transition between different functional states (Mahajan & Sanejouand 2015, McCammon 1984). Low-frequency modes associated with $C\alpha$ vibrations capture large-scale movements, such as domain rotations or hinge-like motions (McCammon 1984) which may be implicated in substrate binding (Catucci et al. 2022).

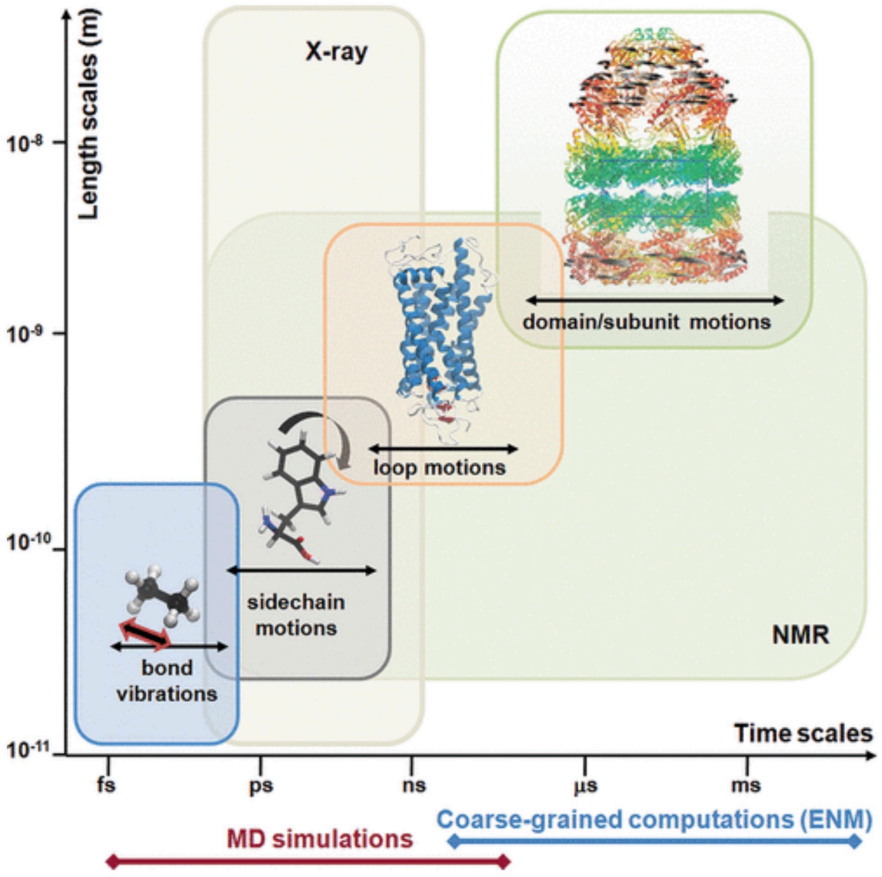


Figure 2.5: **Broad range of equilibrium motions accessible under native state conditions, adapted from Bahar et al. (2010).** Cooperative motions occur at the low frequency end of the mode spectrum, where fluctuations are lower. These modes engage large substructures (domain/subunit motions).

The knowledge and techniques detailed in this literature review have strategically informed the selection of methods used to address the objectives of this study. The chapters that follow delve into the use of structural bioinformatics techniques to address gaps in knowledge concerning the functional effects of CYP star alleles.

Chapter 3

Characterising the combined effects of cytochrome P450 missense variation within star allele definitions

The work presented in this chapter is published in the journal *Pharmacogenomics*. This chapter is self-standing and follows the recommendations given by the Wits Faculty of Health Sciences for a dissertation submitted in 'publication' format.

Khoza, N., Twesigomwe, D., & Othman, H. (2023). Characterizing the combined effects of cytochrome P450 missense variation within star allele definitions. *Pharmacogenomics*, 24(10), 561-578. doi: 10.2217/pgs-2023-0068. PMID: 37503750.

3.1 Summary

Background: Cytochrome P450 (CYP) genetic variation largely impacts drug response. However, many CYP star alleles (haplotypes) lack functional annotation, impeding our understanding of drug metabolism mechanisms. We aimed to investigate the impact of missense variant combinations on CYP protein structures. **Methods:** Normal mode analysis (NMA) was conducted on 261 missense variants within 91 CYP haplotypes. *CYP2D6*2* and *CYP2D6*17* were prioritised for molecular dynamics (MD) simulation. **Results:** NMA and MD highlight the effects of known CYP missense variants on protein stability and conformational dynamics. Missense variants within haplotypes may have intermodulating effects on protein structure and function. **Conclusion:** This study highlights the utility of multi-scale modelling in interpreting CYP missense variants and

particularly their combinations within various star alleles.

3.2 Introduction

Inter-individual drug response is partly influenced by genetic variation in genes that encode proteins involved in the Absorption, Distribution, Metabolism, and Excretion (ADME) of drugs. Of these ADME proteins, the cytochrome P450 (CYP) superfamily is known to be important in drug processing (reported to metabolise over 70% of drugs) (Lynch & Neff 2007, Fanni et al. 2021). CYP genetic variation is therefore crucial in the pharmacogenomics of a wide range of medications (Gaedigk et al. 2020, Zhou & Lauschke 2022). CYP proteins are expressed in the liver and kidney, among other organs, and are generally responsible for phase 1 drug metabolism (Fanni et al. 2021, Anzenbacher & Anzenbacherova 2001).

Human CYP proteins consist of 12 alpha-helices, labelled A to L, starting from the N-terminal and a few beta-sheets (Poulos et al. 1987). Structural regions that play crucial roles in the functioning of CYP proteins are the F- and G-helices, the F/G loop, the B/C loop, as well as the I-helix. These regions are primarily involved in the catalysis of substrates and are closely associated with the active site of the CYP proteins and their haem cofactors. The F- and G- helices, the F/G loop and the B/C loop facilitate access of the substrate into the buried active site (Poulos et al. 1987, Otyepka et al. 2007, Wade et al. 2004). CYP genetic variation tends to occur as haplotypes commonly referred to as star ('*') alleles. These star alleles may be composed of collections of single or small nucleotide variations and/or structural variants. Typically, non-reference star alleles consist of at least one core variant – the variant(s) responsible for the haplotype function and designation – and supplementary variants. A core variant is defined by sequence variations that affect the function of the protein by causing an amino acid change, changing the expression levels, or interfering with splicing (Botton et al. 2021). Different groups and consortia have sought to characterise pharmacogene genetic variation to inform prescription guidelines (Fanni et al. 2021, Lunenburg et al. 2020, Relling & Klein 2011, Chumnumwat et al. 2019). For example, the Pharmacogene Variation (PharmVar) Consortium database (Gaedigk et al. 2020) (https://www.pharmvar.org) curates and catalogues many of the known star alleles (Gaedigk et al. 2018). The Clinical Pharmaco-

genetics Implementation Consortium (CPIC) (Relling & Klein 2011) collaborates with PharmVar to provide functional designations (increased, normal, decreased, no function, and uncertain function) of resolved haplotypes within the PharmVar database.

Many of the star alleles in the PharmVar database are yet to be assigned a CPIC clinical function classification. The process of assigning functional classification is complex and challenging. First, relatively common star alleles are more likely to have clinical and/or experimental functional evidence – this means that most rare and sometimes more deleterious variants are not functionally characterised. Moreover, testing the unequivocal effect of genetic variant(s) on drug metabolism may require expensive and time-consuming laboratory experiments. However, advances in protein computational techniques such as structural bioinformatics have made it possible to study biological systems that would otherwise be challenging in a wet laboratory, while bringing down the time and the associated financial cost. Structural bioinformatics is a branch of science that studies the structure of biological macromolecules in silico, with applications in disease mechanisms. Structural bioinformatics relies on the implementation of physics, chemistry, and mathematics to solve a range of biological problems at rates often quicker than conventional experimental procedures (Peter et al. 2019). It often explores protein structure, stability, and conformational dynamics and the changes that may arise due to mutations.

The protein structure-function relationship explains how the three-dimensional conformation of a protein is intrinsically tied to protein function as specific conformations may allow for specific functions (Ahmad et al. 2021, Sinai & Kelsic 2020). Protein folding is driven by intramolecular forces that form between specific amino acids to ensure that the protein achieves a stable conformation under physiological conditions and maintains specialised spatial arrangements that facilitate intermolecular bonding with the protein's co-factors and specialised function. Protein structure mutations that change the geometry and spatial arrangement of amino acids in a protein, especially those localised within functionally crucial regions of the protein such as the active site, may affect protein function by either impaired the binding of ligands and co-factors or altered protein interaction which results in impaired protein mechanistic function (Ahmad et al. 2021, Sinai & Kelsic 2020, Adamcik & Mezzenga 2018, Dill & MacCallum 2012). Understanding the relationship between protein structure and function holds the potential to unravel mechanisms that add to our knowledge of biology and may potentially elucidate drug

discovery, personalised medicine and diagnostics paradigms.

While structural bioinformatics alone may not be enough to elucidate the full mechanisms at play regarding CYP haplotypes' functional impact, it provides a layer of evidence that can complement other approaches such as Deep Mutational Scanning (Gerasimavicius et al. 2023). This can be useful in determining the functional impact of rare missense-variant-defined star alleles catalogued in PharmVar, and novel star alleles in under-represented global populations e.g. African populations, which are known to be highly diverse genetically. To this end, this study aimed to leverage computational tools to assess the potential impact of missense variant combinations within CYP star alleles on protein stability and conformational dynamics.

3.3 Materials and Methods

3.3.1 Data

At the time of this study, 17 cytochrome P450 genes had been transitioned to the PharmVar pharmacogenomics database and also undergone star allele and nomenclature curations by the respective PharmVar gene expert panels (https://www.pharmvar.org). We focused our analysis on 9/17 of these CYP genes, namely *CYP2D6*, *CYP2B6*, *CYP2C19*, *CYP2C9*, *CYP2C8*, *CYP2F1*, *CYP2A13*, *CYP2W1*, and *CYP3A4*. This was primarily because these 9 genes have multiple reported missense haplotype combinations and provided a relatively comprehensive dataset to assess the utility of the structural bioinformatics approaches described herein (Figure 3.1). A total of 261 CYP missense variants were analysed in the study (Figure 3.2 and Table S1). Variant Call Format (VCF) files for each of these CYP genes were acquired from PharmVar v4.2.7.2 and then filtered for missense variation using the Ensembl Variant Effect Predictor(VEP; https://www.ensembl.org/ToolsVEP) (McLaren et al. 2016).

3.3.2 Normal mode analysis

Normal mode analysis (NMA) methods are a type of structural bioinformatics approach that seeks to describe the pattern of motion of protein systems (Frappier & Najmanovich 2014). They produce a basic set of possible movements based on the assumption that a

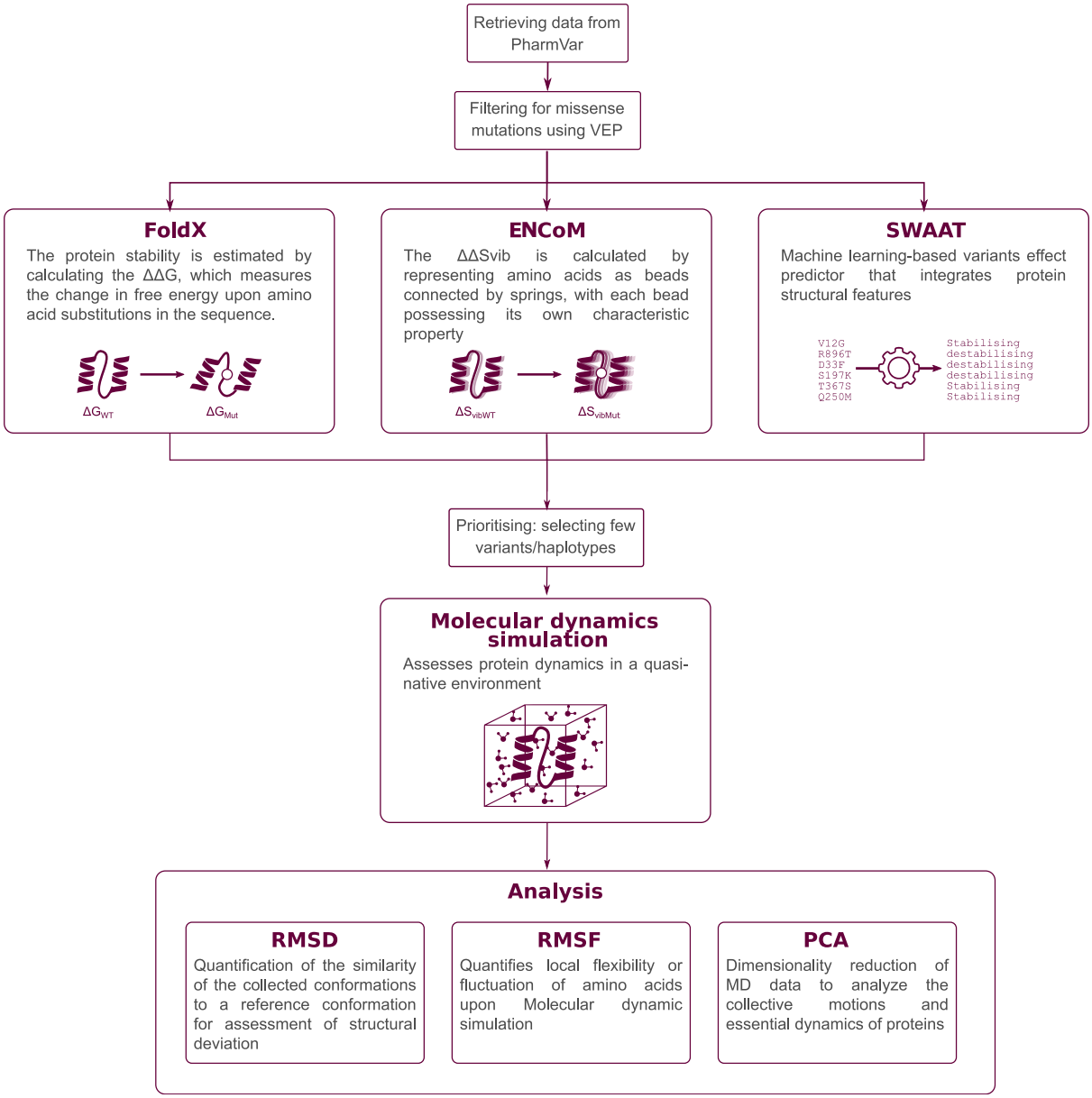


Figure 3.1: **An overview of the employed methodologies and their interrelationships.**

protein harmonically oscillates around an energetically minimum conformation (Cirauqui Diaz et al. 2021, Frappier & Najmanovich 2014). The dynamics of the protein are described as eigenvectors (normal modes) and eigenvalues (frequency of motions), and the movements can be sampled at various points in conformational space (Frappier & Najmanovich 2014).

We used the Elastic Network Contact Model (ENCoM) (Frappier & Najmanovich 2014) to study the effects of selected missense variants and to predict variations in vibrational entropy ($\Delta\Delta S_{\text{vib}}$). ENCoM is a coarse-grained NMA approach used to study variations in the entropy of a protein as a result of variations within the protein (Frappier et al. 2015). Coarse-graining in NMA entails the reduction of atomic representation

by the grouping of atoms, easing the computational burden for relatively quicker and simplified simulation of protein systems (Kmiecik et al. 2016). ENCoM is set apart from other coarse-grained NMA methods because it factors in a pairwise atom-type term by describing the nature of the amino acids (Frappier & Najmanovich 2014).

3.3.3 Calculating Gibbs free energy of folding

We used FoldX (Schymkowitz et al. 2005) to predict the effects of missense variants on protein conformational dynamics — in particular, estimating protein energetics in the form of variation in Gibbs free energy of folding ($\Delta\Delta G$). FoldX is a force field derived from protein experiments. A force field refers to a collection of equations used to describe the energetic properties of a protein in reference to the protein's atomic coordination.

$\Delta\Delta S_{\text{vib}}$ may be used to approximate $\Delta\Delta G$; therefore, we computed the combined

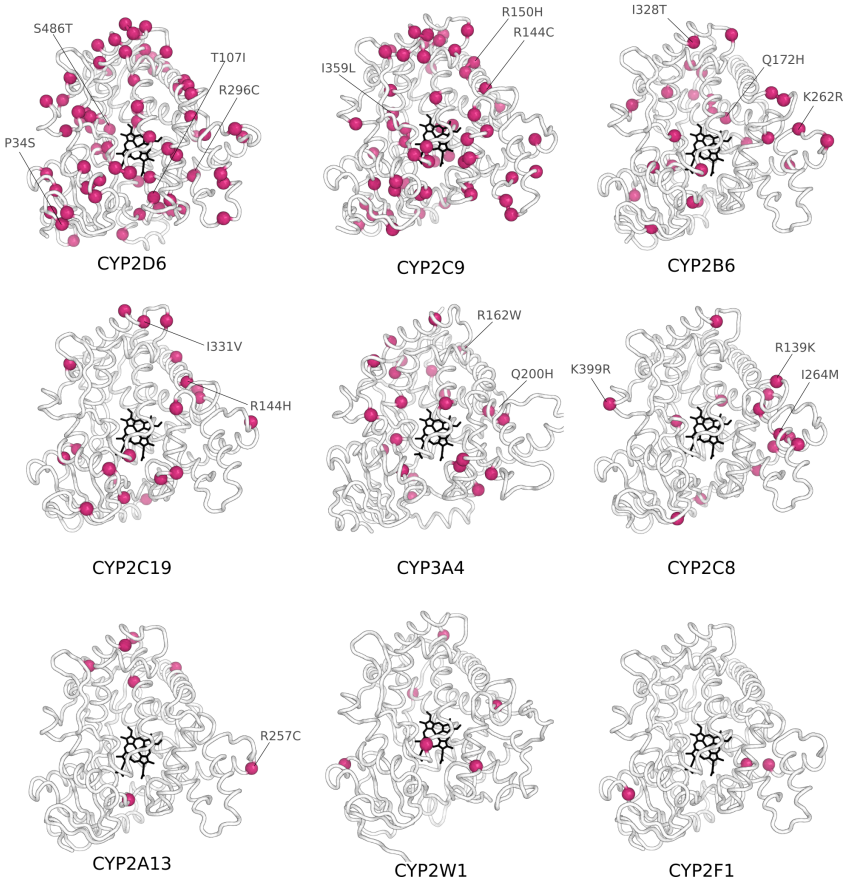


Figure 3.2: **Mapping of single amino acid variants on CYP P450 proteins.** The localisation of the catalytic site is indicated by the haem atoms (black). We have also labelled the positions of some clinically relevant amino acids (magenta) in selected protein structures.

$\Delta\Delta G$ (combination of ENCoM and FoldX) for each variant and star allele analysed in this study. The combination of FoldX and ENCoM was shown to be a better predictor of the effects of protein variation, based on folding energy variation when factoring in parameters to minimise the root mean square error (RMSE) between experimental and predicted $\Delta\Delta G$ values (Frappier et al. 2015). The combined $\Delta\Delta G$ is described by equation 3.1:

$$\Delta\Delta G_{\text{combined}} = \alpha\Delta\Delta G_{\text{ENCoM}} + \beta\Delta\Delta G_{\text{FoldX}} \tag{3.1}$$

where α and β are parameters to minimise the RMSE corresponding to -1.12 and 0.38, respectively.

A comparison of the thermodynamic energies exhibited by the wildtype and the mutant was carried out as follows (equations 3.2 and 3.3):

$$\Delta\Delta S_{\text{vib}} = \Delta S_{\text{vib (mutant)}} - \Delta S_{\text{vib (wildtype)}} \tag{3.2}$$

$$\Delta\Delta G = \Delta G_{\text{(mutant)}} - \Delta G_{\text{(wildtype)}} \tag{3.3}$$

A $\Delta\Delta S_{vib} > 0.5\ kcal \cdot mol^{-1} \cdot K^{-1}$ signifies increased conformational flexibility of the mutant compared to the wild-type protein, $\Delta\Delta S_{\text{vib}} < -0.5\ kcal \cdot mol^{-1} \cdot K^{-1}$ signifies that the wild-type has greater conformational flexibility than the mutant protein. A $\Delta\Delta G > 0.5\ kcal \cdot mol^{-1}$ signifies that the mutant protein is destabilizing compared to the wild-type, while a $\Delta\Delta G < -0.5\ kcal \cdot mol^{-1}$ signifies that the mutant protein is stabilizing. $\Delta\Delta S_{\text{vib}}$ and $\Delta\Delta G$ closest to zero indicate the minimal effect of the protein mutation on protein stability (Frappier & Najmanovich 2014).

As part of NMA, we calculated Root Mean Square Inner Product (RMSIP) values for haplotypes based on eigenvectors and eigenvalues. RMSIP was used to measure the similarity between two conformational subspaces, described by modes/motions, obtained through PCA (Martín-García et al. 2015, David & Jacobs 2011). If the vectors in subspace one correlate to the vectors in subspace two, the protein is expected to have similar motion and, by extension, the function (David & Jacobs 2011).

3.3.4 Comparison of NMA with Deep Mutational Scanning and other approaches

Deep Mutational Scanning (DMS) (Fowler & Fields 2014) has recently emerged as an informative experimental technique for investigating the functional impact of missense vari-

ants at a large scale. At present, among the CYP proteins, only CYP2C9 and CYP2C19 have published DMS data (Amorosi et al. 2021, Zhang et al. 2020). We, therefore, compared our NMA predictions for variants in these proteins with the results from these studies by assessing the concordance. Due to majority of the proteins were not covered by DMS, we used the commonly used variant effect prediction tools, namely SIFT (Ng & Henikoff 2003), Polyphen-2 (Adzhubei et al. 2010) and CADD (Rentzsch et al. 2019) via the Ensembl VEP to obtain the proportion of variants predicted to be functionally damaging by these tools compared to NMA. Furthermore, where possible, we used the Structural Workflow for Annotating ADME Targets (SWAAT) (Othman et al. 2022) to ascertain missense variants that are likely to be structurally damaging. SWAAT employs both sequence-based and structure-based approaches for improved accuracy (Gyulkhandanyan et al. 2020) in its predictions.

3.3.5 Molecular dynamics simulation

We used the Assisted Model Building with Energy Refinement (AMBER) (Case et al. 2005) program to conduct molecular dynamics (MD) simulation to explore the conformational energy landscape and the dynamic impact of star alleles enriched in African populations as well as their defining missense variants. MD seeks to model biological systems as atoms *in silico* and run a simulation while assuming a given potential energy function (Cheatham III & Case 2013).

Protein structures from PDB were prepared for AMBER20 using pdb4amber. We used the ff19SB (Tian et al. 2019) AMBER protein force field to describe the topology of the protein systems. The bound ligand thioridazine was parameterised using the General Amber Force Field (GAFF) (Wang et al. 2004) and the bonded configuration between the haem-bound Cys^{443} sulphur atom and the Fe metal ion were accounted for using parameters from Shahrokh et al (Shahrokh et al. 2012). Prior to production simulation, two-step combined minimisations were conducted: restrained then unrestrained using steepest decent and conjugate gradient algorithms. The first step had a maximum of 10000 cycles with a restrained constant force of 500 $kcal \cdot mol^{-1} \cdot Å^{-2}$ followed by a second minimisation with a maximum of 7000 cycles with no restraints. Minimisation was followed with a heating stage from 50.0K to 300.0K using Langevin dynamics and SHAKE algorithm restraints with a collision frequency of $5.0ps^{-1}$, and 10 $kcal \cdot mol^{-1} \cdot Å^{-2}$ restraint force for

20ps. Restraints were then gradually released in decrements of 2 $kcal \cdot mol^{-1} \cdot Å^{-2}$ using the SHAKE algorithm. Equilibration was performed at constant 1atm pressure maintained by isotropic position scaling with a relaxation time of 2ps for a range of restraints initiating at 10 $kcal \cdot mol^{-1} \cdot Å^{-2}$ and finalising at 0 $kcal \cdot mol^{-1} \cdot Å^{-2}$, with decrements of 1 $kcal \cdot mol^{-1} \cdot Å^{-2}$. Production molecular dynamics simulation was conducted at 300.0K for 600ns. Snapshots of the trajectories were captured in intervals of 10ps.

MD simulation results were analysed using PYTRAJ and MDTRAJ (Nguyen et al. 2016, McGibbon et al. 2015) Python modules. We used Root-mean-square deviation (RMSD) to determine the variant(s) most responsible for the effects of the haplotype. RMSD measures the average distance between corresponding atoms of superimposed protein structures during molecular dynamics simulation (Sargsyan et al. 2017). RMSD calculations were conducted for the entirety of the simulation time (600ns) (Nguyen et al. 2016). Root-mean-square fluctuation (RMSF) was used to study residue-level structural deviations and to infer molecular flexibility. RMSF measures the fluctuation of the alpha carbon ($C\alpha$) atoms of a protein above their average position (Martínez 2015). Furthermore, from the MD simulations, we used a custom Principal Component Analysis (PCA) to extract motions described as 'essential dynamics' from sampled conformations. PCA reduces the dimensionality of data to the most meaningful projection operators (Sittel et al. 2014). The motions we extracted can be viewed as the most prominent motions of a protein which describe the function of the protein.

3.4 Results

Predicting single- and multi-variant star allele impacts with Normal Mode Analysis

From our NMA characterisation, the mean $\Delta\Delta G$ (computed from FoldX) for all 261 CYP variants analysed in this study was 1.60 $kcal \cdot mol^{-1}$, where 82 variants were predicted to be neutral, 150 destabilising, and 29 stabilising (Table S1). On the other hand, a different distribution is observed for $\Delta\Delta S_{\text{vib}}$ (computed from ENCoM) where the mean was 0.11 $kcal \cdot mol^{-1} \cdot K^{-1}$, and the counts for variants predicted to be neutral, destabilising, and stabilising were 183, 57, and 21, respectively (Table S1). Overall, G had a wider

distribution compared to $\Delta\Delta S_{\text{vib}}$ (Figure 3.3A). The combination of FoldX and ENCoM resulted in a combined $\Delta\Delta G$ mean of 0.49 $kcal{\cdot}mol^{-1}$ with 146 missense variants predicted to be neutral to the protein structures, 85 variants predicted to be destabilising, and 30 predicted to be stabilising.

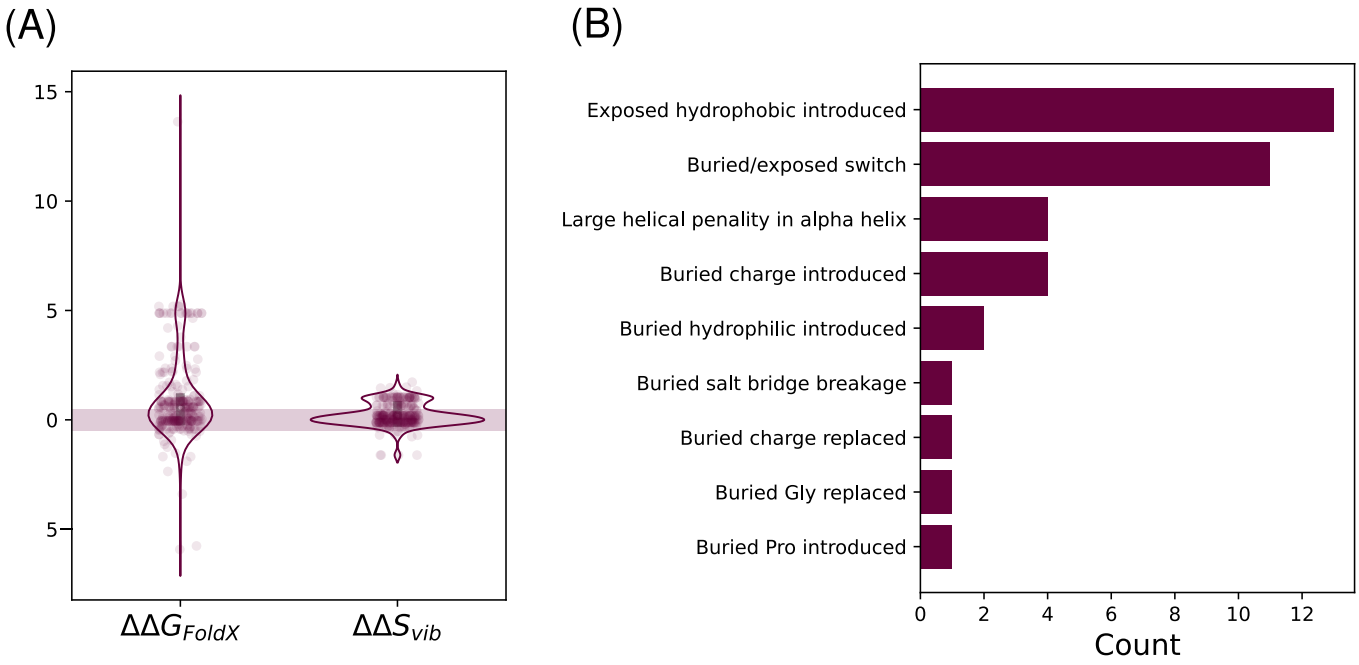


Figure 3.3: **Distribution of Fold-X ($\Delta\Delta G$) and ENCoM-computed ($\Delta\Delta S_{\text{vib}}$) missense variant scores and SWAAT red-flag protein structural feature count.** (A) FoldX and ENCoM-computed scores distribution. The highlighted region represents energies where the missense variants are expected to have neutral effects on protein stability (Frappier & Najmanovich 2014). (B) SWAAT red-flag. Many SWAAT structural consequences are linked to potentially functional altering disruptions in protein electrostatic forces (Gonzalez et al. 2022, Vascon et al. 2020, Gong et al. 2017, Ittisoponpisan et al. 2019).

For multi-variant star alleles (Table S2), according to the $\Delta\Delta G$ output from FoldX, 72/92 haplotypes were destabilising, 4/92 were stabilising, and 16 fell within the neutral range ($-0.5\ kcal \cdot mol^{-1} < \Delta\Delta G < 0.5\ kcal \cdot mol^{-1}$) (Figure S1). On the other hand, based on the $\Delta\Delta S_{\text{vib}}$ output from ENCoM, only 1/92 were found to be stabilising, 51/92 haplotypes were predicted to be destabilising, while the remaining 40 haplotypes fell within the neutral range ($-0.5\ kcal{\cdot}mol^{-1}{\cdot}K^{-1} < \Delta\Delta S < 0.5\ kcal{\cdot}mol^{-1}{\cdot}K^{-1}$) (Figure S1). Combined $\Delta\Delta G$ predicted 36, 24 and 32 haplotypes to be destabilising, stabilising and neutral, respectively.

3.4.1 Comparison with computational variant effect predictors

We observed that 64/261 (25%) variants had concordant deleterious effect predictions according to NMA and the consensus (2 from 3 rule) of SIFT, CADD, PolyPhen-2, while 146/261 (56%) variants had a concordant neutral effect prediction. Therefore, a combined 80% of predictions from the VEP tools were concordant with our NMA results (Table S1). Furthermore, 45/261 (17%) of the variants that had NMA scores that were considerably divergent from the wildtype protein structures had red flags according to predictions from SWAAT (Table S3). The introduction of an exposed hydrophobic residue (13/45 missense variants) and exposure of previous buried amino acids (11/45 missense variants) accounted for many of the observed red-flag protein structural consequences (Figure 3.3B and Table S3). These structural may affect protein structural integrity through disruptions in protein electrostatic forces (Gonzalez et al. 2022, Vascon et al. 2020, Gong et al. 2017) and potentially affect function (Ittisoponpisan et al. 2019).

3.4.2 Comparison with Deep Mutational Scanning

We compared 71 of the 90 *CYP2C9* and *CYP2C19* variants analysed using NMA with corresponding DMS abundance and activity values (Amorosi et al. 2021, Zhang et al. 2020). A total of 66% of our NMA variant structural impact predictions (based on combined $\Delta\Delta G$) showed concordance with DMS for *CYP2C9* and *CYP2C19* missense variants (Table 3.1). Out of the 71 variants with both NMA and DMS results, 29 (41%) were reported to be 'tolerated' according to DMS studies and 'neutral' by our NMA study. Only 18/71 (25%) variants were reported to be damaging/disruptive by both DMS and NMA.

Table 3.1: Comparison of NMA-based CYP2C9 and CYP2C19 missense variant effect predictions with deep mutational scanning results.

Mutation	**Star allele**	**Normal Mode Analysis**		**Deep Mutational Scanning**	
		$\Delta\Delta G$ **(kcal.mol-1)**	**Structural Impact**	**Abundance Score**	**Functional Study**
R433W	*CYP2C19*5*	1.750	Destabilising		
R132Q	*CYP2C19*6*	0.068	Neutral	0.718	Tolerated
W120R	*CYP2C19*8*	0.519	Neutral	0.480	Severely Damaging
R144H	*CYP2C19*9*	0.562	Destabilising	0.621	Damaging
P227L	*CYP2C19*10*	0.731	Destabilising	0.625	Damaging

Table 3.1 continued from previous page

Mutation	Star allele	Normal Mode Analysis		Deep Mutational Scanning	
		$\Delta\Delta G$ (kcal.mol-1)	Structural Impact	Abundance Score	Functional Study
R150H	*CYP2C19*11*	0.168	Neutral	0.693	Tolerated
R410C	*CYP2C19*13*	-0.053	Neutral	0.677	Tolerated
R329H	*CYP2C19*18*	0.179	Neutral		
S51G	*CYP2C19*19*	-0.442	Neutral		
R186P	*CYP2C19*22*	-0.832	Stabilising	0.560	Severely Damaging
G91R	*CYP2C19*23*	0.396	Neutral	0.668	Tolerated
R335Q	*CYP2C19*24*	-0.046	Neutral	0.601	Damaging
D256N	*CYP2C19*26*	0.052	Neutral		
V374I	*CYP2C19*28*	0.565	Destabilising	0.628	Damaging
H78Y	*CYP2C19*31*	-0.053	Neutral		
H99R	*CYP2C19*32*	0.551	Destabilising		
D188N	*CYP2C19*33*	0.188	Neutral	0.678	Tolerated
E122A	*CYP2C19*39*	-0.582	Stabilising	0.754	Tolerated
aI331V	Many CYP2C19 star alleles	0.008	Neutral	V331I=0.600	V331I=Damaging
I359L	*CYP2C9*18*	-0.373	Neutral	0.782	Tolerated
D379A		-1.419	Destabilising		
R125L	*CYP2C9*35*	-0.540	Stabilising	0.917	Tolerated
R144C	*CYP2C9*35, *61*	0.985	Destabilising	0.950	Tolerated
N457S	*CYP2C9*61*	0.134	Neutral		
E272G	*CYP2C9*71*	-1.041	Destabilising	1.037	Tolerated
P489S		0.259	Neutral		
A149T	*CYP2C9*46*	1.373	Destabilising	0.605	Damaging
A149V	*CYP2C9*72*	1.506	Destabilising	0.367	Severely Damaging
D191G	*CYP2C9*69*	1.456	Destabilising	0.278	Severely Damaging
D360E	*CYP2C9*5*	1.025	Destabilising		
D49G	*CYP2C9*37*	2.045	Destabilising	0.787	Tolerated
E300V	*CYP2C9*81*	-0.340	Neutral	0.954	Tolerated
E326D	*CYP2C9*65*	1.133	Destabilising	0.585	Damaging
E354K	*CYP2C9*24*	3.103	Destabilising	-0.047	Severely Damaging
F110S	*CYP2C9*40*	0.447	Neutral	1.096	Tolerated
G40R	*CYP2C9*76*	13.174	Destabilising	0.181	Severely Damaging
G431R	*CYP2C9*84*	7.983	Destabilising	0.864	Tolerated
G442V	*CYP2C9*70*	0.189	Neutral	0.642	Damaging
G70R	*CYP2C9*20*	4.572	Destabilising	0.326	Severely Damaging
G96A	*CYP2C9*38*	0.537	Destabilising	1.102	Tolerated
G96R	*CYP2C9*77*	2.301	Destabilising	0.632	Damaging
G98V	*CYP2C9*39*	3.229	Destabilising	0.764	Tolerated
H251R	*CYP2C9*9*	-0.155	Neutral	1.009	Tolerated
I207T	*CYP2C9*48*	0.291	Neutral	0.737	Tolerated
I222V	*CYP2C9*49*	0.208	Neutral	1.022	Tolerated
I284V	*CYP2C9*51*	0.024	Neutral	0.998	Tolerated
I327T	*CYP2C9*31*	0.410	Neutral	0.553	Severely Damaging

Table 3.1 continued from previous page

Mutation	Star allele	Normal Mode Analysis		Deep Mutational Scanning	
		$\Delta\Delta G$ (kcal.mol-1)	Structural Impact	Abundance Score	Functional Study
I359T	*CYP2C9*4*	0.607	Destabilising	1.141	Tolerated
I387V	*CYP2C9*56*	0.233	Neutral	1.122	Tolerated
I434F	*CYP2C9*59*	-0.209	Neutral		
K119R	*CYP2C9*41*	-0.221	Neutral	0.865	Tolerated
L128R	*CYP2C9*78*	0.120	Neutral	1.025	Tolerated
L361I	*CYP2C9*55*	0.449	Neutral	0.954	Tolerated
L362V	*CYP2C9*66*	0.318	Neutral	0.953	Tolerated
L36F	*CYP2C9*83*	0.536	Destabilising	1.206	Tolerated
L90P	*CYP2C9*13*	1.963	Destabilising	0.534	Severely Damaging
N204H	*CYP2C9*57*	1.141	Destabilising	0.589	Damaging
N418T	*CYP2C9*75*	0.169	Neutral		
N41D	*CYP2C9*22*	-0.092	Neutral	0.377	Severely Damaging
P163L	*CYP2C9*47*	0.333	Neutral	0.685	Tolerated
P227L	*CYP2C9*82*	0.348	Neutral	0.573	Severely Damaging
P227S	*CYP2C9*50*	0.157	Neutral		
P279T	*CYP2C9*29*	0.302	Neutral		
P317S	*CYP2C9*53*	0.439	Neutral	0.762	Tolerated
P337T	*CYP2C9*58*	1.322	Destabilising	0.631	Damaging
P346L	*CYP2C9*79*	1.333	Destabilising		
P382S	*CYP2C9*17*	0.997	Destabilising	0.667	Damaging
Q214H	*CYP2C9*74*	1.688	Destabilising	0.761	Tolerated
Q214L	*CYP2C9*28*	-0.181	Stabilising	0.895	Tolerated
Q454H	*CYP2C9*19*	1.151	Destabilising	0.460	Severely Damaging
R124Q	*CYP2C9*42*	-0.918	Stabilising	0.833	Tolerated
R124W	*CYP2C9*43*	0.079	Neutral		
R125C	*CYP2C9*62*	0.109	Neutral	0.962	Tolerated
R125H	*CYP2C9*14*	0.206	Neutral	0.999	Tolerated
R132Q	*CYP2C9*33*	0.138	Neutral		
R132W	*CYP2C9*45*	0.391	Neutral	0.738	Tolerated
R144H	*CYP2C9*63*	1.125	Destabilising	0.988	Tolerated
R150C	*CYP2C9*73*	-0.222	Neutral	0.803	Tolerated
R150H	*CYP2C9*8*	0.100	Neutral	0.739	Tolerated
R150L	*CYP2C9*27*	-0.455	Neutral	0.900	Tolerated
R335Q	*CYP2C9*34*	-0.027	Neutral		
R335W	*CYP2C9*11*	0.154	Neutral	0.244	Severely Damaging
R433W	*CYP2C9*67*	0.219	Neutral		
S280C	*CYP2C9*64*	-0.068	Neutral	0.890	Tolerated
S343R	*CYP2C9*54*	-0.153	Neutral	0.901	Tolerated
T130M	*CYP2C9*44*	0.971	Destabilising	0.759	Tolerated
T130R	*CYP2C9*26*	3.746	Destabilising	0.954	Tolerated
T299A	*CYP2C9*16*	-0.571	Stabilising	1.174	Tolerated
T299R	*CYP2C9*52*	1.877	Destabilising	0.300	Severely Damaging
V76M	*CYP2C9*23*	1.034	Destabilising	0.295	Severely Damaging

[a]Zhang et al. (2020) assayed effects of V33I; however, we know from PharmVar that I331V (rs3758581) is the most common allele for CYP2C19. I331V is known to be normal function Relling & Klein (2011). Normal Mode Analysis (NMA) predicts the protein structural stability impact while DMS reports the functional assessment results measured in abundance scores.

CYP2C9 abundance score thresholds:

Abundance score $< 0.578 \rightarrow$ severely damaging

Abundance score $< 0.670 \rightarrow$ damaging

Abundance score $> 0.670 \rightarrow$ tolerated

CYP2C19 abundance score thresholds:

Abundance score $< 0.597 \rightarrow$ severely damaging

Abundance score $< 0.635 \rightarrow$ damaging

Abundance score $> 0.635 \rightarrow$ tolerated

3.4.3 Comparison with star allele CPIC clinical functions

Comparison of CYP star allele impact predictions based on NMA (combined $\Delta\Delta G$) with known clinical functional impacts are highlighted in Figure 3.4A. Five of the 9 normal function star alleles exhibited neutral protein structural effects as expected while the other 4 were slightly stabilising. Twelve of the known 19 decreased function star alleles exhibited either stabilising or destabilising structural effects, while the other 7 unexpectedly exhibited neutral effects. Of the known non-functional star alleles, 14 were perturbing to the protein structure, with the majority (11/14) being destabilising. Conversely, four of the known non-functional star alleles exhibited neutral effects on the protein structure. Of the existing star alleles with unknown/uncertain function, **37*, **64*, **72*, **83*, **87*, **93*, **94*, **95*, **105*, **110*, **111*, **113*, **127*, **139*, **141* and **142* in *CYP2D6*; **18*, **20*, **58*, **69* **72*, **74*, **76*, **77*, **79* and **84* in *CYP2C9*; **2* and **4* in *CYP2F1*; **2*, **5* and **6* in *CYP2W1*; **21* and **45* in *CYP2B6*; **14* in *CYP2C8*; and **2*, **12*, **13* in *CYP3A4* were found to be the most perturbing to their respective protein structures, based on NMA (Figure 3.4A and Table S1).

Given some of the above inconsistencies, we further assessed whether NMA predictions based on RMSIP overlaps with the wildtype protein structures provided better concordance with known clinical functional impacts (Figure 3.4B). An RMSIP value of $\geq$0.7 is

indicative of significant concordance with the wildtype (David & Jacobs 2011). All the CYP star alleles analysed in this study had RMSIP values >0.99 (see also Discussion).

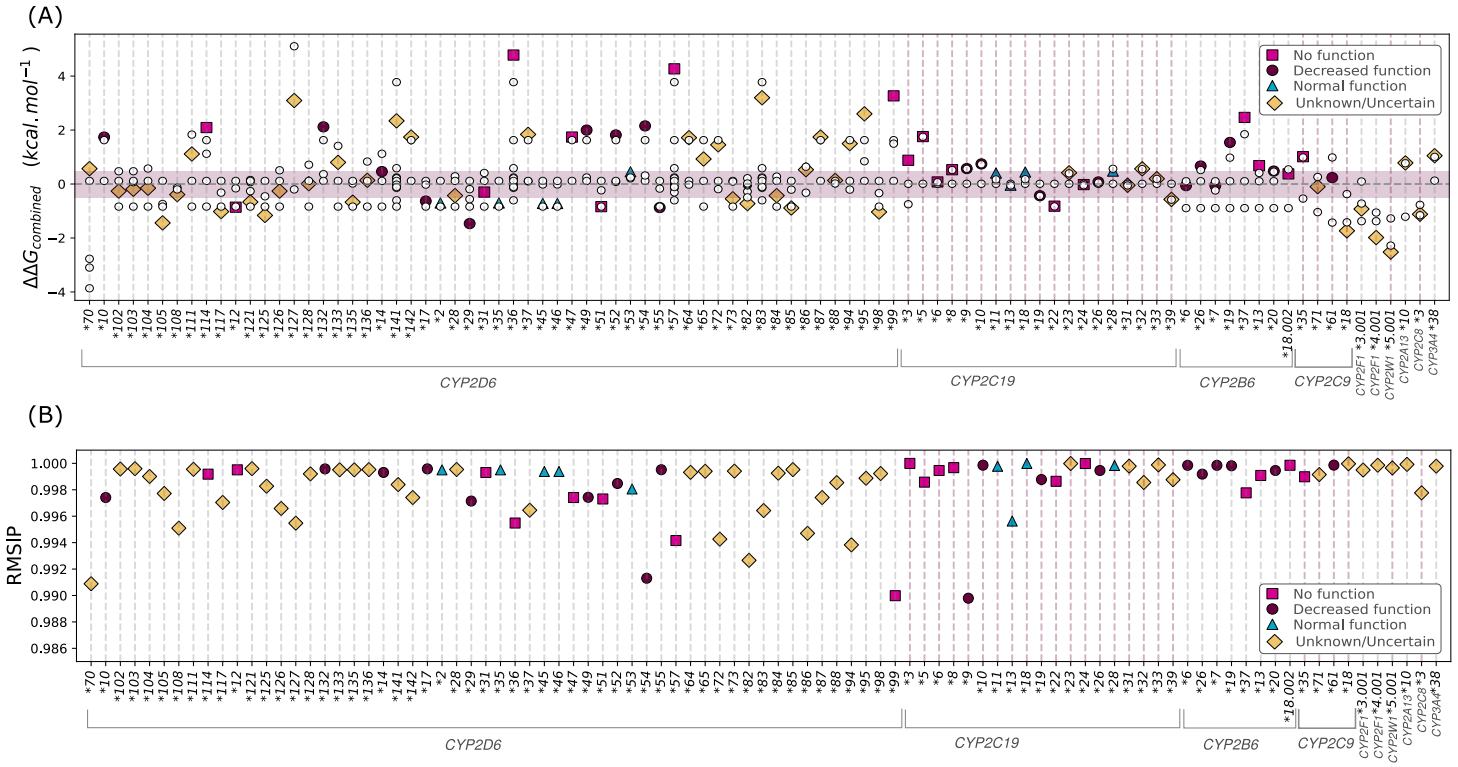


Figure 3.4: **Assessment of the concordance of CPIC clinical functional classification with NMA predications.** This figure focuses on star alleles defined by multiple missense variants as $\Delta\Delta G$ scores are provided in Table S1. (A) Concordance based on haplotype combined $\Delta\Delta G$ scores. Haplotypes with Combined $\Delta\Delta G$ scores that fall within the range $0.5 \leq combined\Delta\Delta G \leq 0.5$ are expected to have neutral effects on protein stability (Frappier & Najmanovich 2014). The black-outlined white circles represent defining missense variants that make up each haplotype, and a key to the haplotype CPIC clinical function classification is provided. (B) Concordance based on NMA-derived RMSIP scores for star alleles defined by multiple missense variants. Mutant proteins with RMSIP values closer to 1 are interpreted as having increased structural overlap with the wildtype protein structure and are thus expected to exhibit wildtype-like function. A key of the haplotype CPIC clinical function classification is provided.

3.4.4 Molecular dynamics simulation of *CYP2D6*2* and *CYP2D6*17*

Although we were able to run SWAAT and NMA on a larger set of missense variants from multiple CYP proteins, MD simulation was performed only for *CYP2D6*2* and

*CYP2D6*17* as a model given that it takes a long time and requires considerable computational resources.

Root-mean-square deviation

Our RMSD results are presented in Figure 3.5. Although R296C (in both *CYP2D6*2* and *CYP2D6*17*) converged right at the start of the simulation and were able to maintain a stable RMSD of 1.5Å for the entirety of the simulated time, R296C exhibited fluctuations that were most deviant from the wildtype protein system when compared to all other simulated systems. Variants S486T and T107I also showed deviations from the wildtype; however, these deviations were moderate and resembled the wildtype for most of the simulation. S486T achieved convergence at 100 ns and remained relatively stable for the remainder of the simulation time. T107I had fluctuations close to those of the wild-type but had more deviations than S486T. No clear convergence was observed for T107I for the entirety of our MD simulation (0 to 600 ns). *CYP2D6*17* (i.e. combination of T107I, R296C, and S486T) had higher RMSD values compared to *CYP2D6*2* (combination of R296C and S486T), alluding to it being the less stable of the two star alleles, and the one with more altered conformational dynamics.

Root-mean-square fluctuation

RMSF measures the fluctuations of the alpha carbon (Cα) atoms of a protein above their average position which is informative of residue-level structural deviations and molecular flexibility (Martínez 2015). There were no observable major deviations in RMSF for variant S486T compared to the wildtype (Figure 3.6). CYP2D6 R296C and T107I mutant proteins exhibit fluctuations that largely resemble those of the wildtype for the most part; however, significant deviations are observed around the F- and G-helix region that also covers the FG loop (Figure 3.6). This indicates that R296C and T107I mutants have reduced flexibility in this region compared to the wild-type. Comparison of the same F- and G-helix region of star alleles *CYP2D6*2* and *CYP2D6*17* reveals that the *CYP2D6*17*-encoded mutant is more rigid in the crucial (Poulos & Johnson 2005) FG-loop region compared to the *CYP2D6*2*-encoded mutant which has no significant deviations in residue fluctuations.

Principal component analysis

We used PCA to determine the mutant protein(s) with atomic coordinates most similar or divergent to those exhibited by the wildtype (Figure 3.7). Variant S486T was excluded because, according to all previous measures, it did not confer any deviations from the wild-type that may suggest it has a significant impact on protein stability and conformational dynamics. The R296C mutant's subspace shows great overlap with the wildtype compared to the mutant with T107I as the latter's atoms sample a greater range of subspaces (Figure 3.7). Regarding the full haplotypes, both *CYP2D6*2* and *CYP2D6*17*-encoded mutants confer deviant motions to the wild-type. However, a more deviant pattern is observed for *CYP2D6*17*.

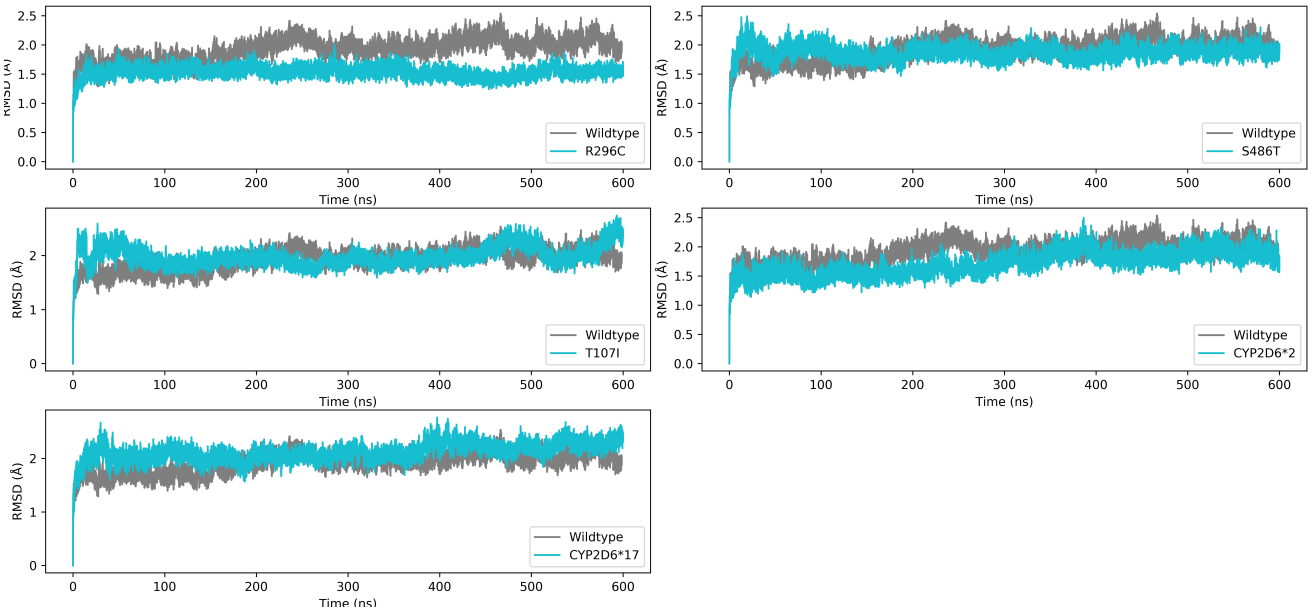


Figure 3.5: **Analysis of backbone protein structural deviations by selected CYP2D6 variants and haplotypes.** RMSD of the CYP2D6 mutant protein systems is shown in blue while that for the wild-type is shown in grey. The protein structure encoded by R296C exhibited the most deviations from the wildtype protein structure; however, it converged early in the simulation time (50ns). No clear convergence was observed for T107I. RMSD for the two haplotypes, *CYP2D*2* (R296C and S486T) and **17* (**2*-backbone + T107I), was inconclusive regarding the characterisation of protein structural deviations.

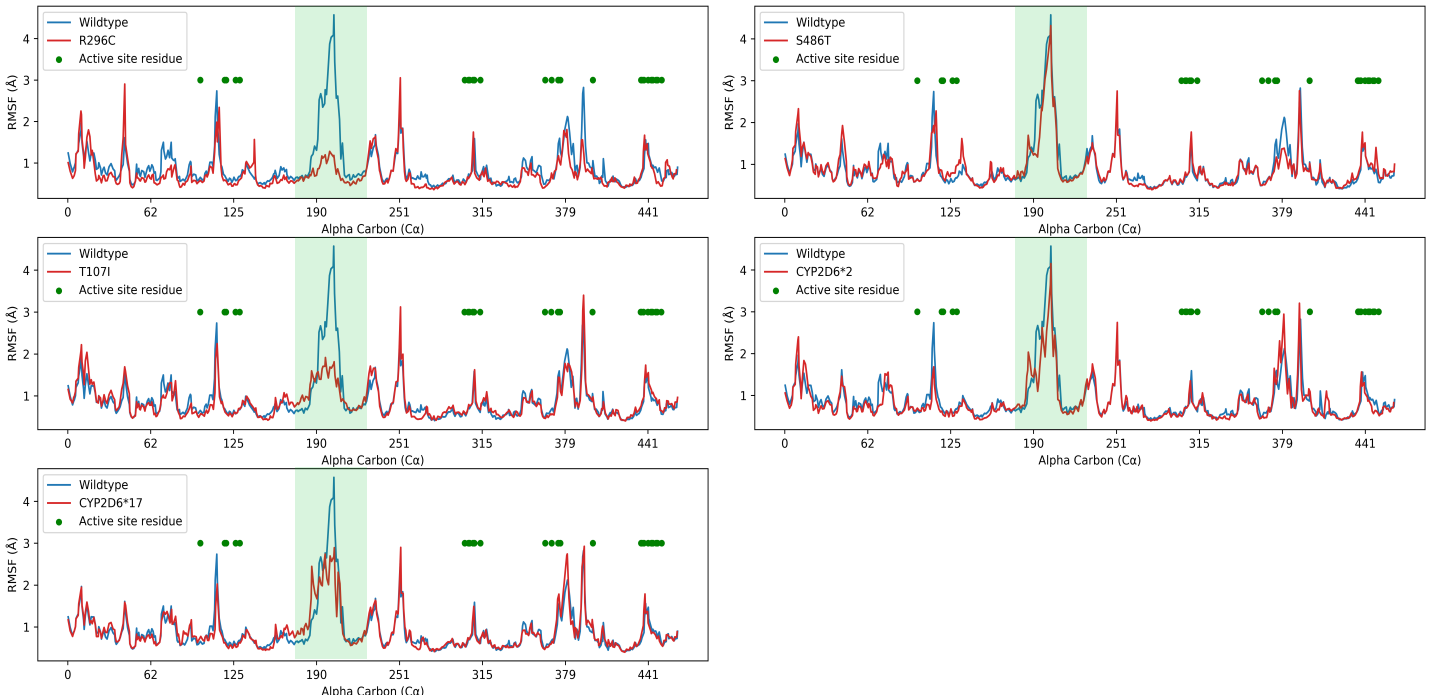


Figure 3.6: **Analysis of residue-level fluctuations induced by selected CYP2D6 variants and haplotypes.** Fluctuations for mutant protein structures (encoded by *CYP2D6*2*, *CYP2D6*17*, R296C, S486T and T107I) are indicated in red while fluctuations for the wildtype are shown in blue. Spikes in root-mean-square fluctuations (RMSF) represent increased flexibility. The analysis is directed towards the functionally crucial F-G loop region (green). Additionally, we have represented active site residues within a 4Å radius from the catalytic haem group. There were no significant deviations of the highlighted mutant active site residues relative to the wild-type residues.

3.5 Discussion

The functional impact of the majority of the known star alleles in ADME genes is currently unknown or uncertain (Gaedigk et al. 2018). This is compounded by the fact that the rate of star allele discovery is increasing given the enhanced accessibility to genome sequencing and presents challenges in clinical pharmacogenomics implementation globally. Clinical and experimental functional assays are expensive and may not be scalable, especially for rare star alleles. However, advances in structural bioinformatics provide a pathway for studying the functional impact of missense variants from a protein structural analysis perspective. Therefore, the main question we aimed to address is whether structural bioinformatics approaches can provide insights into the impact of single and/or combinations of missense variants that define star alleles with regard to CYP protein

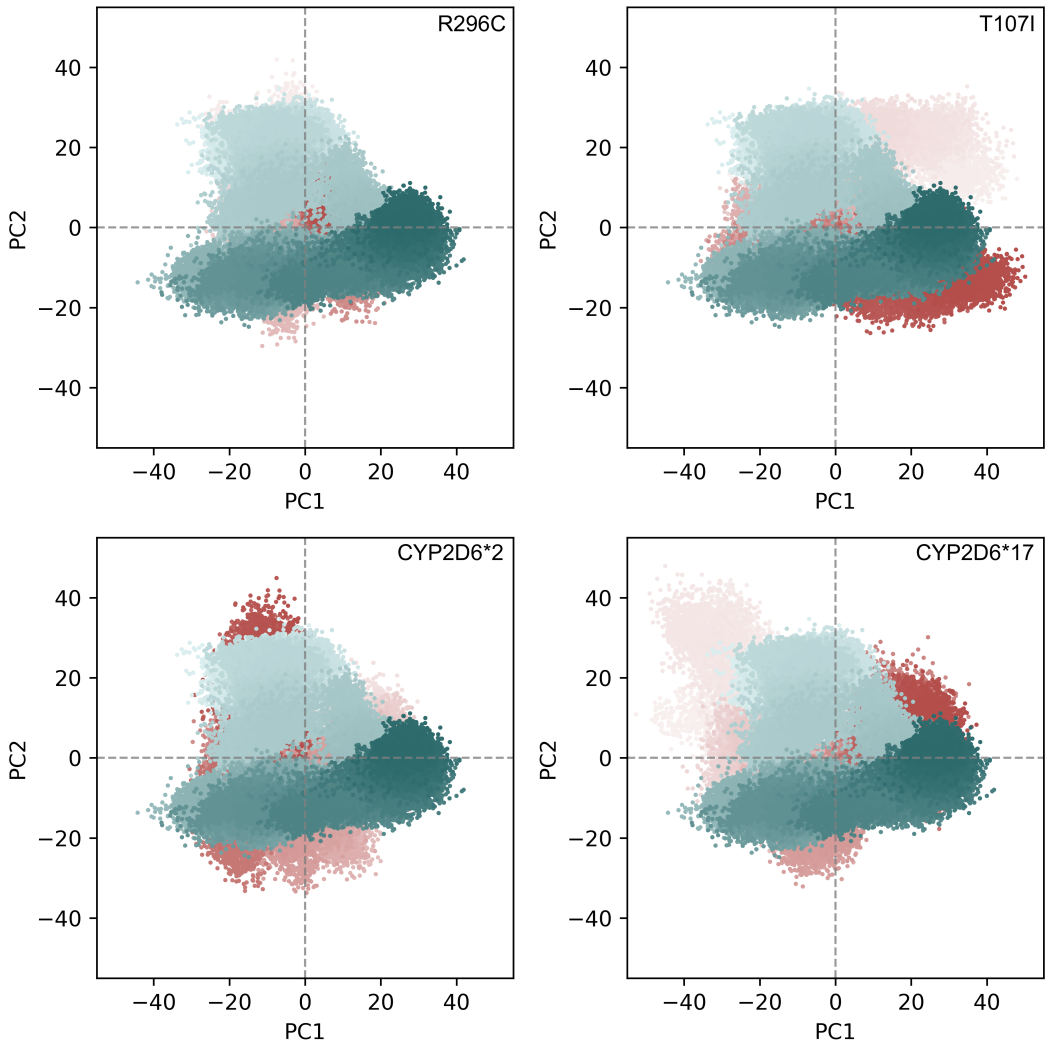


Figure 3.7: **Principal Component Analysis from Molecular Dynamics simulations.** The coordinates of C atoms of mutant (red) CYP2D6 protein systems (*CYP2D6*2*, *CYP2D6*17*, R296C and T107I) overlapped with CYP2D6 wildtype (blue) protein coordinate, on the first and second principal components, for 600ns trajectories. The structure encoded by T107I sampled a larger subspace along the two principal components compared to the other mutant protein structures.

structure and function. Herein we focus on the application of these techniques to star alleles in 9 CYP proteins included in the PharmVar database.

In our normal mode analysis, we computed $\Delta\Delta G$ changes caused by CYP missense variants using FoldX and ENCoM and this enabled the classification of 261 variants as either perturbing (115/261; 44%) to their respective CYP protein structures or neutral (56%). The perturbations could either be in the form of protein structure stabilisation or destabilisation which may translate into alterations to interactions with ligands (i.e.

drugs) (Koulgi et al. 2022, Zhang et al. 2012). The majority of variants with NMA values that are highly divergent from those of wild-type proteins are also predicted to be deleterious by SWAAT (Othman et al. 2022) and flagged to cause potentially function-altering changes such as exposing residues that are buried in wildtype structures (Table S3). The use of $\Delta\Delta G$ for assessing the functional impact of variants within CYP star alleles via protein structural conformational changes has been previously exemplified by Borba et al. (Borba et al. 2016). However, this previous study only focused on variants defining *CYP2D6*7* and **114* (formerly **14A*). Our results for the *CYP2D6*7*-defining variant (H324P; destabilising) and the *CYP2D6*114*-defining variants (Q169R; - destabilising; P34S - destabilising; R296C - stabilising; S486T - neutral) are consistent with the findings by Borba et al., and they also support the known CPIC functional impacts for *CYP2D6*7* (non-functional) and *CYP2D6*114* (non-functional) (Relling & Klein 2011). In the present study, we build on these efforts by providing an extensive assessment of predicted CYP protein structure perturbations for over 200 missense-variant-defined star alleles, 91 of which are defined by multiple missense variants.

Regarding comparisons between NMA and other variant effect prediction and/or quantification approaches, we observed an 80% (208/261) concordance between NMA-derived predictions and the consensus of SIFT, Polyphen-2 and CADD, while the concordance between NMA and DMS was 66% (47/71). However, the latter was computed only for missense variants in *CYP2C9* and *CYP2C19* (the only CYP proteins with published DMS data). Compared to DMS, NMA categorised 8% more missense variants as deleterious. However, it should be noted that these techniques assay different aspects related to missense variant functional effects. NMA provides a measure of protein structure perturbations based on $\Delta\Delta G$ changes while DMS is typically used for massively parallel assessment of missense variant effect on CYP enzyme expression (e.g. via a fluorescent-based abundance assay) or activity in the presence of a probe drug within a suitable model (e.g. yeast) (Zhang et al. 2020, Amorosi et al. 2021). In effect, these approaches are complementary with regard to providing insights and additional context on missense variant functional impact.

For multi-mutant star alleles, in some cases, we found that the NMA-derived $\Delta\Delta G$ of the respective protein structures is the combined effect of some or all of their defining missense variants. For other star alleles, the $\Delta\Delta G$ values suggest that their effects on

protein structural stability and conformation are driven by one of the missense variants. This may imply that such variants potentially have a more dominant impact on the protein function. *CYP2D6*17* and *CYP2D6*29* are examples of star alleles that exhibit a somewhat additive interaction of their respective missense variants. On the other hand, folding energies corresponding to star alleles such as *CYP2D6*10* may be defined (according to our combined $\Delta\Delta G$ analysis) by a single missense variant. The apparent differences in the interaction of variants that make up these star alleles present complexity in the analysis of CYP genetic variation. Therefore, for multi-variant star alleles, looking at each variant in isolation may not capture the full picture of the star allele's impact on protein structure stability and potential functional implications.

In addition to NMA, we performed a proof-of-concept investigation into the utility of MD simulations as part of the structural bioinformatics techniques potentially capable of informing CYP star allele function. However, based on previous experience (Othman et al. 2021, da Rocha et al. 2022), MD simulations are typically lengthy to perform and therefore we could only analyse two model haplotypes with this approach within the time constraints of our study. We prioritised two well-characterised alleles (**2* and **17*) (Caudle et al. 2020, Relling & Klein 2011) in CYP2D6 as it is one of the major enzymes in pharmacogenetics (involved in the processing of approximately 25% of clinical drugs (Ingelman-Sundberg et al. 2007)). *CYP2D6*2* was prioritised because it forms the backbone of many other CYP2D6 star alleles, while CYP2D6*17 was included to assess the impact of mutating a *CYP2D6*2*-backbone and also given its well-documented decreased function annotation (Masimirembwa & Hasler 1997) that served as a benchmark. *CYP2D6*2* has a frequency of 7-20% across global populations, while *CYP2D6*17* is particularly frequent across sub-Saharan African (SSA) populations (frequency of 6-30%) (Twesigomwe et al. 2023).

From our MD simulations, RMSF output indicated that of the missense variants in *CYP2D6*2* (R296C and S486T) and *CYP2D6*17* (**2* missense variants + T107I), R296C and T107I have detrimental effects on the F-G loop region and would thus make it more difficult for a substrate to interact with the enzyme active site (Poulos et al. 1987, Fischer et al. 2018, Kumar 2010). However, although R296C individually rigidifies the F-G loop to a larger extent compared to T107I (Figure 3.6), in addition to causing more profound RMSD deviations from the wild-type protein system (Figure 3.5), the

combination of R296C and S486T in *CYP2D6*2* has a wildtype-like F-G loop fluctuation. In comparison, the combination of R296C, S486T and T107I in *CYP2D6*17* has a relatively rigid F-G loop. Moreover, the PCA plots depicting the atom trajectories in our MD simulation indicate that the *CYP2D6*17*-encoded protein mutant samples a wider subspace (Figure 3.7) compared the *CYP2D6*2*-encoded mutant, implying that **17* confers a larger divergence from wildtype-like fitness compared to **2*. These findings imply that there may be complex inter-modulating effects on protein structure across missense variants within star alleles as hypothesised by others (Bapiro et al. 2002). While not comprehensive, these results also provide context—from a structural bioinformatics perspective—supporting the CPIC clinical functions observed for *CYP2D6*2* (normal function) and **17* (decreased function), respectively. However, the MD results for the R296C occurring as the single missense variant in a haplotype (as in *CYP2D6*34*) are contradictory to the known CPIC clinical function of *CYP2D6*34* (normal function) (Relling & Klein 2011). This implies, that structural bioinformatics approaches alone are not sufficient to address the gaps in knowledge on the functional impact of missense variation in a pharmacogenomics context, but should rather be used as one of the sources of evidence in this endeavour.

There were some limitations in this study. Firstly, we did not investigate the substrate-specific effects of the proteins included in our analysis. However, the insights from the generic effects of the investigated variants provide an important baseline for drug-gene pair-focused analysis. Secondly, the structural bioinformatics approach is only suitable for star alleles defined by missense variants. Nonetheless, our findings are highly relevant to the pharmacogenomics community since most star alleles without CPIC functional curation are defined by missense variants. Lastly, we obtained MD insights from only *CYP2D6*2* and *CYP2D6*17* among the existing star alleles. Future studies involving other key star alleles are warranted to obtain a more comprehensive benchmark for the application of structural bioinformatics in assessing star allele function.

From a broader perspective, we acknowledge the limitations associated with structural bioinformatics approaches. Firstly, protein function is multifactorial, and aspects such as gene regulation and gene-environmental interactions, among others, cannot be captured by these approaches. This could explain why the structural stability effects predicted for a number of star alleles did not match expected predictions given their

known CPIC clinical functions (Figure 3.4). For example, *CYP2B6*6* was predicted to be neutral according to combined $\Delta\Delta G$ computed from NMA, yet it would be expected to be damaging given that it is a well-characterised decreased function star allele (Desta et al. 2019). Furthermore, we found metrics such as NMA-computed RMSIP to be relatively uninformative regarding categorising star alleles from a CYP structural and/or functional impact perspective (except for a few such as *CYP2D6*99*). For predictions based on combined $\Delta\Delta G$, one way to mitigate discordances with observed clinical impacts could involve adjusting the thresholds for determining protein structure perturbing haplotypes as the current thresholds ($\Delta\Delta G <$ -0.5 for stabilising; $\Delta\Delta G >$ 0.5 for destabilising) were not validated in pharmacogenomics-focused studies. Secondly, structural bioinformatics relies on protein models, some determined experimentally (Burley et al. 2021) and others predicted from amino acid sequences (Jumper et al. 2021). The quality of protein structures varies and heavily influences the accuracy of the predictions (Ko & Lee 2021). Lastly, many structural bioinformatics tools seek to assess protein conformational states and dynamics (such as the tools used in this study). However, it is challenging to capture the full range of conformational flexibility (de Almeida Paiva et al. 2022, Palamini et al. 2016, Samish et al. 2015). Nonetheless, techniques such as MD simulations are complementary to static protein models derived from the commonly used experimental techniques (Torrens-Fontanals et al. 2020). Moreover, structural bioinformatics is particularly useful for hypothesis generation and prioritisation of potentially damaging genetic variation due to its high-throughput capabilities and it is important in pharmacogenomics where the majority of the recently discovered star alleles are without functional annotations.

3.6 Conclusion

This study provides insights regarding the utility of structural bioinformatics approaches in addressing the variant/haplotype functional impact prediction problem in the context of pharmacogenomics. Based on NMA, we observe that quantifying the protein stability and energy changes resulting from missense mutations has the potential to elucidate their functional consequence. For example, we predict that 37 of the existing CYP star alleles with unknown/uncertain functions are perturbing to their respective protein struc-

tures. This protein conformational information can be used in addition to predictions by computational variant effect predictors and/or DMS to identify uncharacterised variants/haplotypes with potentially significant impact on CYP enzyme function. Furthermore, we present insights into how MD simulations can inform the mechanisms by which either single or combinations of missense variants impact CYP protein structure (and by extension potential function) using two CYP2D6 star alleles (**2* and **17*) as a model. MD for these star alleles indicated that the missense variants within multi-mutant carrier proteins tend to have cross-modulating effects on the stability and conformational dynamics properties. MD simulation is therefore a relatively inexpensive and highly useful technique for prioritising star alleles lacking functional curation for further experimental or clinical phenotype studies.

Author contributions

Conceptualization: N Khoza and H Othman. Data curation and analysis: N Khoza and H Othman. Data interpretation or technical support: N Khoza, D Twesigomwe and H Othman. Software: H Othman. Methodology: N Khoza and H Othman. Writing – original draft: N Khoza. Writing – review and editing: N Khoza, H Othman and D Twesigomwe. Supervision: H Othman. All authors approved the final version of the manuscript.

Acknowledgments

The authors acknowledge the Centre for High Performance Computing (CHPC), South Africa, for providing computational resources for this research project. The authors would like to thank S Hazelhurst (at the Sydney Brenner Institute for Molecular Bioscience) for reviewing the manuscript and the Wits-GSK ADME team for their support and guidance.

Supplementary

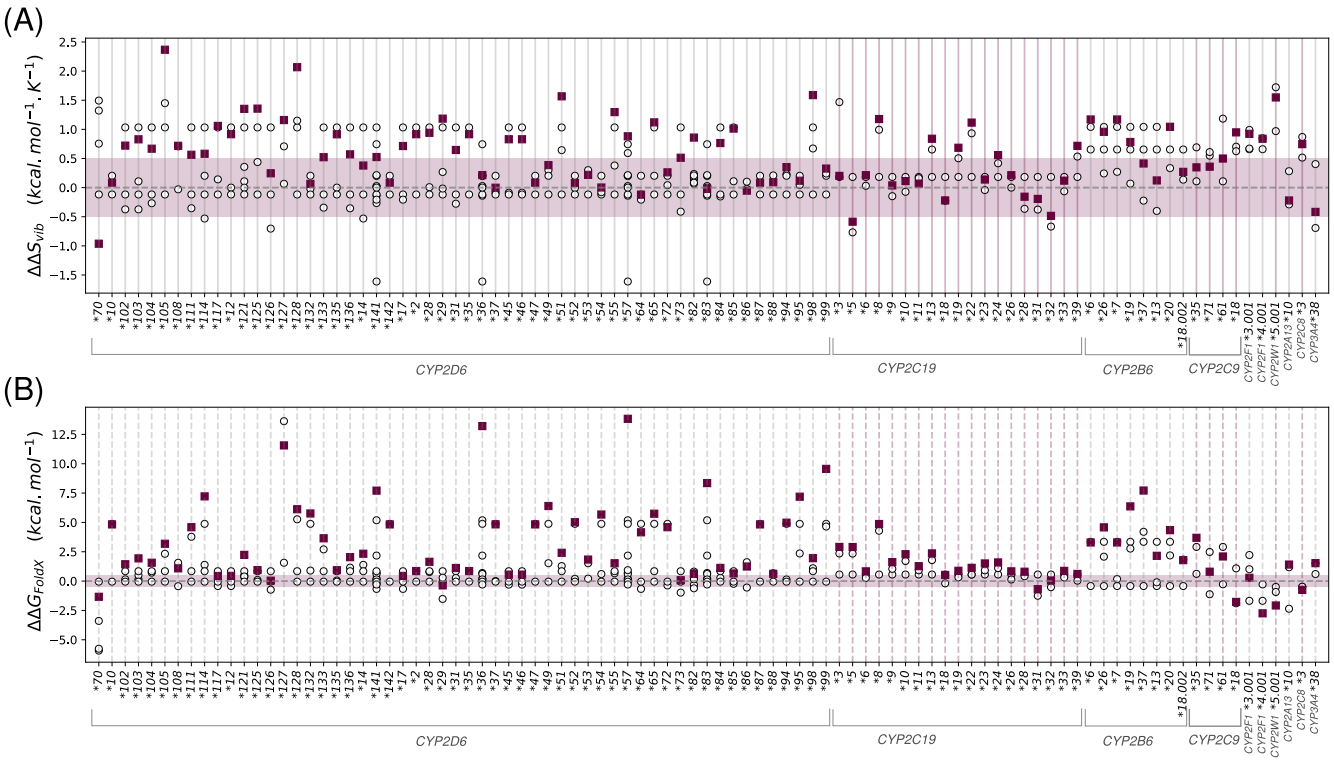


Figure S1. **Comparison of ENCoM-computed ($\Delta\Delta S_{\mathbf{vib}}$) and FoldX-computed ($\Delta\Delta G$) star allele and missense variant energy scores.** (A) ENCoM $\Delta\Delta S_{\text{vib}}$ scores. (B) FoldX $\Delta\Delta G$ scores. Each haplotype is represented by a maroon square and their defining missense variants are represented by circles. Figure points that fall within the highlighted region are expected to have neutral effects on protein stability (Frappier & Najmanovich 2014). $\Delta\Delta G$ predicts most of the variants and haplotypes as destabilising to the protein-encoded protein structures. Both $\Delta\Delta G$ and $\Delta\Delta S_{\text{vib}}$ seem to exhibit a destabilising bias for both haplotype and variants. However, $\Delta\Delta S_{\text{vib}}$ predicted more haplotypes and variants to have neutral effects on protein stability than $\Delta\Delta G$ (see also Figure 3.3A).

Table S1: Missense variant NMA analysis.

Gene	Mutation	Normal Mode Analysis				Computational Variant Effect Predictors			
		$\Delta\Delta G$ ($kcal \cdot mol^{-1}$)	$\Delta\Delta S$ ($kcal \cdot mol^{-1} \cdot K^{-1}$	combined $\Delta\Delta G$ ($kcal \cdot mol^{-1}$)	Prediction[s]	SIFT [a]	PolyPhen-2[a]	CADD[a]	Consensus[a]
CYP2A13	D158E	-0,693498	0,012	-0,2769692					
CYP2A13	F453Y	2,59731	-0,087	1,0844178	X	X	X	X	X
CYP2A13	R101Q	-0,137141	0,678	-0,8114736	X	X	X	X	X
CYP2A13	R494C	2,68964	1,046	-0,1494568		X	X	X	X
CYP2A13	V323L	-0,218444	-0,104	0,03347128		X		X	X
CYP2A13	R257C	-2,3597	0,282227	-1,2127802	X		X		
CYP2A13	I331T	1,1793	-0,288086	0,77079032	X				
CYP2B6	A102T	0,816358	-0,142	0,46925604					
CYP2B6	A407T	0,0666466	-0,007	0,03316571					
CYP2B6	G476D	12,937	-0,312	5,2655	X	X	X	X	X
CYP2B6	G99E	17,3674	-0,964	7,679292	X	X	X	X	X
CYP2B6	H247P	1,78199	0,253	0,3937962					
CYP2B6	I328T	2,55874	0,382	0,5444812	X	X	X	X	X
CYP2B6	I382N	4,51067	0,789	0,8303746	X				
CYP2B6	I391N	2,9144	0,47	0,581072	X	X	X	X	X
CYP2B6	K139E	0,332693	0,909	-0,8916567	X	X	X	X	X
CYP2B6	K262R	-0,11128	-0,148	0,1234736					
CYP2B6	L313I	2,40625	0,154	0,741895	X		X		
CYP2B6	M198T	2,02589	0,719	-0,0354418		X		X	X
CYP2B6	M459V	1,17419	0,246	0,1706722					
CYP2B6	M46V	3,41641	0,667	0,5511958	X		X		
CYP2B6	P428T	3,86486	-0,429	1,9491268	X	X	X	X	X
CYP2B6	Q485L	-0,584528	-0,388	0,21243936		X	X	X	X
CYP2B6	R140Q	0,0819975	0,106	-0,087561					
CYP2B6	R434Q	0,466131	0,495	-0,3772702					
CYP2B6	R487S	1,58759	0,58	-0,0463158					

Table S1 continued from previous page

Gene	Mutation	Normal Mode Analysis				Computational Variant Effect Predictors			
		$\Delta\Delta G$ ($kcal \cdot mol^{-1}$)	$\Delta\Delta S$ ($kcal \cdot mol^{-1} \cdot K^{-1}$)	combined $\Delta\Delta G$ ($kcal \cdot mol^{-1}$)	Prediction[a]	SIFT[a]	PolyPhen-2[a]	CADD[a]	Consensus[a]
CYP2B6	S259R	0,15035	0,003	0,053773		X	X		X
CYP2B6	V183I	-0,432545	-0,259	0,1257129					
CYP2B6	Q172H	3,345	1,04199	0,1040712					
CYP2B6	K262R	-0,421	0,6543	-0,892796	X				
CYP2B6	P167A	2,079	0,24316	0,5176808	X	X	X	X	X
CYP2B6	R487C	0,17	0,27051	-0,2383712					
CYP2B6	R336C	2,77	0,06836	0,9760368	X	X	X	X	X
CYP2B6	V183G	4,2	-0,22363	1,8464656	X	X	X	X	X
CYP2B6	K139E	-0,102	-0,40039	0,4096768		X	X	X	X
CYP2B6	T168I	2,188	0,33496	0,4562848					
CYP2B6	I328T	1,832	0,13672	0,5430336	X	X	X	X	X
CYP2C19	I331V	0,560533	0,183	0,00804254					
CYP2C19	W212X	2,35528	1,468	-0,7491536	X				
CYP2C19	R433W	2,3425	-0,768	1,75031	X	X	X	X	X
CYP2C19	R132Q	0,269454	0,031	0,06767252		X	X	X	X
CYP2C19	W120R	4,29131	0,993	0,5185378	X	X	X	X	X
CYP2C19	R144H	1,0455	-0,147	0,56193	X	X	X	X	X
CYP2C19	P227L	1,71065	-0,072	0,730687	X	X	X	X	X
CYP2C19	R150H	0,920755	0,162	0,1684469					
CYP2C19	R410C	1,78773	0,654	-0,0531426		X	X		X
CYP2C19	R329H	-0,192877	-0,225	0,17870674					
CYP2C19	S51G	0,322177	0,504	-0,4420527					
CYP2C19	R186P	0,558345	0,932	-0,8316689	X				
CYP2C19	G91R	0,917492	-0,042	0,39568696		X	X	X	X
CYP2C19	R335Q	1,10988	0,418	-0,0464056		X	X	X	X
CYP2C19	D256N	0,136223	0	0,05176474		X		X	X

Table S1 continued from previous page

Gene	Mutation	Normal Mode Analysis				Computational Variant Effect Predictors			
		$\Delta\Delta G$ ($kcal \cdot mol^{-1}$)	$\Delta\Delta S$ ($kcal \cdot mol^{-1} \cdot K^{-1}$)	combined $\Delta\Delta G$ ($kcal \cdot mol^{-1}$)	Prediction[a]	SIFT[a]	PolyPhen-2[a]	CADD[a]	Consensus[a]
CYP2C19	V374I	0,413279	-0,364	0,56472602	X				
CYP2C19	H78Y	-1,25142	-0,377	-0,0532996					
CYP2C19	H99R	-0,520979	-0,669	0,55130798	X				
CYP2C19	D188N	0,308117	-0,063	0,18764446		X		X	X
CYP2C19	E122A	0,0390276	0,533	-0,5821295	X	X		X	X
CYP2C8	A238P	3,9232	-0,076	1,575936	X				
CYP2C8	G171S	-0,938687	-0,546	0,25481894					
CYP2C8	I223M	-0,783911	-0,159	-0,1198062					
CYP2C8	I244V	1,07585	0,253	0,125463					
CYP2C8	I264M	-0,261521	-0,03	-0,065778		X	X		X
CYP2C8	I269F	1,4958	0,134	0,418324		X	X	X	X
CYP2C8	I331T	3,35775	0,322	0,915305	X	X	X	X	X
CYP2C8	I331T	3,35775	0,322	0,915305	X	X	X	X	X
CYP2C8	K247R	0,396455	0,015	0,1338529					
CYP2C8	K383N	1,32454	0,589	-0,1563548		X	X	X	X
CYP2C8	L361F	3,84962	0,018	1,4426956	X	X			
CYP2C8	R186G	2,09486	1,053	-0,3833132		X	X	X	X
CYP2C8	V181I	-0,468388	-0,368	0,23417256					
CYP2C8	R139K	-0,5106	0,51563	-0,7715336	X		X		
CYP2C8	K399R	-0,4808	0,86914	-1,1561408	X				
CYP2C9	A149T	2,24048	-0,466	1,3733024	X	X	X	X	X
CYP2C9	A149V	2,94853	-0,344	1,5057214	X				
CYP2C9	D191G	5,14328	0,445	1,4560464	X	X	X	X	X
CYP2C9	D360E	2,0968	-0,204	1,025264	X	X			
CYP2C9	D49G	5,17648	-0,07	2,0454624	X	X		X	X
CYP2C9	E300V	-0,933367	-0,013	-0,3401195					

Table S1 continued from previous page

Gene	Mutation	Normal Mode Analysis				Computational Variant Effect Predictors			
		$\Delta\Delta G$ ($kcal \cdot mol^{-1}$)	$\Delta\Delta S$ ($kcal \cdot mol^{-1} \cdot K^{-1}$)	combined $\Delta\Delta G$ ($kcal \cdot mol^{-1}$)	Prediction[a]	SIFT[a]	PolyPhen-2[a]	CADD[a]	Consensus[a]
CYP2C9	E326D	5,00395	0,686	1,133181	X	X	X	X	X
CYP2C9	E354K	8,01916	-0,05	3,1032808	X	X	X	X	X
CYP2C9	F110S	2,36825	0,404	0,447455					
CYP2C9	G40R	31,8313	-0,963	13,174454	X				
CYP2C9	G431R	18,7053	-0,781	7,982734	X				
CYP2C9	G442V	-0,657883	-0,392	0,18904446		X		X	X
CYP2C9	G70R	10,2831	-0,593	4,571738	X	X	X	X	X
CYP2C9	G96A	0,79321	-0,21	0,5366198	X				
CYP2C9	G96R	2,21022	-1,305	2,3014836	X				
CYP2C9	G98V	8,74468	0,084	3,2288984	X	X	X	X	X
CYP2C9	H251R	-0,245571	0,055	-0,154917		X	X	X	X
CYP2C9	I207T	3,01685	0,764	0,290723		X	X		X
CYP2C9	I222V	0,584734	0,013	0,20763892					
CYP2C9	I284V	1,08465	0,347	0,023527					
CYP2C9	I327T	2,60966	0,519	0,4103908		X	X	X	X
CYP2C9	I359T	2,26965	0,228	0,607107	X	X	X	X	X
CYP2C9	I387V	0,778488	0,056	0,23310544					
CYP2C9	I434F	-0,333141	0,074	-0,2094736			X		
CYP2C9	K119R	-1,08235	-0,17	-0,220893					
CYP2C9	L128R	1,40852	0,371	0,1197176					
CYP2C9	L361I	0,783158	-0,135	0,44880004					
CYP2C9	L362V	1,54403	0,24	0,3179314					
CYP2C9	L36F	0,60469	-0,273	0,5355422	X				
CYP2C9	L90P	6,26266	0,372	1,9631708	X		X		
CYP2C9	N204H	3,42673	0,144	1,1408774	X	X			
CYP2C9	N418T	0,127028	-0,108	0,16923064					

Table S1 continued from previous page

Gene	Mutation	Normal Mode Analysis				Computational Variant Effect Predictors			
		$\Delta\Delta G$ ($kcal \cdot mol^{-1}$)	$\Delta\Delta S$ ($kcal \cdot mol^{-1} \cdot K^{-1}$)	combined $\Delta\Delta G$ ($kcal \cdot mol^{-1}$)	Prediction[a]	SIFT[a]	PolyPhen-2[a]	CADD[a]	Consensus[a]
CYP2C9	N41D	-0,430025	-0,064	-0,0917295		X	X	X	X
CYP2C9	P163L	1,8849	0,342	0,333222		X	X	X	X
CYP2C9	P227L	0,604233	-0,106	0,34832854					
CYP2C9	P227S	-0,0795527	-0,167	0,15680997		X	X	X	X
CYP2C9	P279T	0,793963	0	0,30170594		X			
CYP2C9	P317S	1,78971	0,215	0,4392898		X	X	X	X
CYP2C9	P337T	2,61775	-0,292	1,321785	X	X	X	X	X
CYP2C9	P346L	2,11666	-0,472	1,3329708	X				
CYP2C9	P382S	1,91924	-0,239	0,9969912	X	X	X	X	X
CYP2C9	Q214H	3,17459	-0,43	1,6879442	X				
CYP2C9	Q214L	-0,840573	-0,124	-0,1805377		X		X	X
CYP2C9	Q454H	2,30593	-0,245	1,1506534	X	X	X	X	X
CYP2C9	R124Q	-1,87437	0,184	-0,9183406	X	X	X	X	X
CYP2C9	R124W	-1,66312	-0,635	0,0792144		X	X	X	X
CYP2C9	R125C	0,45835	0,058	0,109213		X	X	X	X
CYP2C9	R125H	0,00512823	-0,183	0,20690873			X	X	X
CYP2C9	R132Q	0,331003	-0,011	0,13810114		X	X	X	X
CYP2C9	R132W	0,250043	-0,264	0,39069634		X	X		X
CYP2C9	R144H	2,67756	-0,096	1,1249928	X		X	X	X
CYP2C9	R150C	1,06524	0,56	-0,2224088					
CYP2C9	R150H	1,11331	0,288	0,1004978					
CYP2C9	R150L	0,130176	0,45	-0,4545331					
CYP2C9	R335Q	0,72199	0,269	-0,0269238			X	X	X
CYP2C9	R335W	0,387574	-0,006	0,15399812		X	X	X	X
CYP2C9	R433W	-0,986376	-0,53	0,21877712		X	X	X	X
CYP2C9	S280C	-0,120844	0,02	-0,0683207		X			

Table S1 continued from previous page

Gene	Mutation	Normal Mode Analysis				Computational Variant Effect Predictors			
		$\Delta\Delta G$ ($kcal \cdot mol^{-1}$)	$\Delta\Delta S$ ($kcal \cdot mol^{-1} \cdot K^{-1}$)	combined $\Delta\Delta G$ ($kcal \cdot mol^{-1}$)	Prediction[a]	SIFT[a]	PolyPhen-2[a]	CADD[a]	Consensus[a]
CYP2C9	S343R	-0,398508	0,001	-0,152553		X			
CYP2C9	T130M	2,36101	-0,066	0,9711038	X	X	X	X	X
CYP2C9	T130R	7,36524	-0,846	3,7463112	X	X	X	X	X
CYP2C9	T299A	0,177293	0,57	-0,5710287	X	X	X	X	X
CYP2C9	T299R	2,10234	-0,963	1,8774492	X	X	X	X	X
CYP2C9	V76M	1,55461	-0,396	1,0342718	X	X	X	X	X
CYP2C9	R125L	0,6178	0,69141	-0,5396152	X	X	X	X	X
CYP2C9	R144C	2,9126	0,1084	0,98538	X		X		
CYP2C9	E272G	-1,123	0,54883	-1,0414296	X	X			
CYP2C9	P489S	2,4805	0,61035	0,258998		X	X	X	X
CYP2C9	N457S	-0,2715	1,18359	-1,4287908	X				
CYP2C9	I359L	1,0875	0,70215	-0,373158					
CYP2C9	D397A	-1,8932	0,625	-1,419416	X	X	X	X	X
CYP2D6	A226V	3,16647	1,128	-0,0601014					
CYP2D6	C191L	-3,40228	-0,563	-0,6623064	X				
CYP2D6	E156A	2,22624	0,967	-0,2370688		X	X	X	X
CYP2D6	E215K	4,42362	-0,227	1,9352156	X				
CYP2D6	F457L	0,746963	0,138	0,12928594		X	X	X	X
CYP2D6	G42E	-0,39499	0	-0,1500962		X	X		X
CYP2D6	G439D	9,85274	-0,153	3,9154012	X	X	X	X	X
CYP2D6	G445R	-0,459624	-1,51	1,51654288	X	X	X	X	X
CYP2D6	K147R	0,143681	-0,219	0,29987878		X		X	X
CYP2D6	L142S	3,03872	0,743	0,3225536		X	X	X	X
CYP2D6	N175S	0,484066	0,181	-0,0187749					
CYP2D6	Q117R	1,56328	-0,11	0,7172464	X				
CYP2D6	R330P	1,87993	0,796	-0,1771466		X	X	X	X

Table S1 continued from previous page

Gene	Mutation	Normal Mode Analysis				Computational Variant Effect Predictors			
		$\Delta\Delta G$ ($kcal \cdot mol^{-1}$)	$\Delta\Delta S$ ($kcal \cdot mol^{-1} \cdot K^{-1}$)	combined $\Delta\Delta G$ ($kcal \cdot mol^{-1}$)	Prediction[a]	SIFT[a]	PolyPhen-2[a]	CADD[a]	Consensus[a]
CYP2D6	R344Q	-0,0977968	-0,305	0,30443722					
CYP2D6	R441C	0,11692	1,144	-1,2368504	X	X	X	X	X
CYP2D6	R441H	0,621678	0,025	0,20823764		X	X	X	X
CYP2D6	R497H	0,42382	-0,324	0,5239316	X	X	X	X	X
CYP2D6	S488F	-0,982676	-0,69	0,39938312					
CYP2D6	T249P	3,47729	-0,157	1,4972102	X		X		
CYP2D6	T310A	0,466573	0,683	-0,5876623	X	X	X	X	X
CYP2D6	V136M	-2,70624	-0,227	-0,7741312	X				
CYP2D6	V370I	-0,852284	-0,16	-0,1446679					
CYP2D6	P34S	4,87717	0,203	1,6259646	X	X	X	X	X
CYP2D6	S486T	-0,0484366	-0,118	0,11375409					
CYP2D6	A90V	0,147028	-0,373	0,47363064					
CYP2D6	R296C	0,843081	1,033	-0,8365892	X				
CYP2D6	N166D	0,50018	0,108	0,0691084					
CYP2D6	E156V	0,719275	-0,268	0,5734845	X	X	X	X	X
CYP2D6	F366S	2,32534	1,45	-0,7403708	X	X	X	X	X
CYP2D6	H352R	-0,435635	-0,029	-0,1330613					
CYP2D6	Y355C	1,56085	0,708	-0,199837		X	X	X	X
CYP2D6	G111S	3,77519	-0,354	1,8310522	X	X	X	X	X
CYP2D6	G169R	1,38451	-0,531	1,1208338	X	X	X	X	X
CYP2D6	D337N	-0,421284	0,142	-0,3191279					
CYP2D6	G42R	-0,42162	0	-0,1602156			X	X	X
CYP2D6	H167Q	0,355311	0,356	-0,2637018					
CYP2D6	S168A	0,390422	-0,001	0,14948036					
CYP2D6	R450H	0,0954001	0,439	-0,455428		X	X	X	X
CYP2D6	S135F	-0,73469	-0,704	0,5092978	X	X	X	X	X

Table S1 continued from previous page

Gene	Mutation	Normal Mode Analysis				Computational Variant Effect Predictors			
		$\Delta\Delta G$ ($kcal \cdot mol^{-1}$)	$\Delta\Delta S$ ($kcal \cdot mol^{-1} \cdot K^{-1}$)	combined $\Delta\Delta G$ ($kcal \cdot mol^{-1}$)	Prediction[a]	SIFT[a]	PolyPhen-2[a]	CADD[a]	Consensus[a]
CYP2D6	R365H	13,6283	0,065	5,105954	X	X	X	X	X
CYP2D6	V68G	5,26442	1,149	0,7135996	X	X	X	X	X
CYP2D6	N285S	0,893283	-0,023	0,36520754					
CYP2D6	G340R	2,70247	-0,344	1,4122186	X	X	X	X	X
CYP2D6	R414C	0,0498072	0	0,01892674			X		
CYP2D6	P430L	1,14417	-0,355	0,8323846	X		X	X	X
CYP2D6	A482G	0,215937	0,2	-0,1419439					
CYP2D6	A482T	0,690884	0,018	0,24237592					
CYP2D6	E410K	-0,232234	-0,265	0,20855108					
CYP2D6	F481V	0,605643	0,745	-0,6042557	X				
CYP2D6	G479R	5,18459	-1,612	3,7755842	X				
CYP2D6	H478P	0,0320266	0,026	-0,0169499		X	X		X
CYP2D6	H478Y	0,14391	-0,139	0,2103658					
CYP2D6	P469A	2,17337	0,214	0,5862006	X	X	X		X
CYP2D6	T107I	-0,668954	-0,207	-0,0223625					
CYP2D6	T470A	-0,291731	0	-0,1108578					
CYP2D6	Q151E	0,773877	0,008	0,28511326					
CYP2D6	V136I	-1,52252	-0,016	-0,5606376	X				
CYP2D6	V338M	0,287774	0,268	-0,1908059		X	X	X	X
CYP2D6	R440H	0,262725	-0,277	0,4100755			X	X	X
CYP2D6	R201H	0,00359482	-0,087	0,09880603					
CYP2D6	E155K	-0,302358	-0,088	-0,016336					
CYP2D6	F120I	1,52379	0,3	0,2430402					
CYP2D6	E334A	1,28008	0,643	-0,2337296		X	X	X	X
CYP2D6	E418K	0,188964	-0,006	0,07852632			X	X	X
CYP2D6	A122S	0,204611	-0,118	0,20991218		X	X		X

Table S1 continued from previous page

Gene	Mutation	Normal Mode Analysis				Computational Variant Effect Predictors			
		$\Delta\Delta G$ ($kcal \cdot mol^{-1}$)	$\Delta\Delta S$ ($kcal \cdot mol^{-1} \cdot K^{-1}$)	combined $\Delta\Delta G$ ($kcal \cdot mol^{-1}$)	Prediction[a]	SIFT[a]	PolyPhen-2[a]	CADD[a]	Consensus[a]
CYP2D6	T261I	0,598285	-0,079	0,3158283		X			
CYP2D6	K404Q	0,66692	0,379	-0,1710504		X	X	X	X
CYP2D6	R62W	0,932277	0,593	-0,3098947		X	X	X	X
CYP2D6	E383K	-0,373296	0,042	-0,1888925		X			
CYP2D6	V104M	-0,981051	-0,413	0,08976062					
CYP2D6	H94R	-0,306898	0,144	-0,2779012					
CYP2D6	I109V	0,377463	0,174	-0,0514441					
CYP2D6	L91M	-0,612478	0,234	-0,4948216		X	X		X
CYP2D6	T107N	0,153259	0,069	-0,0190416					
CYP2D6	T107S	0,808942	0,084	0,21331796					
CYP2D6	V104A	0,668115	0,209	0,0198037					
CYP2D6	P267H	0,274668	-0,152	0,27461384		X	X	X	X
CYP2D6	H478Q	-0,231298	0,109	-0,2099732					
CYP2D6	E278K	-0,552539	0,101	-0,3230848		X			
CYP2D6	M279K	1,61274	-0,023	0,6386012	X	X		X	X
CYP2D6	D337G	0,135803	0,231	-0,2071149		X		X	X
CYP2D6	R388H	2,35542	0,046	0,8435396	X				
CYP2D6	H463D	1,09752	0,672	-0,3355824		X		X	X
CYP2D6	R88P	4,64978	0,242	1,4958764	X	X	X	X	X
CYP2D6	H324P	5,18382	0,201	1,7447316	X	X	X	X	X
CYP2F1	D218N	-1,682791	0,65918	-1,3777422	X				
CYP2F1	Q266H	1,01377	0,99023	-0,723825	X	X			
CYP2F1	P490L	2,21995	0,66992	0,0932706		X			
CYP2F1	S38P	-0,281256	0,84766	-1,0562565	X				
CYP2W1	V432I	-0,498	0,969727	-1,2753342	X				
CYP2W1	Q482H	-0,929	1,722656	-2,2823947	X				

Table S1 continued from previous page

Gene	Mutation	Normal Mode Analysis				Computational Variant Effect Predictors			
		$\Delta\Delta G$ ($kcal \cdot mol^{-1}$)	$\Delta\Delta S$ ($kcal \cdot mol^{-1} \cdot K^{-1}$)	combined $\Delta\Delta G$ ($kcal \cdot mol^{-1}$)	Prediction[a]	SIFT[a]	PolyPhen-2[a]	CADD[a]	Consensus[a]
CYP2W1	P488L	0,950687	-0,232	0,62110106	X				
CYP2W1	E58A	-0,104029	0,324	-0,402411		X	X	X	X
CYP2W1	A181T	12,9333	-0,68	5,676254	X				
CYP2W1	A125T	1,70379	-0,351	1,0405602	X				
CYP3A4	A370S	0,45083	-0,192	0,3863554					
CYP3A4	D174H	0,228664	-0,376	0,50801232	X				
CYP3A4	F113I	2,101	0,515	0,22158					
CYP3A4	F189S	6,26981	1,37	0,8481278	X	X	X	X	X
CYP3A4	G56D	2,04455	-0,068	0,853089	X	X	X	X	X
CYP3A4	H324Q	1,32278	0,471	-0,0248636		X	X	X	X
CYP3A4	I118V	0,572509	0,661	-0,5227666	X				
CYP3A4	I335T	1,87632	0,521	0,1294816		X	X	X	X
CYP3A4	I427V	0,642671	0,138	0,08965498		X			
CYP3A4	L293P	0,986763	0,698	-0,4067901			X		
CYP3A4	L373F	5,22524	-0,436	2,4739112	X	X			
CYP3A4	P218R	1,14656	-0,24	0,7044928	X	X	X	X	X
CYP3A4	P416L	6,31746	-0,642	3,1196748	X	X	X	X	X
CYP3A4	P467S	2,69737	0,187	0,8155606	X	X			
CYP3A4	Q200H	0,424505	0	0,1613119					
CYP3A4	R130Q	-0,231126	0,711	-0,8841479	X	X	X	X	X
CYP3A4	R162Q	0,332387	0,288	-0,1962529					
CYP3A4	R162W	0,896397	-0,231	0,59935086	X	X	X	X	X
CYP3A4	S222P	3,42096	0,04	1,2551648	X				
CYP3A4	T185S	1,74183	0,187	0,4524554		X	X	X	X
CYP3A4	V170I	-0,16228	-0,188	0,1488936					
CYP3A4	Y319C	4,23306	1,008	0,4796028		X	X	X	X

Table S1 continued from previous page

Gene	Mutation	Normal Mode Analysis				Computational Variant Effect Predictors			
		$\Delta\Delta G$ $(kcal \cdot mol^{-1})$	$\Delta\Delta S$ $(kcal \cdot mol^{-1} \cdot K^{-1})$	combined $\Delta\Delta G$ $(kcal \cdot mol^{-1})$	Prediction[a]	SIFT[a]	PolyPhen-2[a]	CADD[a]	Consensus[a]
CYP3A4	T363M	0,600548	-0,693	1,00436824	X	X	X		X
CYP3A4	M445T	1,5203	0,404	0,125234		X	X	X	X

[a]'X' and blanks represent variants predicted to be deleterious and neutral, respectively, by the prediction tools NMA: Normal Mode Analysis.

This table presents all 261 missense variants analysed in this study, together with their FoldX and ENCoM-derived energy scores. A combined $\Delta\Delta G$ (combination of ENCoM and FoldX) is presented, from which a prediction of the effect of each variant is made based on the threshold (-0.5 < combined $\Delta\Delta G$ < 0.5) (Frappier & Najmanovich 2014). Predictions from commonly used Ensembl Variant Effect Predictor (VEP; https://www.ensembl.org/ToolsVEP) (McLaren et al. 2016) plugins as well as a consensus ('2 from 3' rule) has been included for comparison with combined $\Delta\Delta G$ predictions.

Table S2: CYP star allele/haplotype NMA analysis.

Star Allele/ Haplotype	Haplotype Combined ΔΔG (kcal·mol^{-1})	Predicted haplotype structural effect	CPIC functional classification	Variants defining haplotype	Variant Combined ΔΔG (kcal·mol^{-1})	Predicted variant functional effect
				D218N	-1,3777422	stabilising
				Q266H	-0,723825	stabilising
*CYP2F1*3.001*	-0,9262366	stabilising		P490L	0,0932706	neutral
				S38P	-1,0562565	stabilising
*CYP2F1*4.001*	-1,9834974	stabilising		D218N	-1,3777422	stabilising
				V432I	-1,2753342	stabilising
*CYP2W1*5.001*	-2,5236142	stabilising		Q482H	-2,2823947	stabilising
				R257C	-1,2127802	stabilising
*CYP2A13*10*	0,78203736	destabilising	function not assigned	I331T	0,7707903	destabilising
				Q172H	0,1040712	neutral
*CYP2B6*6*	-0,0601104	neutral	decreased function	K262R	-0,892796	stabilising
				K262R	-0,892796	stabilising
				Q172H	0,1040712	neutral
*CYP2B6*26*	0,6662888	destabilising	decreased function	P167A	0,5176808	destabilising
				Q172H	0,1040712	neutral
				K262R	-0,892796	stabilising
*CYP2B6*7*	-0,0601104	neutral	decreased function	R487C	-0,2383712	neutral
				Q172H	0,1040712	neutral
				K262R	-0,892796	stabilising
*CYP2B6*19*	1,541516	destabilising	decreased function	R336C	0,9760368	destabilising
				Q172H	0,1040712	neutral
				K262R	-0,892796	stabilising
*CYP2B6*37*	2,4683328	destabilising	no function	V183G	1,8464656	destabilising
				Q172H	0,1040712	neutral
				K262R	-0,892796	stabilising
*CYP2B6*13*	0,680704	destabilising	no function	K139E	0,4096768	neutral
				K262R	-0,892796	stabilising
				Q172H	0,1040712	neutral
*CYP2B6*20*	0,480356	neutral	decreased function	T168I	0,4562848	neutral
				K262R	-0,892796	stabilising
*CYP2B6*18.002*	0,3749488	neutral	no function	I328T	0,5430336	destabilising
				R139K	-0,7715336	stabilising
*CYP2C8*3*	-1,1189533	stabilising	function not assigned	K399R	-1,1561408	stabilising
				R125L	-0,5396152	stabilising
*CYP2C9*35*	1,01259728	destabilising	no function	R144C	0,98538	destabilising
				E272G	-1,0414296	stabilising
*CYP2C9*71*	-0,098842	neutral	uncertain function	P489S	0,258998	neutral
*CYP2C9*61*	0,23829624	neutral		R144C	0,98538	destabilising

			decreased function	N457S	-1,4287908	stabilising
*CYP2C9*18*	-1,735581	stabilising	uncertain function	I359L	-0,373158	neutral
				D397A	-1,419416	stabilising
*CYP2D6*70*	0,5700364	destabilising	uncertain function	V338M	-3,0965856	stabilising
				V136I	-3,8637948	stabilising
				V119M	-2,7709664	stabilising
				S486T	0,1137541	neutral
*CYP2D6*10*	1,7397212	destabilising	decreased function	P34S	1,6259646	destabilising
				S486T	0,1137541	neutral
*CYP2D6*102*	-0,2623378	neutral	unknown function	A90V	0,4736306	neutral
				R296C	-0,8365892	stabilising
				S486T	0,1137541	neutral
*CYP2D6*103*	-0,1932294	neutral	unknown function	A90V	0,4736306	neutral
				N166D	0,0691084	neutral
				R296C	-0,8365892	stabilising
				S486T	0,1137541	neutral
*CYP2D6*104*	-0,1539084	neutral	unknown function	EA156V	0,5734845	destabilising
				R296C	-0,8365892	stabilising
				S486T	0,1137541	neutral
*CYP2D6*105*	-1,441462	stabilising	unknown function	F366S	-0,7403708	stabilising
				R296C	-0,8365892	stabilising
				S486T	0,1137541	neutral
*CYP2D6*108*	-0,3916966	neutral	unknown function	H352R	-0,1330613	neutral
				Y355C	-0,199837	neutral
*CYP2D6*111*	1,1128658	destabilising	uncertain function	G111S	1,8310522	destabilising
				R296C	-0,8365892	stabilising
				S486T	0,1137541	neutral
*CYP2D6*114*	2,09231	destabilising	no function	G169R	1,1208338	destabilising
				P34S	1,6259646	destabilising
				R296C	-0,8365892	stabilising
				S486T	0,1137541	neutral
*CYP2D6*117*	-1,0210867	stabilising	unknown function	D337N	-0,3191279	neutral
				R296C	-0,8365892	stabilising
				S486T	0,1137541	neutral
*CYP2D6*12*	-0,8583694	stabilising	no function	G42R	-0,1602156	neutral
				R296C	-0,8365892	stabilising
				S486T	0,1137541	neutral
*CYP2D6*121*	-0,665719	stabilising	unknown function	H167Q	-0,2637018	neutral
				N166D	0,0691084	neutral
				R296C	-0,8365892	stabilising
				S168A	0,1494804	neutral
				S486T	0,1137541	neutral
*CYP2D6*125*	-1,1629252	stabilising	unknown function	R296C	-0,8365892	stabilising
				R450H	-0,455428	neutral
				S486T	0,1137541	neutral
*CYP2D6*126*	-0,2621316	neutral	unknown function	R296C	-0,8365892	stabilising
				S135F	0,5092978	destabilising

				S486T	0,1137541	neutral
			unknown	R365H	5,105954	destabilising
*CYP2D6*127*	3,09417	destabilising	function	Y355C	-0,199837	neutral
				R296C	-0,8365892	stabilising
			uncertain	S486T	0,1137541	neutral
*CYP2D6*128*	0,0112326	neutral	function	V68G	0,7135996	destabilising
				N285S	0,3652075	neutral
			decreased	P34S	1,6259646	destabilising
*CYP2D6*132*	2,1189126	destabilising	function	S486T	0,1137541	neutral
				G340R	1,4122186	destabilising
			uncertain	R296C	-0,8365892	stabilising
*CYP2D6*133*	0,8011668	destabilising	function	S486T	0,1137541	neutral
				R296C	-0,8365892	stabilising
			uncertain	RA414C	0,0189267	neutral
*CYP2D6*135*	-0,6727194	stabilising	function	S486T	0,1137541	neutral
				P430L	0,8323846	destabilising
			uncertain	R296C	-0,8365892	stabilising
*CYP2D6*136*	0,1354938	neutral	function	S486T	0,1137541	neutral
				G169R	1,1208338	destabilising
			decreased	R296C	-0,8365892	stabilising
*CYP2D6*14*	0,4613912	neutral	function	S486T	0,1137541	neutral
				A482G	-0,1419439	neutral
				A482T	0,2423759	neutral
				E410K	0,2085511	neutral
				F481V	-0,6042557	stabilising
				G479R	3,7755842	destabilising
				H478P	-0,0169499	neutral
				H478Y	0,2103658	neutral
				P469A	0,5862006	destabilising
				R296C	-0,8365892	stabilising
				S486T	0,1137541	neutral
			uncertain	T107I	-0,0223625	neutral
*CYP2D6*141*	2,34104	destabilising	function	T470A	-0,1108578	neutral
			uncertain	P34S	1,6259646	destabilising
*CYP2D6*142*	1,7397212	destabilising	function	S486T	0,1137541	neutral
				R296C	-0,8365892	stabilising
			decreased	S486T	0,1137541	neutral
*CYP2D6*17*	-0,6321365	stabilising	function	T107I	-0,0223625	neutral
			normal	R296C	-0,8365892	stabilising
*CYP2D6*2*	-0,7025097	stabilising	function	S486T	0,1137541	neutral
				Q151E	0,2851133	neutral
			uncertain	R296C	-0,8365892	stabilising
*CYP2D6*28*	-0,4291658	neutral	function	S486T	0,1137541	neutral
				R296C	-0,8365892	stabilising
				S486T	0,1137541	neutral
			decreased	V136I	-0,5606376	stabilising
*CYP2D6*29*	-1,4660549	stabilising	function	V338M	-0,1908059	neutral
				R296C	-0,8365892	stabilising
				R440H	0,4100755	neutral
*CYP2D6*31*	-0,2989764	neutral	no function	S486T	0,1137541	neutral
*CYP2D6*35*	-0,7025097	stabilising		R296C	-0,8365892	stabilising

			normal function	S486T	0,1137541	neutral
*CYP2D6*36*	4,77919	destabilising	no function	A482G	-0,1419439	neutral
				A482T	0,2423759	neutral
				F481V	-0,6042557	stabilising
				G479R	3,7755842	destabilising
				H478P	-0,0169499	neutral
				H478Y	0,2103658	neutral
				P34S	1,6259646	destabilising
				P469A	0,5862006	destabilising
				S486T	0,1137541	neutral
				T470A	-0,1108578	neutral
*CYP2D6*37*	1,8385254	destabilising	uncertain function	P34S	1,6259646	destabilising
				R201H	0,098806	neutral
				S486T	0,1137541	neutral
*CYP2D6*45*	-0,7198924	stabilising	normal function	E155K	-0,016336	neutral
				R296C	-0,8365892	stabilising
				S486T	0,1137541	neutral
*CYP2D6*46*	-0,7198924	stabilising	normal function	E155K	-0,016336	neutral
				R296C	-0,8365892	stabilising
				S486T	0,1137541	neutral
*CYP2D6*47*	1,7397212	destabilising	no function	P34S	1,6259646	destabilising
				S486T	0,1137541	neutral
*CYP2D6*49*	1,9982854	destabilising	decreased function	F120I	0,2430402	neutral
				P34S	1,6259646	destabilising
				S486T	0,1137541	neutral
*CYP2D6*51*	-0,8386966	stabilising	no function	E334A	-0,2337296	neutral
				R296C	-0,8365892	stabilising
				S486T	0,1137541	neutral
*CYP2D6*52*	1,818246	destabilising	decreased function	E418K	0,0785263	neutral
				P34S	1,6259646	destabilising
				S486T	0,1137541	neutral
*CYP2D6*53*	0,451982	neutral	normal function	A122S	0,2099122	neutral
				F120I	0,2430402	neutral
*CYP2D6*54*	2,153272	destabilising	decreased function	P34S	1,6259646	destabilising
				S486T	0,1137541	neutral
				T261I	0,3158283	neutral
*CYP2D6*55*	-0,873559	stabilising	decreased function	K404Q	-0,1710504	neutral
				R296C	-0,8365892	stabilising
				S486T	0,1137541	neutral
*CYP2D6*57*	4,268034	destabilising	no function	A482G	-0,1419439	neutral
				A482T	0,2423759	neutral
				F481V	-0,6042557	stabilising
				G479R	3,7755842	destabilising
				H478P	-0,0169499	neutral
				H478Y	0,2103658	neutral
				P34S	1,6259646	destabilising
				P469A	0,5862006	destabilising
				R62W	-0,3098947	neutral
				S486T	0,1137541	neutral
				T470A	-0,1108578	neutral

*CYP2D6*64*	1,715655	destabilising	uncertain function	P34S	1,6259646	destabilising
				S486T	0,1137541	neutral
				T107I	-0,0223625	neutral
*CYP2D6*65*	0,9272902	destabilising	uncertain function	P34S	1,6259646	destabilising
				R296C	-0,8365892	stabilising
				S486T	0,1137541	neutral
*CYP2D6*72*	1,4516606	destabilising	uncertain function	E383K	-0,1888925	neutral
				P34S	1,6259646	destabilising
				S486T	0,1137541	neutral
*CYP2D6*73*	-0,5471407	stabilising	unknown function	R296C	-0,8365892	stabilising
				S486T	0,1137541	neutral
				V104M	0,0897606	neutral
*CYP2D6*82*	-0,7269554	stabilising	unknown function	H94R	-0,2779012	neutral
				I109V	-0,0514441	neutral
				L91M	-0,4948216	neutral
				T107N	-0,0190416	neutral
				T107S	0,213318	neutral
				V104A	0,0198037	neutral
*CYP2D6*83*	3,1973284	destabilising	uncertain function	A482G	-0,1419439	neutral
				A482T	0,2423759	neutral
				F481V	-0,6042557	stabilising
				G479R	3,7755842	destabilising
				H478P	-0,0169499	neutral
				H478Y	0,2103658	neutral
				P469A	0,5862006	destabilising
				S486T	0,1137541	neutral
				T470A	-0,1108578	neutral
*CYP2D6*84*	-0,430364	neutral	uncertain function	P267H	0,2746138	neutral
				R296C	-0,8365892	stabilising
				S486T	0,1137541	neutral
*CYP2D6*85*	-0,8875665	stabilising	unknown function	H478Q	-0,2099732	neutral
				R296C	-0,8365892	stabilising
				S486T	0,1137541	neutral
*CYP2D6*86*	0,5302152	destabilising	unknown function	E278K	-0,3230848	neutral
				M279K	0,6386012	destabilising
*CYP2D6*87*	1,7397212	destabilising	uncertain function	P34S	1,6259646	destabilising
				S486T	0,1137541	neutral
*CYP2D6*88*	0,13216822	neutral	uncertain function	S486T	0,1137541	neutral
				V104A	0,0198037	neutral
*CYP2D6*94*	1,4972556	destabilising	uncertain function	D337G	-0,2071149	neutral
				P34S	1,6259646	destabilising
				S486T	0,1137541	neutral
*CYP2D6*95*	2,600268	destabilising	uncertain function	P34S	1,6259646	destabilising
				R388H	0,8435396	destabilising
				S486T	0,1137541	neutral
*CYP2D6*98*	-1,0368608	stabilising	uncertain function	H463D	-0,3355824	neutral
				R296C	-0,8365892	stabilising
				S486T	0,1137541	neutral
*CYP2D6*99*	3,266807	destabilising	no function	P34S	1,6259646	destabilising
				R88P	1,4958764	destabilising
				S486T	0,1137541	neutral

				I331V	0,0080425	neutral
*CYP2C19*3*	0,8806458	destabilising	no function	W212X	-0,7491536	stabilising
				I331V	0,0080425	neutral
*CYP2C19*5*	1,7583514	destabilising	no function	R433W	1,75031	destabilising
				I331V	0,0080425	neutral
*CYP2C19*6*	0,07683506	neutral	no function	R132Q	0,0676725	neutral
				I331V	0,0080425	neutral
*CYP2C19*8*	0,5296734	destabilising	no function	W120R	0,5185378	destabilising
			decreased	I331V	0,0080425	neutral
*CYP2C19*9*	0,5719456	destabilising	function	R144H	0,56193	destabilising
			decreased	I331V	0,0080425	neutral
*CYP2C19*10*	0,7448702	destabilising	function	P227V	0,730687	destabilising
			normal	I331V	0,0080425	neutral
*CYP2C19*11*	0,405794	neutral	function	R150H	0,1684469	neutral
			normal	I331V	0,0080425	neutral
*CYP2C19*13*	-0,043127	neutral	function	R410C	-0,0531426	neutral
			normal	I331V	0,0080425	neutral
*CYP2C19*18*	0,4445719	neutral	function	R329H	0,1787067	neutral
			decreased	I331V	0,0080425	neutral
*CYP2C19*19*	-0,4329313	neutral	function	S51G	-0,4420527	neutral
				I331V	0,0080425	neutral
*CYP2C19*22*	-0,8205922	stabilising	no function	R186P	-0,8316689	stabilising
			uncertain	I331V	0,0080425	neutral
*CYP2C19*23*	0,4155246	neutral	function	G91R	0,395687	neutral
				I331V	0,0080425	neutral
*CYP2C19*24*	-0,020874	neutral	no function	R335Q	-0,0464056	neutral
			decreased	I331V	0,0080425	neutral
*CYP2C19*26*	0,07683506	neutral	function	D256N	0,0517647	neutral
			normal	I331V	0,0080425	neutral
*CYP2C19*28*	0,47827198	neutral	function	V374I	0,564726	destabilising
			uncertain	I331V	0,0080425	neutral
*CYP2C19*31*	-0,0452559	neutral	function	H78Y	-0,0532996	neutral
			uncertain	I331V	0,0080425	neutral
*CYP2C19*32*	0,56244312	destabilising	function	H99R	0,551308	destabilising
			uncertain	I331V	0,0080425	neutral
*CYP2C19*33*	0,19557604	neutral	function	D188N	0,1876445	neutral
			function not	I331V	0,0080425	neutral
*CYP2C19*39*	-0,5709945	stabilising	assigned	E122A	-0,5821295	stabilising
			function not	T363M	1,0043682	destabilising
*CYP3A4*38*	1,0478738	destabilising	assigned	M445T	0,125234	neutral

NMA: Normal Mode Analysis

This table presents 92 CYP star alleles together with the defining missense variants for each star allele. Moreover, a CPIC functional classification is included together with the NMA-based combined ΔΔG analysis.

Table S3: SWAAT analysis of CYP missense variant structural consequence.

Gene	**Mutation**	**SWAAT prediction[a]**	**SWAAT red-flag protein structural consequences**
CYP2A13	I331T	X	
CYP2A13	R257C	X	exposed hydrophilic introduced, salt bridge breakage
CYP2B6	P167A	X	
CYP2B6	K139E	X	Salt bridge formation
CYP2B6	G110V	X	
CYP2B6	R336C	X	exposed hydrophilic introduced, Salt bridge breakage
CYP2B6	R487C		exposed hydrophilic introduced
CYP2B6	Q172H	X	buried charge introduced, Salt bridge formation
CYP2B6	I328T	X	
CYP2B6	M198T	X	
CYP2B6	I114T	X	
CYP2B6	R29S		
CYP2B6	R487S	X	
CYP2B6	V183G	X	
CYP2B6	T168I	X	buried exposed switch
CYP2B6	R29P	X	
CYP2B6	K262R		
CYP2B6	E148D	X	
CYP2B6	A279P		
CYP2C19	E122A		exposed hydrophilic introduced
CYP2C19	V374I	X	buried exposed switch
CYP2D6	R62W	X	exposed hydrophilic introduced, Salt bridge breakage
CYP2D6	H167Q	X	
CYP2D6	N166D		
CYP2D6	T107N	X	buried exposed switch
CYP2D6	P469A	X	buried exposed switch
CYP2D6	T107S	X	buried exposed switch
CYP2D6	L460V	X	
CYP2D6	R129L		exposed hydrophilic introduced
CYP2D6	D337G		
CYP2D6	E410K		
CYP2D6	V119M	X	
CYP2D6	A90V	X	
CYP2D6	E418Q		salt bridge breakage
CYP2D6	E383K		salt bridge breakage
CYP2D6	H352R		
CYP2D6	P325L	X	large helical penality in alpha helix
CYP2D6	T107I	X	
CYP2D6	H478Y		
CYP2D6	T470A		
CYP2D6	A122S	X	
CYP2D6	R388H	X	salt bridge formation
CYP2D6	R201H		salt bridge breakage
CYP2D6	E418K	X	salt bridge breakage

Table S3 continued from previous page

Gene	Mutation	SWAAT prediction[a]	SWAAT red-flag protein structural consequences
CYP2D6	L91M	X	
CYP2D6	G479R	X	
CYP2D6	P268S	X	
CYP2D6	P34S	X	
CYP2D6	V136I		
CYP2D6	R365H	X	buried salt bridge breakage, salt bridge breakage
CYP2D6	V338M	X	
CYP2D6	I109V	X	
CYP2D6	G42R		
CYP2D6	V104A	X	
CYP2D6	R269Q	X	salt bridge breakage
CYP2D6	A482G		
CYP2D6	T261I	X	
CYP2D6	E278K		
CYP2D6	E155K		
CYP2D6	A482T	X	
CYP2D6	R173C		exposed hydrophilic introduced
CYP2D6	A449D	X	
CYP2D6	Q151E	X	
CYP2D6	Y355C	X	
CYP2D6	H94R	X	
CYP2D6	F120I	X	
CYP2D6	R88P	X	
CYP2D6	S168A		
CYP2D6	N285S		
CYP2D6	F481V	X	
CYP2D6	M279K	X	salt bridge formation
CYP2D6	H478P		
CYP3A4	I334T	X	
CYP3A4	A369S	X	
CYP3A4	I118V	X	
CYP3A4	G56D	X	buried glycine replaced, buried charge introduced
CYP3A4	L372F	X	
CYP3A4	R162W	X	exposed hydrophilic introduced, salt bridge breakage
CYP3A4	P218R	X	large helical penalty in alpha helix
CYP3A4	S222P	X	
CYP3A4	P415L	X	
CYP3A4	Y318C	X	
CYP3A4	R162Q		salt bridge breakage
CYP3A4	M444T	X	
CYP3A4	L292P	X	
CYP3A4	H323Q	X	salt bridge breakage
CYP3A4	I426V	X	
CYP3A4	F113I	X	
CYP3A4	T185S	X	

Table S3 continued from previous page

Gene	Mutation	SWAAT prediction[a]	SWAAT red-flag protein structural consequences
CYP3A4	F189S	X	buried exposed switch
CYP3A4	R130Q	X	
CYP3A4	V170I	X	
CYP3A4	T362M	X	
CYP3A4	P466S	X	
CYP3A4	Q200H		
CYP2C8	K399R		
CYP2C8	R139K		
CYP2C9	E272G	X	salt bridge breakage
CYP2C9	P489S	X	
CYP2C9	R144C	X	buried charge replaced, buried exposed switch
CYP2C9	N457S		
CYP2C9	I359L	X	
CYP2C9	R125L		exposed hydrophilic introduced
CYP2C9	D397A	X	exposed hydrophilic introduced, salt bridge breakage
CYP2F1	Q266H	X	buried charge introduced
CYP2F1	D218N	X	buried exposed switch
CYP2F1	P490L	X	
CYP2F1	S38P		
CYP2J2	D342N		salt bridge breakage
CYP2J2	G312R	X	buried exposed switch, large helical penality in alpha helix
CYP2J2	T143A		
CYP2J2	N404Y	X	
CYP2J2	P351L	X	
CYP2J2	P115L	X	buried exposed switch, large helical penalty in alpha helix
CYP2J2	I192N	X	buried hydrophilic introduced
CYP2J2	R158C	X	exposed hydrophilic introduced, salt bridge breakage
CYP2R1	L99P	X	buried proline introduced
CYP2S1	P466L	X	
CYP2S1	L230R	X	
CYP2S1	S61N	X	
CYP2S1	R380C		exposed hydrophilic introduced
CYP2W1	V432I		
CYP2W1	Q482H		
CYP3A5	F446S	X	buried exposed switch
CYP3A5	L82R	X	buried hydrophilic introduced, buried charge introduced
CYP3A5	I456V	X	
CYP3A5	H30Y	X	
CYP3A5	Q200R		
CYP3A5	I488T	X	
CYP3A5	T398N	X	
CYP3A5	F203C	X	
CYP3A5	A337T		
CYP3A5	Y53C	X	
CYP3A5	R28C		exposed hydrophilic introduced

[a] 'X' and blanks represent variants predicted to be deleterious and neutral by SWAAT, respectively.

The likely structural events are based on a set of features that were assessed for structural impact with the likelihood to influence protein function (Ittisoponpisan et al. 2019).

Chapter 4

Discussion and conclusion

Structural bioinformatics has emerged as a viable technique for predicting the impact of genetic variation on protein structure and function, revolutionising computational variant effect prediction. The knowledge generated by these techniques may be utilized to create novel pharmaceuticals, develop individualised therapy, and be used to learn more about the molecular-level interactions between genetic variations and drug response. Furthermore, computational variant impact prediction may be used to predict the effects of genetic variants on protein-protein and protein-ligand interactions.

This study aimed to evaluate the impact of star allele missense variant combinations on protein stability and conformational dynamics while also highlighting the utility of structural bioinformatics/multiscale modelling in this context. This final chapter consolidates the findings of the study, explores the emerging themes from this study, discusses the contextual relevance of this work, emphasises the study's strengths and limitations, outlines future directions, and sums the conclusions.

4.1 Consolidated findings

Objective 1:

To collect data regarding known CYP missense variants and star alleles catalogued in the PharmVar database.

Main findings:

- Majority of the star alleles catalogued in PharmVar are yet to be assigned CPIC clinical functional annotations.
- Rare star alleles are least likely to have CPIC clinical function annotations.

Objective 2:

To assess the effects of individual and missense-variant combinations defined by star alleles on protein stability and conformational dynamics.

Main findings:

- Differences in the change in Gibbs free energy ($\Delta\Delta G$) between the mutant and the wild-type protein have the tendency to predict CYP missense variants and star alleles to have significantly aberrant effects on the proteins' thermodynamic stability.
- Differences in the change in vibrational entropy ($\Delta\Delta S_{\text{vib}}$) predicted more variants to have neutral effects on protein vibrational entropy; however, many star alleles were predicted to increase the proteins' vibrational entropy.
- The recommended combination $\Delta\Delta G$ and $\Delta\Delta S_{\text{vib}}$ into a "combined $\Delta\Delta G$" energy term modulated the effects of the two energies and more variants (56%) were predicted to have neutral effects on protein stability.

Objective 3:

To compare our *in silico* derived results with other commonly used variant interpretation tools and results from experimental procedures.

Main findings:

- We observed 80% concordance of our 'combined $\Delta\Delta G$' results with the consensus results from commonly used variants effect predictors (SIFT, PolyPhen-2 and CADD).

- Deep Mutational Scanning (DMS) results were only available for only CYP2C9 and CYP2C19 (and only 71 variants from our study). 66% concordance was observed for our 'combined $\Delta\Delta G$' and DMS functional assay results.
- Inconsistencies were noted when our Normal Mode Analysis and 'combined $\Delta\Delta G$' were compared to known CPIC clinical functional impacts.

Objective 4:

To perform multi-scale molecular dynamics simulation to further analyse variant impact on protein structural properties.

Main findings:

- Due to time constraints associated with MD simulation, only two star alleles were simulated – *CYP2D6*2* (S486T and R296C) and *CYP2D6*17* (**2* variants + T107I).
- Most deviations between the wild-type and the mutant structures are concentrated in the F-G helix region. Active site residues did not confer any structural divergences from the wild-type.
- R296C significantly perturbs the F-G loop region; however, when in CYP2D6*2, the combination of R296C and S486T results in a normal function star allele.
- In the case of CYP2D6*17 (composed of variants T107I, S486T, and R296C), T107I alone is not enough to explain the function exhibited by the CYP2D6*17. We observed a modulation between T107I and R296C.

4.2 Emerging themes

Two main themes emerge from this study. Firstly, the utility of structural bioinformatics in missense variant functional impact analysis, particularly in pharmacogenomics, particularly pharmacogenomics analysis. Secondly, the mechanism of variant interactions in CYP star alleles. The specific results that lead to the emerging themes have been discussed in the published article and re-emphasised in the following subsections.

4.2.1 Utility of structural bioinformatics in variant analysis

The five main benefits are detailed below:

1. **High-throughput analysis:**

   A total of 91 CYP star alleles and 261 variants from 9 CYP genes were analysed. Structural bioinformatics and computational efficiency allow for the processing of large amounts of data simultaneously, significantly accelerating data processing to perform high-throughput analysis. This in turn facilitates the comprehensive understanding of dysfunctions in biological systems.

2. **Time-effectiveness:**

   Unlike conventional wet-laboratory procedures which often require sample preparation, time-consuming experimental procedures, and data collection, computational methods provide rapid results utilising resources on high-performance computing clusters and also implement scalable features e.g. parallelisation (Peña-Varas et al. 2022). Furthermore, conventional wet-laboratory experiments often require specialised equipment, samples, reagents and personnel which is often expensive and not easily accessible (Koeppe et al. 2020, Peña-Varas et al. 2022), structural bioinformatics approaches often utilise relatively affordable computational resources; eliminating the cost of samples, reagent and making use of fewer specialised equipment. However, complex biological systems may present limited accuracy of prediction. This limitation may be addressed by integrated and collaborative approaches to draw consensus predictions and generate robust insights into the mechanisms of functional impact resulting from CYP missense variants.

3. **Reproducibility:**

   Structural bioinformatics fosters transparency within the scientific community. Procedures can be shared in the form of automated programs which are able to replicate the environment in which analysis is carried out, thus eliminating external variables that may lead to inconsistencies in the results. This in turn may facilitate collaborations between scientists which accelerates research progress and foster innovation.

4. **Hypothesis generation:**

   Results from structural bioinformatics studies can guide experimental designs for

wet-laboratory-based studies by prioritising potential targets from a large pool of targets. This is aided by the time-effectiveness of structural bioinformatics approaches. We are able to use our findings and prioritise missense variants that are likely to affect structural functioning thus fast-tracking the research process while saving on resource costs. We observed high concordance between our results and other variant effect predictors (VEPs) and deep mutational scanning functional assays (e.g., *CYP2D6*10* predictions from this study correlated with predictions from commonly used VEPs). Moreover, We note that different methods assay different aspects relating to protein function and therefore it is apt to have some inconsistencies. Furthermore, there are various external factors that may affect protein function which unfortunately are challenging to model in computational systems. Nonetheless, structural bioinformatics informs about a crucial aspect of protein function and therefore cautions about potential aberrations in function. This analysis may be extended to include other functional assays for a holistic view of the effects of missense variation on protein function.

These advantages highlight the utility of structural bioinformatics approaches in identifying potentially important missense variations in CYP proteins from which novel therapeutic strategies may be designed to improve healthcare.

4.2.2 Mechanisms of variant interaction in CYP star alleles

Two main scenarios are presented by our Normal Mode Analysis results: (i) The energy exhibited by the star allele may be accounted for by a single missense variant; (ii) The energy of the star allele presents as the sum of energies of individual missense variants that make up the star allele. The latter is further exemplified by our molecular dynamics simulation results.

Studying the way in which combinations of variants affect CYP protein structure is crucial for various reasons. Variants are inherited in haplotypes which highlights the complexities of evolutionary history; studying a single variant may not fully illuminate the different mechanisms and factors resulting in altered biological states to a degree matching that of haplotype studies. Variants may have a limited impact on the structure and function of a protein when acting individually but produce more significant effects when presented in haplotype form. This opens a new avenue to variant interpretation.

Suffice it to say that the cumulative effect of variants making up a star allele is neither simple nor trivial.

4.3 Contextual relevance

Despite numerous efforts to characterise African-specific genetic variation, there remains a gap in knowledge pertaining to the genetic variation harboured on the African continent (Da Rocha et al. 2021, Wang et al. 2022). This is evident in commonly used genetic variation databases. African populations exhibit distinct CYP star allele frequencies and patterns when compared to other populations (Kane 2021, Wang et al. 2022). In recent years, there has been a surge in sequencing studies carried out in African populations accompanied by a wealth of novel genetic variation data. However, these endeavours have led to a large volume of genetic data and a backlog in the interpretation of genetic variation.

We have illuminated the impact that resource limitation has on the interpretation of genetic variation in the previous section [Emerging themes]. A majority of African countries are classified as low- and middle-income countries (LMICs). Due to insufficient investments in genomics research by LMICs (WHO 2021), there may not be enough trained personnel, infrastructure and other research resources to effectively perform large-scale genetic interpretation studies.

Structural bioinformatics as a means to prioritise genetic variation that is most likely to result in protein function change could potentially reduce the financial cost associated with studies that seek to interpret genetic variation. This enables the processing of more genetic variation in Africa while easing the financial burden in resource-limited Africa.

4.4 Strengths and limitations

The strengths and limitations of this study have been discussed in-depth in Chapter 3 (the published article). This section expands on a select few and gives a broader view of the strengths and limitations of this MSc study.

Although there have been significant improvements in computational variant effect prediction over the years, a key issue remains to be predicting accuracy, which is de-

pendent on the quality of the data and the methodologies used. There remains a lack of standardisation in computational variant effect prediction (Livesey & Marsh 2022). Structural bioinformatics employs numerous methods and approaches, making it difficult to compare and assess outcomes. A consistent set of benchmarks and evaluation criteria could aid in improving the accuracy and dependability of computational predictions. However, our NMA integrated FoldX, which is a top-performing protein stability predictor (Gerasimavicius et al. 2020), and ENCoM. The combination of both these algorithms has been found to yield highly reliable results (Frappier et al. 2015).

It should be noted that many genetic variant stability predictor algorithms have low accuracy when predicting stabilising variants (Buß et al. 2018). In this study, we combined FoldX and ENCoM to find coincident predictors and as a means to minimise inaccuracies in predictions.

Another limitation of this study is the absence of consideration for substrate-specific effects. These are interactions unique to different substrates and the partner protein active sites. Substrate-specific effects are determined by various factors: complementary shapes, electrostatic interactions, hydrogen bonding, hydrophobic interactions, and other molecular forces all play a role. Due to the diverse nature of substrates, we were unable to capture specific substrate interactions within the time frame of the MSc study. However, describing substrate-specific effects more rapidly using force fields and scoring functions is an ongoing research interest (Meli et al. 2022). Findings from this study serve as a critical foundation for drug-gene pair-focused studies.

We were able to obtain molecular dynamic (MD) properties of only CYP2D6*2 and CYP2D6*17, out of all the existing star alleles. MD simulation requires significant time to execute. Conducting MD simulation for a large number of star alleles would not be feasible due to the constraints of the MSc time frame. Additionally, The computational complexity of MD simulations presents inherent limitations. MD simulations are divided into small time steps that measure the movements of atoms and each time step captures a snapshot of fast-occurring motions. Smaller time steps are more accurate but are more computationally demanding. Moreover, lengthier simulations (often biologically relevant) necessitate more computational resources and require time to simulate the protein system. There is often a trade-off between the number of protein systems one can simulate and the length of the simulation. In this study, we simulated our protein systems for 600ns

which is biologically relevant but also requires significant time spent preparing and conducting the simulation, resulting in the limit to CYP2D6*2 and CYP2D6*17 (including the simulations of individual missense variants that make up these two star alleles).

4.5 Future work and directions

This study exemplifies the use of integrated approaches in pharmacogenomic research to address gaps in CYP variant effect prediction. This research serves as the foundation for subsequent explorations:

Structural bioinformatics as a complementary approach to other commonly used VEPs

This research highlights the utility of structural bioinformatics approaches and presents them as a key layer of evidence in ADME missense variant prediction. Structural bioinformatics samples unique measures that inform about the mechanistic effect of missense variants on protein structural properties. The findings from this study may be added to a searchable database such as the Database for Non-Synonymous SNPs Functional Predictions (dbNSFP) (Liu et al. 2020) to be used complementary with other Ensembl VEP plugins (www.ensembl.org).

Scaling study to cover all possible CYP missense variants

This study can be expanded to include all possible missense variants from CYP genes to inform about their potential functional impact. Regarding MD simulations and the acknowledgement that the execution of a simulation requires a substantial investment of time, advancements in computational power in the form of graphics processing units enable researchers to perform longer and high-resolution protein simulations. MD simulation is expensive if run on dedicated hardware. However, the large uptake of cloud-based high-performance computing can reduce the cost associated with advanced computational hardware by providing users with access to high-performance computing resources (Zhou et al. 2021). This direction entails the implementation of MD simulations, and other computationally demanding structural bioinformatics approaches, in a scalable manner.

Furthermore, the scaling of structural bioinformatics approaches will enable the de-

tailed interpretation of rare variants and haplotypes which may go on to support their inclusion in other types of analysis such as their inclusion in genotyping panels.

Interpretation of protein-ligand interactions

The inclusion of a ligand in protein models can provide distinct insights relating to substrate-specific effects of CYP proteins. It has been noted that *CYP2D6*17* exhibits substrate-specific properties (van der Lee et al. 2021). Furthermore, in a study conducted by Dong et al. (2022), it was observed that the combined influence of CYP2D6*14 variants (P34S, G169R, R296C, and S486T) played a role in altering the conformation of both access channels and the active site, consequently leading to variable ligand interactions. Another study by Othman et al. (2021) noted that CYP proteins missense affect the plasticity of the active site and ligand access channels and the conformations of other functionally crucial regions of the protein. This not only underscores the significance of examining protein-ligand interactions but also underscores the importance of assessing missense variants in star allele combinations as opposed to individual variants. The incorporation of ligand information may inform about how missense genetic variation can affect the stability of CYP proteins and affect ligand binding affinities. A future study may include binding pocket analysis and docking analysis to identify and analyse ligand-binding sites in the CYP protein and to predict the binding affinity of ligands. Ultimately, this may be used to provide additional insights into altered catalysis and contribute to the advancement of precision medicine.

Expansion of interpretation techniques

This study focused on three key post-MD simulation analyses, namely RMSD, RMSF, and PCA. This selection was influenced by the time constraints associated with the MSc program. Despite this limitation, the chosen analytical methods yielded valuable insights into protein structure and function. It is noteworthy that an additional post-MD analysis method, specifically the evaluation of the radius of gyration, could have enriched our study findings by offering additional insights into the overall size and shape of the protein systems as they evolve through simulation time (Ahmed et al. 2020).

Another analysis technique of interest could be the assessment of the distribution of RMSD. It encompasses all sampled conformations, providing insight into the overall

structural dynamics of the system. The distribution of RMSD for modelled star alleles could provide insight into the range of structural changes across the protein ensemble, highlighting patterns of variability among haplotypes and informing hypotheses regarding their functional consequences.

4.6 Conclusion

We demonstrated the utility of structural bioinformatics techniques in screening a large number of CYP variants and star alleles for their impact on structural features that may inform about protein function. In particular, multi-scale modelling can enable the prediction of cross-modulating effects of CYP missense variant combinations within star alleles. Structural bioinformatics techniques are therefore complementary to existing VEPs and experimental techniques such as DMS with regards to assessing CYP star allele functional alterations influenced by missense variation.

Bibliography

Adamcik, J. & Mezzenga, R. (2018), 'Amyloid polymorphism in the protein folding and aggregation energy landscape', *Angewandte Chemie International Edition* **57**(28), 8370–8382.

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S. & Sunyaev, S. R. (2010), 'A method and server for predicting damaging missense mutations', *Nature methods* **7**(4), 248–249.

Ahmad, H. I., Afzal, G., Jamal, A., Kiran, S., Khan, M. A., Mehmood, K., Kamran, Z., Ahmed, I., Ahmad, S., Ahmad, A. et al. (2021), '*In silico* structural, functional, and phylogenetic analysis of cytochrome (CYPD) protein family', *BioMed Research International* **2021**.

Ahmed, M. C., Crehuet, R. & Lindorff-Larsen, K. (2020), 'Computing, analyzing, and comparing the radius of gyration and hydrodynamic radius in conformational ensembles of intrinsically disordered proteins', *Intrinsically disordered proteins: methods and protocols* pp. 429–445.

Amorosi, C. J., Chiasson, M. A., McDonald, M. G., Wong, L. H., Sitko, K. A., Boyle, G., Kowalski, J. P., Rettie, A. E., Fowler, D. M. & Dunham, M. J. (2021), 'Massively parallel characterization of CYP2C9 variant enzyme activity and abundance', *The American Journal of Human Genetics* **108**(9), 1735–1751.

Ampadu, H. H., Hoekman, J., de Bruin, M. L., Pal, S. N., Olsson, S., Sartori, D., Leufkens, H. G. & Dodoo, A. N. (2016), 'Adverse drug reaction reporting in Africa and a comparison of individual case safety report characteristics between Africa and the rest of the world: analyses of spontaneous reports in VigiBase®', *Drug Safety* **39**(4), 335–345.

Anzenbacher, P. & Anzenbacherova, E. (2001), 'Cytochromes P450 and metabolism of xenobiotics', *Cellular and Molecular Life Sciences CMLS* **58**(5), 737–747.

B Tata, E., A Ambele, M. & S Pepper, M. (2020), 'Barriers to Implementing Clinical Pharmacogenetics Testing in Sub-Saharan Africa. A Critical Review', *Pharmaceutics* **12**(9), 809.

Bahar, I., Lezon, T. R., Bakan, A. & Shrivastava, I. H. (2010), 'Normal mode analysis of biomolecular structures: functional mechanisms of membrane proteins', *Chemical reviews* **110**(3), 1463–1497.

Bapiro, T. E., Hasler, J. A., Ridderström, M. & Masimirembwa, C. M. (2002), 'The molecular and enzyme kinetic basis for the diminished activity of the cytochrome P450 2D6. 17 (CYP2D6. 17) variant: potential implications for CYP2D6 phenotyping studies and the clinical use of CYP2D6 substrate drugs in some African populations', *Biochemical pharmacology* **64**(9), 1387–1398.

Borba, M. A., Melo-Neto, R. P., Leitão, G. M., Castelletti, C. H., Lima-Filho, J. L. & Martins, D. B. (2016), 'Evaluating the impact of missenses mutations in *CYP2D6*7* and *CYP2D6*14A*: does it compromise tamoxifen metabolism?', *Pharmacogenomics* **17**(6), 561–570.

Botton, M. R., Whirl-Carrillo, M., Del Tredici, A. L., Sangkuhl, K., Cavallari, L. H., Agúndez, J. A., Duconge, J., Lee, M. T. M., Woodahl, E. L., Claudio-Campos, K. et al. (2021), 'PharmVar GeneFocus: *CYP2C19*', *Clinical Pharmacology & Therapeutics* **109**(2), 352–366.

Bousman, C. A., Stevenson, J. M., Ramsey, L. B., Sangkuhl, K., Hicks, J. K., Strawn, J. R., Singh, A. B., Ruaño, G., Mueller, D. J., Tsermpini, E. E. et al. (2023), 'Clinical Pharmacogenetics Implementation Consortium (CPIC) guideline for CYP2D6, CYP2C19, CYP2B6, SLC6A4, and HTR2A genotypes and serotonin reuptake inhibitor antidepressants', *Clinical Pharmacology & Therapeutics* .

Brini, E., Simmerling, C. & Dill, K. (2020), 'Protein storytelling through physics', *Science* **370**(6520), eaaz3041.

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E. et al. (2019), 'The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019', *Nucleic acids research* **47**(D1), D1005–D1012.

Burley, S. K., Bhikadiya, C., Bi, C., Bittrich, S., Chao, H., Chen, L., Craig, P. A., Crichlow, G. V., Dalenberg, K., Duarte, J. M. et al. (2023), 'RCSB Protein Data Bank (RCSB. org): delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning', *Nucleic acids research* **51**(D1), D488–D508.

Burley, S. K., Bhikadiya, C., Bi, C., Bittrich, S., Chen, L., Crichlow, G. V., Christie, C. H., Dalenberg, K., Di Costanzo, L., Duarte, J. M. et al. (2021), 'RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences', *Nucleic acids research* **49**(D1), D437–D451.

Buß, O., Rudat, J. & Ochsenreither, K. (2018), 'FoldX as protein engineering tool: better than random based approaches?', *Computational and structural biotechnology journal* **16**, 25–33.

Campbell, M. C. & Tishkoff, S. A. (2008), 'African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping', *Annual review of genomics and human genetics* **9**, 403.

Case, D. A., Cheatham III, T. E., Darden, T., Gohlke, H., Luo, R., Merz Jr, K. M., Onufriev, A., Simmerling, C., Wang, B. & Woods, R. J. (2005), 'The Amber biomolecular simulation programs', *Journal of computational chemistry* **26**(16), 1668–1688.

Catucci, G., Ciaramella, A., Di Nardo, G., Zhang, C., Castrignanò, S. & Gilardi, G. (2022), 'Molecular Lego of human cytochrome P450: The key role of heme domain flexibility for the activity of the chimeric proteins', *International Journal of Molecular Sciences* **23**(7), 3618.

Caudle, K. E., Sangkuhl, K., Whirl-Carrillo, M., Swen, J. J., Haidar, C. E., Klein, T. E., Gammal, R. S., Relling, M. V., Scott, S. A., Hertz, D. L. et al. (2020), 'Standardizing

CYP 2D6 genotype to phenotype translation: consensus recommendations from the Clinical Pharmacogenetics Implementation Consortium and Dutch Pharmacogenetics Working Group', *Clinical and translational science* **13**(1), 116–124.

Cavallari, L., Beitelshees, A., Blake, K., Dressler, L., Duarte, J., Elsey, A., Eichmeyer, J., Empey, P., Franciosi, J., Hicks, J. et al. (2017), 'The IGNITE Pharmacogenetics Working Group: an opportunity for building evidence with pharmacogenetic implementation in a real-world setting', *Clinical and translational science* **10**(3), 143.

Cheatham III, T. E. & Case, D. A. (2013), 'Twenty-five years of nucleic acid simulations', *Biopolymers* **99**(12), 969–977.

Cheng, J., Novati, G., Pan, J., Bycroft, C., Žemgulytė, A., Applebaum, T., Pritzel, A., Wong, L. H., Zielinski, M., Sargeant, T. et al. (2023), 'Accurate proteome-wide missense variant effect prediction with AlphaMissense', *Science* **381**(6664), eadg7492.

Chumnumwat, S., Lu, Z. H., Sukasem, C., Winther, M. D., Capule, F. R., Hamid, A. A. A., Bhandari, B., Chanhom, N., Chantarangsu, S., Charoenyingwattana, A. et al. (2019), 'Southeast asian pharmacogenomics research network (SEAPharm): current status and perspectives', *Public health genomics* **22**(3-4), 132–139.

Cirauqui Diaz, N., Frezza, E. & Martin, J. (2021), 'Using normal mode analysis on protein structural models. How far can we go on our predictions?', *Proteins: Structure, Function, and Bioinformatics* **89**(5), 531–543.

Collins, F. S., Morgan, M. & Patrinos, A. (2003), 'The Human Genome Project: lessons from large-scale biology', *Science* **300**(5617), 286–290.

da Rocha, J. B., Othman, H. & Hazelhurst, S. (2022), 'Molecular dynamics of G6PD variants from sub-Saharan Africa', *Biochemistry and Biophysics Reports* **30**, 101236.

Da Rocha, J. E., Othman, H., Botha, G., Cottino, L., Twesigomwe, D., Ahmed, S., Drögemöller, B. I., Fadlelmola, F. M., Machanick, P., Mbiyavanga, M. et al. (2021), 'The extent and impact of variation in ADME genes in sub-Saharan African populations', *Frontiers in Pharmacology* **12**, 634016.

David, C. C. & Jacobs, D. J. (2011), 'Characterizing protein motions from structure', *Journal of Molecular Graphics and Modelling* **31**, 41–56.

de Almeida Paiva, V., de Souza Gomes, I., Monteiro, C. R., Mendonça, M. V., Martins, P. M., Santana, C. A., Gonçalves-Almeida, V., Izidoro, S. C., de Melo-Minardi, R. C. & de Azevedo Silveira, S. (2022), 'Protein structural bioinformatics: An overview', *Computers in Biology and Medicine* p. 105695.

Department of Higher Education and Training (2014), 'National Scarce Skills List: Top 100 Occupations in Demand'.
URL: *https://www.gov.za/sites/default/files/gcis$_d$ocument*/201409/37678*gen*380*.pdf*

Desta, Z., Gammal, R. S., Gong, L., Whirl-Carrillo, M., Gaur, A. H., Sukasem, C., Hockings, J., Myers, A., Swart, M., Tyndale, R. F. et al. (2019), 'Clinical Pharmacogenetics Implementation Consortium (CPIC) guideline for CYP2B6 and efavirenz-containing antiretroviral therapy', *Clinical Pharmacology & Therapeutics* **106**(4), 726–733.

Dill, K. A. & MacCallum, J. L. (2012), 'The protein-folding problem, 50 years on', *science* **338**(6110), 1042–1046.

Dong, A. N., Ahemad, N., Pan, Y., Palanisamy, U. D., Yiap, B. C. & Ong, C. E. (2022), 'Role of P34S, G169R, R296C, and S486T Substitutions in Ligand Access and Catalysis for Cytochrome P450 2D6 Allelic Variants CYP2D6* 14A and CYP2D6* 14B', *Drug Metabolism and Bioanalysis Letters Formerly: Drug Metabolism Letters* **15**(1), 51–63.

Drake, J. A. & Pettitt, B. M. (2020), 'Physical chemistry of the protein backbone: Enabling the mechanisms of intrinsic protein disorder', *The Journal of Physical Chemistry B* **124**(22), 4379–4390.

Engler, N., Ostermann, A., Niimura, N. & Parak, F. G. (2003), 'Hydrogen atoms in proteins: positions and dynamics', *Proceedings of the National Academy of Sciences* **100**(18), 10243–10248.

Fanni, D., Pinna, F., Gerosa, C., Paribello, P., Carpiniello, B., Faa, G. & Manchia, M. (2021), 'Anatomical distribution and expression of CYP in humans: Neuropharmacological implications', *Drug Development Research* **82**(5), 628–667.

Fischer, A., Don, C. G. & Smiesko, M. (2018), 'Molecular dynamics simulations reveal structural differences among allelic variants of membrane-anchored cytochrome P450 2D6', *Journal of chemical information and modeling* **58**(9), 1962–1975.

Fowler, D. M. & Fields, S. (2014), 'Deep mutational scanning: a new style of protein science', *Nature methods* **11**(8), 801–807.

Frappier, V., Chartier, M. & Najmanovich, R. J. (2015), 'ENCoM server: exploring protein conformational space and the effect of mutations on protein function and stability', *Nucleic acids research* **43**(W1), W395–W400.

Frappier, V. & Najmanovich, R. J. (2014), 'A coarse-grained elastic network atom contact model and its use in the simulation of protein dynamics and the prediction of the effect of mutations', *PLoS computational biology* **10**(4), e1003569.

Fujikura, K., Ingelman-Sundberg, M. & Lauschke, V. M. (2015), 'Genetic variation in the human cytochrome P450 supergene family', *Pharmacogenetics and genomics* **25**(12), 584–594.

Gaedigk, A., Ingelman-Sundberg, M., Miller, N. A., Leeder, J. S., Whirl-Carrillo, M., Klein, T. E. & Committee, P. S. (2018), 'The Pharmacogene Variation (PharmVar) Consortium: incorporation of the human cytochrome P450 (CYP) allele nomenclature database', *Clinical Pharmacology & Therapeutics* **103**(3), 399–401.

Gaedigk, A., Sangkuhl, K., Whirl-Carrillo, M., Twist, G. P., Klein, T. E., Miller, N. A., Committee, P. S. et al. (2019), 'The evolution of PharmVar', *Clinical pharmacology and therapeutics* **105**(1), 29.

Gaedigk, A., Whirl-Carrillo, M., Pratt, V. M., Miller, N. A. & Klein, T. E. (2020), 'PharmVar and the landscape of pharmacogenetic resources', *Clinical pharmacology and therapeutics* **107**(1), 43.

Gerasimavicius, L., Liu, X. & Marsh, J. A. (2020), 'Identification of pathogenic missense mutations using protein stability predictors', *Scientific Reports* **10**(1), 1–10.

Gerasimavicius, L., Livesey, B. J. & Marsh, J. A. (2023), 'Correspondence between functional scores from deep mutational scans and predicted effects on protein stability', *Protein Science* p. e4688.

Gong, H., Zhang, H., Zhu, J., Wang, C., Sun, S., Zheng, W.-M. & Bu, D. (2017), 'Improving prediction of burial state of residues by exploiting correlation among residues', *BMC bioinformatics* **18**(3), 165–175.

Gonzalez, N. A., Li, B. A. & McCully, M. E. (2022), 'The stability and dynamics of computationally designed proteins', *Protein Engineering, Design and Selection* **35**.

Gyulkhandanyan, A., Rezaie, A. R., Roumenina, L., Lagarde, N., Fremeaux-Bacchi, V., Miteva, M. A. & Villoutreix, B. O. (2020), 'Analysis of protein missense alterations by combining sequence-and structure-based methods', *Molecular genetics & genomic medicine* **8**(4), e1166.

Hamdi, Y., Zass, L., Othman, H., Radouani, F., Allali, I., Hanachi, M., Okeke, C. J., Chaouch, M., Tendwa, M. B., Samtal, C. et al. (2021), 'Human OMICs and computational biology research in Africa: current challenges and prospects', *Omics: a journal of integrative biology* **25**(4), 213–233.

Harmalkar, A. & Gray, J. J. (2021), 'Advances to tackle backbone flexibility in protein docking', *Current opinion in structural biology* **67**, 178–186.

Holm, L. (2020), 'Structural bioinformatics, methods and protocols', *Methods Mol Biology* **2112**(29-1106), 42.

Horne, J. & Shukla, D. (2022), 'Recent advances in machine learning variant effect prediction tools for protein engineering', *Industrial & engineering chemistry research* **61**(19), 6235–6245.

Ingelman-Sundberg, M., Sim, S. C., Gomez, A. & Rodriguez-Antona, C. (2007), 'Influence of cytochrome P450 polymorphisms on drug therapies: pharmacogenetic, pharmacoepigenetic and clinical aspects', *Pharmacology & therapeutics* **116**(3), 496–526.

Isvoran, A., Louet, M., Vladoiu, D. L., Craciun, D., Loriot, M.-A., Villoutreix, B. O. & Miteva, M. A. (2017), 'Pharmacogenomics of the cytochrome P450 2C family: impacts of amino acid variations on drug metabolism', *Drug Discovery Today* **22**(2), 366–376.

Ittisoponpisan, S., Islam, S. A., Khanna, T., Alhuzimi, E., David, A. & Sternberg, M. J. (2019), 'Can predicted protein 3D structures provide reliable insights into whether missense variants are disease associated?', *Journal of Molecular Biology* **431**(11), 2197–2212.

Jisna, V. & Jayaraj, P. (2021), 'Protein structure prediction: conventional and deep learning perspectives', *The protein journal* **40**(4), 522–544.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A. et al. (2021), 'Highly accurate protein structure prediction with AlphaFold', *Nature* **596**(7873), 583–589.

Kaledin, M., Kaledin, A. L., Brown, A. & Bowman, J. M. (2005), Driven molecular dynamics for normal modes of biomolecules without the Hessian, and beyond, *in* 'Normal Mode Analysis', Chapman and Hall/CRC, pp. 305–324.

Kalman, L. V., Agúndez, J. A., Appell, M. L., Black, J. L., Bell, G. C., Boukouvala, S., Bruckner, C., Bruford, E., Caudle, K., Coulthard, S. et al. (2016), 'Pharmacogenetic allele nomenclature: International workgroup recommendations for test result reporting', *Clinical Pharmacology & Therapeutics* **99**(2), 172–185.

Kane, M. (2021), *CYP2D6* overview: allele and phenotype frequencies, *in* 'Medical Genetics Summaries [Internet]', National Center for Biotechnology Information (US).

Kmiecik, S., Gront, D., Kolinski, M., Wieteska, L., Dawid, A. E. & Kolinski, A. (2016), 'Coarse-grained protein models and their applications', *Chemical reviews* **116**(14), 7898–7936.

Ko, J. & Lee, J. (2021), 'Can AlphaFold2 predict protein-peptide complex structures accurately?', *BioRxiv* pp. 2021–07.

Koeppe, J. R., McDonald, A. R., Roberts, R. & Craig, P. A. (2020), 'Integrating Computation and Wet Lab Methods in a Biochemistry Lab Course-Based Undergraduate Research Experience (Cure)', *Biophysical Journal* **118**(3), 318a.

Koulgi, S., Jani, V., Phukan, S., Sonavane, U., Joshi, R., Kamboj, R. K. & Palle, V. (2022), 'A Deep Dive into the Conformational Dynamics of CYP3A4: Understanding the Binding of Homotropic and Non-homotropic Ligands for Mitigating Drug-Drug interaction (DDI)', *ChemistrySelect* **7**(17), e202200249.

Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. & Moult, J. (2021), 'Critical assessment of methods of protein structure prediction (CASP)—Round XIV', *Proteins: Structure, Function, and Bioinformatics* **89**(12), 1607–1617.

Kumar, S. (2010), 'Engineering cytochrome P450 biocatalysts for biotechnology, medicine and bioremediation', *Expert opinion on drug metabolism & toxicology* **6**(2), 115–131.

Li, C., Zhang, L., Zhuo, Z., Su, F., Li, H., Xu, S., Liu, Y., Zhang, Z., Xie, Y., Yu, X. et al. (2023), 'Artificial intelligence-based recognition for variant pathogenicity of BRCA1 using AlphaFold2-predicted structures', *Theranostics* **13**(1), 391.

Lin, J. & Ngiam, K. Y. (2023), 'How data science and AI-based technologies impact genomics', *Singapore Medical Journal* **64**(1), 59.

Liu, X., Li, C., Mou, C., Dong, Y. & Tu, Y. (2020), 'dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human non-synonymous and splice-site SNVs', *Genome medicine* **12**(1), 1–8.

Livesey, B. J. & Marsh, J. A. (2022), 'Interpreting protein variant effects with computational predictors and deep mutational scanning', *Disease Models & Mechanisms* **15**(6), dmm049510.

Lunenburg, C. A., van der Wouden, C. H., Nijenhuis, M., Crommentuijn-van Rhenen, M. H., de Boer-Veger, N. J., Buunk, A. M., Houwink, E. J., Mulder, H., Rongen, G. A., van Schaik, R. H. et al. (2020), 'Dutch Pharmacogenetics Working Group (DPWG) guideline for the gene–drug interaction of DPYD and fluoropyrimidines', *European Journal of Human Genetics* **28**(4), 508–517.

Lynch, T. & Neff, A. P. (2007), 'The effect of cytochrome P450 metabolism on drug response, interactions, and adverse effects', *American family physician* **76**(3), 391–396.

Mahajan, S. & Sanejouand, Y.-H. (2015), 'On the relationship between low-frequency normal modes and the large-scale conformational changes of proteins', *Archives of biochemistry and biophysics* **567**, 59–65.

Mahmood, K., Jung, C.-h., Philip, G., Georgeson, P., Chung, J., Pope, B. J. & Park, D. J. (2017), 'Variant effect prediction tools assessed using independent, functional assay-based datasets: implications for discovery and diagnostics', *Human genomics* **11**(1), 1–8.

Maranville, J. C. & Cox, N. J. (2016), 'Pharmacogenomic variants have larger effect sizes than genetic variants associated with other dichotomous complex traits', *The pharmacogenomics journal* **16**(4), 388–392.

Martín-García, F., Papaleo, E., Gomez-Puertas, P., Boomsma, W. & Lindorff-Larsen, K. (2015), 'Comparing molecular dynamics force fields in the essential subspace', *PLoS One* **10**(3), e0121114.

Martínez, L. (2015), 'Automatic identification of mobile and rigid substructures in molecular dynamics simulations and fractional structural fluctuation analysis', *PloS one* **10**(3), e0119264.

Masimirembwa, C. M. & Hasler, J. A. (1997), 'Genetic polymorphism of drug metabolising enzymes in African populations: implications for the use of neuroleptics and antidepressants', *Brain research bulletin* **44**(5), 561–571.

McCammon, J. (1984), 'Protein dynamics', *Reports on Progress in Physics* **47**(1), 1.

McGibbon, R. T., Beauchamp, K. A., Harrigan, M. P., Klein, C., Swails, J. M., Hernández, C. X., Schwantes, C. R., Wang, L.-P., Lane, T. J. & Pande, V. S. (2015), 'MDTraj: a modern open library for the analysis of molecular dynamics trajectories', *Biophysical journal* **109**(8), 1528–1532.

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., Flicek, P. & Cunningham, F. (2016), 'The ensembl variant effect predictor', *Genome biology* **17**(1), 1–14.

Meli, R., Morris, G. M. & Biggin, P. C. (2022), 'Scoring functions for protein-ligand binding affinity prediction using structure-based deep learning: A review', *Frontiers in bioinformatics* **2**, 57.

Meystre, S. M. & Gouripeddi, R. (2019), Clinical research in the postgenomic era, *in* 'Clinical Research Informatics', Springer, pp. 147–168.

Mhalu, F. S. (2005), 'Burden of diseases in poor resource countries: meeting the challenges of combating HIV/AIDS, tuberculosis and malaria', *Tanzania Journal of Health Research* **7**(3), 179–184.

Mpye, K., Matimba, A., Dzobo, K., Chirikure, S., Wonkam, A. & Dandara, C. (2017), 'Disease burden and the role of pharmacogenomics in African populations', *Global health, epidemiology and genomics* **2**.

Mulder, N., Abimiku, A., Adebamowo, S. N., de Vries, J., Matimba, A., Olowoyo, P., Ramsay, M., Skelton, M. & Stein, D. J. (2018), 'H3Africa: current perspectives', *Pharmacogenomics and personalized medicine* **11**, 59.

Mulder, N., Zass, L., Hamdi, Y., Othman, H., Panji, S., Allali, I. & Fakim, Y. J. (2021), 'African Global Representation in Biomedical Sciences', *Annual Review of Biomedical Data Science* **4**, 57–81.

Ng, P. C. & Henikoff, S. (2003), 'SIFT: Predicting amino acid changes that affect protein function', *Nucleic acids research* **31**(13), 3812–3814.

Nguyen, H., Roe, D. R., Swails, J. & Case, D. A. (2016), 'PYTRAJ: Interactive data analysis for molecular dynamics simulations', *New Brunswick, NJ: Rutgers University* .

Othman, H., da Rocha, J. E. & Hazelhurst, S. (2021), 'Single nucleotide polymorphism induces divergent dynamic patterns in CYP3A5: a microsecond scale biomolecular simulation of variants identified in Sub-Saharan African populations', *International Journal of Molecular Sciences* **22**(15), 7786.

Othman, H., Jemimah, S. & da Rocha, J. E. B. (2022), 'SWAAT Bioinformatics Workflow for Protein Structure-Based Annotation of ADME Gene Variants', *Journal of Personalized Medicine* **12**(2), 263.

Otyepka, M., Skopalík, J., Anzenbacherová, E. & Anzenbacher, P. (2007), 'What common structural features and variations of mammalian P450s are known to date?', *Biochimica et Biophysica Acta (BBA)-General Subjects* **1770**(3), 376–389.

Palamini, M., Canciani, A. & Forneris, F. (2016), 'Identifying and visualizing macromolecular flexibility in structural biology', *Frontiers in molecular biosciences* **3**, 47.

Peña-Varas, C., Kanstrup, C., Vergara-Jaque, A., González-Avendaño, M., Crocoll, C., Mirza, O., Dreyer, I., Nour-Eldin, H. & Ramírez, D. (2022), 'Structural insights into the substrate transport mechanisms in GTR transporters through ensemble docking', *International Journal of Molecular Sciences* **23**(3), 1595.

Peter, S. C., Dhanjal, J. K., Malik, V., Radhakrishnan, N., Jayakanthan, M. & Sundar, D. (2019), Quantitative Structure-Activity Relationship (QSAR): Modeling Approaches to

Biological Applications, *in* S. Ranganathan, M. Gribskov, K. Nakai & C. Schönbach, eds, 'Encyclopedia of Bioinformatics and Computational Biology', Academic Press, Oxford, pp. 661–676.

Porta-Pardo, E., Ruiz-Serra, V., Valentini, S. & Valencia, A. (2022), 'The structural coverage of the human proteome before and after AlphaFold', *PLOS Computational Biology* **18**(1), e1009818.

Poulos, T. & Johnson, E. (2005), 'Cytochrome P450: structure, mechanism, and biochemistry', *New York: Springer US* pp. 87–114.

Poulos, T. L., Finzel, B. C. & Howard, A. J. (1987), 'High-resolution crystal structure of cytochrome P450cam', *Journal of molecular biology* **195**(3), 687–700.

Relling, M. & Klein, T. (2011), 'CPIC: clinical pharmacogenetics implementation consortium of the pharmacogenomics research network', *Clinical Pharmacology & Therapeutics* **89**(3), 464–467.

Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. (2019), 'CADD: predicting the deleteriousness of variants throughout the human genome', *Nucleic acids research* **47**(D1), D886–D894.

Robarge, J. D., Li, L., Desta, Z., Nguyen, A. & Flockhart, D. A. (2007), 'The star-allele nomenclature: retooling for translational genomics', *Clinical Pharmacology & Therapeutics* **82**(3), 244–248.

Samish, I., Bourne, P. E. & Najmanovich, R. J. (2015), 'Achievements and challenges in structural bioinformatics and computational biophysics', *Bioinformatics* **31**(1), 146–150.

Sargsyan, K., Grauffel, C. & Lim, C. (2017), 'How molecular size impacts RMSD applications in molecular dynamics simulations', *Journal of chemical theory and computation* **13**(4), 1518–1524.

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F. & Serrano, L. (2005), 'The FoldX web server: an online force field', *Nucleic acids research* **33**(suppl_2), W382–W388.

Shahrokh, K., Orendt, A., Yost, G. S. & Cheatham III, T. E. (2012), 'Quantum mechanically derived AMBER-compatible heme parameters for various states of the cytochrome P450 catalytic cycle', *Journal of computational chemistry* **33**(2), 119–133.

Sinai, S. & Kelsic, E. D. (2020), 'A primer on model-guided exploration of fitness landscapes for biological sequence design', *arXiv preprint arXiv:2010.10614* .

Sinner, C., Lutz, B., John, S., Reinartz, I., Verma, A. & Schug, A. (2014), 'Simulating Biomolecular Folding and Function by Native-Structure-Based/Go-Type Models', *Israel Journal of Chemistry* **54**(8-9), 1165–1175.

Sittel, F., Jain, A. & Stock, G. (2014), 'Principal component analysis of molecular dynamics: On the use of Cartesian vs. internal coordinates', *The Journal of Chemical Physics* **141**(1), 07B605_1.

Smith, L. M., Agar, J. N., Chamot-Rooke, J., Danis, P. O., Ge, Y., Loo, J. A., Paša-Tolić, L., Tsybin, Y. O., Kelleher, N. L. & for Top-Down Proteomics, C. (2021), 'The human proteoform project: defining the human proteome', *Science advances* **7**(46), eabk0734.

Somody, J. C., MacKinnon, S. S. & Windemuth, A. (2017), 'Structural coverage of the proteome for pharmaceutical applications', *Drug discovery today* **22**(12), 1792–1799.

Šrejber, M., Navrátilová, V., Paloncỳová, M., Bazgier, V., Berka, K., Anzenbacher, P. & Otyepka, M. (2018), 'Membrane-attached mammalian cytochromes P450: An overview of the membrane's effects on structure, drug binding, and interactions with redox partners', *Journal of inorganic biochemistry* **183**, 117–136.

Theken, K. N., Lee, C. R., Gong, L., Caudle, K. E., Formea, C. M., Gaedigk, A., Klein, T. E., Agúndez, J. A. & Grosser, T. (2020), 'Clinical Pharmacogenetics Implementation Consortium Guideline (CPIC) for CYP2C9 and nonsteroidal anti-inflammatory drugs', *Clinical Pharmacology & Therapeutics* **108**(2), 191–200.

Tian, C., Kasavajhala, K., Belfon, K. A., Raguette, L., Huang, H., Migues, A. N., Bickel, J., Wang, Y., Pincay, J., Wu, Q. et al. (2019), 'ff19SB: Amino-acid-specific protein backbone parameters trained against quantum mechanics energy surfaces in solution', *Journal of chemical theory and computation* **16**(1), 528–552.

Torrens-Fontanals, M., Stepniewski, T. M., Aranda-García, D., Morales-Pastor, A., Medel-Lacruz, B. & Selent, J. (2020), 'How do molecular dynamics data complement static structural data of GPCRs', *International Journal of Molecular Sciences* **21**(16), 5933.

Twesigomwe, D., Drögemöller, B. I., Wright, G. E., Adebamowo, C., Agongo, G., Boua, P. R., Matshaba, M., Paximadis, M., Ramsay, M., Simo, G. et al. (2023), 'Characterization of CYP2D6 Pharmacogenetic Variation in Sub-Saharan African Populations', *Clinical Pharmacology & Therapeutics* **113**(3), 643–659.

Twesigomwe, D., Wright, G. E., Drögemöller, B. I., da Rocha, J., Lombard, Z. & Hazelhurst, S. (2020), 'A systematic comparison of pharmacogene star allele calling bioinformatics algorithms: a focus on CYP2D6 genotyping', *NPJ genomic medicine* **5**(1), 1–11.

van der Lee, M., Guchelaar, H.-J. & Swen, J. J. (2021), 'Substrate specificity of CYP2D6 genetic variants', *Pharmacogenomics* **22**(16), 1081–1089.

Vascon, F., Gasparotto, M., Giacomello, M., Cendron, L., Bergantino, E., Filippini, F. & Righetto, I. (2020), 'Protein electrostatics: From computational and structural analysis to discovery of functional fingerprints and biotechnological design', *Computational and structural biotechnology journal* **18**, 1774–1789.

Wade, R. C., Winn, P. J., Schlichting, I. et al. (2004), 'A survey of active site access channels in cytochromes P450', *Journal of inorganic biochemistry* **98**(7), 1175–1182.

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

Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. (2004), 'Development and testing of a general amber force field', *Journal of computational chemistry* **25**(9), 1157–1174.

Wang, W. Y., Twesigomwe, D., Nofziger, C., Turner, A. J., Helmecke, L.-S., Broeckel, U., Derezinski, A. D., Hazelhurst, S. & Gaedigk, A. (2022), 'Characterization of novel CYP2D6 alleles across sub-Saharan African populations', *Journal of Personalized Medicine* **12**(10), 1575.

WHO (2021), 'Gross domestic Ramp;D expenditure on Health (Health Gerd) as a % of GDP'.

URL: *https://www.who.int/observatories/global-observatory-on-health-research-and-development/indicators/gross-domestic-r-d-expenditure-on-health-as-a-percent-of-gross-domestic-product*

Wright, G., Carleton, B., Hayden, M. & Ross, C. (2018), 'The global spectrum of protein-coding pharmacogenomic diversity', *The pharmacogenomics journal* **18**(1), 187–195.

Zanger, U. M. & Klein, K. (2013), 'Pharmacogenetics of cytochrome P450 2B6 (CYP2B6): advances on polymorphisms, mechanisms, and clinical relevance', *Frontiers in genetics* **4**, 24.

Zhang, L., Sarangi, V., Moon, I., Yu, J., Liu, D., Devarajan, S., Reid, J. M., Kalari, K. R., Wang, L. & Weinshilboum, R. (2020), 'CYP2C9 and CYP2C19: deep mutational scanning and functional characterization of genomic missense variants', *Clinical and translational science* **13**(4), 727–742.

Zhang, Z., Wang, L., Gao, Y., Zhang, J., Zhenirovskyy, M. & Alexov, E. (2012), 'Predicting folding free energy changes upon single point mutations', *Bioinformatics* **28**(5), 664–671.

Zhou, A. C., Lao, J., Ke, Z., Wang, Y. & Mao, R. (2021), 'Farspot: Optimizing monetary cost for HPC applications in the cloud spot market', *IEEE Transactions on Parallel and Distributed Systems* **33**(11), 2955–2967.

Zhou, Y. & Lauschke, V. M. (2022), 'The genetic landscape of major drug metabolizing cytochrome P450 genes—an updated analysis of population-scale sequencing data', *The Pharmacogenomics Journal* **22**(5-6), 284–293.

Appendices

Appendix A: Ethics waiver certificate

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

Human Research Ethics Committee (Medical)

Research Office Secretariat:
Faculty of Health Sciences, Phillip Tobias Health Sciences Building, 3rd Floor, Office 301/2/4, 29 Princess of Wales Terrace, Parktown, 2193
Private Bag 3, Wits 2050
Office email: HREC-Medical.ResearchOffice@wits.ac.za
Website: www.wits.ac.za/research/about-our-research/ethics-and-research-integrity/

Ref: W-CBP-210615-01 15/06/2021

TO WHOM IT MAY CONCERN:

Waiver:	This certifies that the following research does not require clearance from the Human Research Ethics Committee (Medical)
Investigator:	Mr Nhlamulo Khoza
Supervisor:	Dr Houcemeddine Othman
Department:	Human Genetics
Project title:	Characterising the combined effects of cytochrome P450 missense variants within the star allele nomenclature
Reason:	Review of information in the public domain. No human participants will be involved in the study.

CB Penny

Dr CB Penny

Chairperson: Human Research Ethics Committee (Medical)

Copy – HREC (Medical) Secretariat: Ms Zanele Ndlovu, Ms Mapula Ramaila and Mr Rhulani Mkansi

Appendix B: Authors' agreement

Declaration: Student's contribution to articles and agreement of co-authors

I, Nhlamulo Khoza, student number 1620584, declare that this dissertation is my own work and that I contributed adequately towards research findings published in the article stated below which is included in my dissertation.

Signature of student: _____ Date: 31/08/2023

Name of primary supervisor: Houcemeddine Othman

Signature of primary supervisor: Houcemeddine Othman (Sep 8, 2023 04:45 GMT+1) Date: 08/09/2023

Agreement by co-authors: By signing this declaration, the co-authors listed below agree to the use of the article by the student as part of his dissertation.

Article

Title: Characterising the combined effects of cytochrome P450 missense variation within star allele definitions

Journal name, year, doi: *Pharmacogenomics*, 2023. doi: 10.2217/pgs-2023-0068.

Authors	Name	Signature	Date
1st	Nhlamulo Khoza		31/08/2023
2nd	David Twesigomwe		31/08/2023
3rd	Houcemeddine Othman	Houcemeddine Othman (Sep 8, 2023 04:45 GMT+1)	08/09/2023

Comments by primary supervisor:

Mr. Nhlamulo Khoza played a key role in formulating the concept, designing the experiments, conducting data analysis, and taking the lead in writing the paper.

Appendix C: Plagiarism declaration and similarity index



PLAGIARISM DECLARATION TO BE SIGNED BY ALL HIGHER DEGREE STUDENTS

SENATE PLAGIARISM POLICY: APPENDIX ONE

I Nhlamulo Khoza (Student number: 1620584) am a student registered for the degree of MSc Medicine (Human Genetics) in the academic year 2023.

I hereby declare the following:

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

Signature: _____ Date: 29/08/2023

1

Nhlamulo_by_publication.pdf

ORIGINALITY REPORT

16%	10%	14%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Nhlamulo Khoza, David Twesigomwe, Houcemeddine Othman. "Characterizing the combined effects of cytochrome P450 missense variation within star allele definitions", Pharmacogenomics, 2023 Publication	6%
2	Submitted to University of Witwatersrand Student Paper	1%
3	www.frontiersin.org Internet Source	1%
4	medworm.com Internet Source	1%