ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ
UNIVERSITY OF CRETE

FORTH

MEDICAL SCHOOL OF UNIVERSITY OF CRETE

FOUNDATION FOR RESEARCH AND TECHNOLOGY

INTERDEPARTMENTAL PROGRAM OF POSTGRADUATE STUDIES

BIOINFORMATICS

Correlation between CYP2D6 genetic variants and their metabolic activity insight from Molecular Dynamics Simulations

MASTER'S THESIS

DANAI MARIA KOTZAMPASI

SUPERVISOR: Dr. GEORGE POTAMIAS

CO-SUPERVISOR: Assist. Prof. VANGELIS DASKALAKIS

HERAKLION, JUNE 2021

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ
UNIVERSITY OF CRETE

FORTH

ΙΑΤΡΙΚΗ ΣΧΟΛΗ ΠΑΝΕΠΙΣΤΗΜΙΟΥ ΚΡΗΤΗΣ
ΙΝΣΤΙΤΟΥΤΟ ΤΕΧΝΟΛΟΓΙΑΣ ΚΑΙ ΕΡΕΥΝΑΣ

ΔΙΙΔΡΥΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ

Συσχέτιση των γενετικών παραλλαγών του CYP2D6 με τη μεταβολική τους δραστηριότητα μέσω προσομοιώσεων Μοριακής Δυναμικής

ΜΕΤΑΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΔΑΝΑΗ ΜΑΡΙΑ ΚΟΤΖΑΜΠΑΣΗ

ΕΠΙΒΛΕΠΩΝ: Δρ. ΓΕΩΡΓΙΟΣ ΠΟΤΑΜΙΑΣ

ΣΥΝΕΠΙΒΛΕΠΩΝ: Επικ. Καθ. ΒΑΓΓΕΛΗΣ ΔΑΣΚΑΛΑΚΗΣ

ΗΡΑΚΛΕΙΟ, ΙΟΥΝΙΟΣ 2021

Επιβλέπων

Ποταμιάς Γεώργιος, Κύριος Ερευνητής (Ερευνητής Β'), Ινστιτούτο Πληροφορικής (ΙΠ), Ίδρυμα Τεχνολογίας και Έρευνας (ΙΤΕ)

Τριμελής Επιτροπή

Ποταμιάς Γεώργιος, Κύριος Ερευνητής (Ερευνητής Β'), Ινστιτούτο Πληροφορικής (ΙΠ), Ίδρυμα Τεχνολογίας και Έρευνας (ΙΤΕ)

Δασκαλάκης Βαγγέλης, Επίκουρος καθηγητής, Τμήμα Επιστήμης και Τεχνολογίας Περιβάλλοντος, Τεχνολογικό Πανεπιστήμιο Κύπρου

Ηλιόπουλος Ιωάννης, Αναπληρωτής καθηγητής, Τμήμα Ιατρικής, Πανεπιστήμιο Κρήτης

Πρόλογος

Η παρούσα μεταπτυχιακή εργασία με τίτλο «Συσχέτιση των γενετικών παραλλαγών του CYP2D6 με τη μεταβολική τους δραστηριότητα μέσω προσομοιώσεων Μοριακής Δυναμικής» πραγματοποιήθηκε στα πλαίσια του διιδρυματικού μεταπτυχιακού προγράμματος της Ιατρικής Σχολής του Πανεπιστημίου Κρήτης και του ΙΤΕ με τίτλο: Βιοπληροφορική, στο Εργαστήριο Υπολογιστικής Βιο-Ιατρικής (CBML) του Ινστιτούτου Τεχνολογίας και Έρευνας, κατά το Ακαδημαϊκό έτος 2020-2021. Καταρχάς, θα ήθελα να ευχαριστήσω τον επιβλέποντα Ερευνητή κύριο Ποταμιά, για την υποστήριξη του, την ενθάρρυνση και την εμπιστοσύνη που μου έδειξε σε όλη τη διάρκεια της μεταπτυχιακής μου εργασίας. Θα ήθελα επίσης να ευχαριστήσω τον Επίκουρο καθηγητή κύριο Δασκαλάκη για την καθοδήγηση του και την σημαντική συμβολή του σε θέματα Μοριακής Δυναμικής. Επίσης, ευχαριστώ τον Αναπληρωτή Καθηγητή κύριο Ηλιόπουλο που δέχτηκε να συμμετάσχει στην τριμελή εξεταστική επιτροπή. Στη συνέχεια, θα ήθελα να ευχαριστήσω τον ερευνητή Αλέξανδρο Καντεράκη για όλη την υποστήριξη και τις συζητήσεις που κάναμε καθόλη τη διάρκεια της μεταπτυχιακής εργασίας. Τέλος, δεν θα μπορούσα να παραλείψω όλους τους συμφοιτητές του Μεταπτυχιακού Προγράμματος που έκαναν αυτά τα δύο χρόνια του μεταπτυχιακού ακόμα πιο όμορφα.

Περίληψη

Τα ένζυμα του κυτοχρώματος P450 ανήκουν σε μια υπεροικογένεια ενζύμων που περιέχουν μια αίμη συμπαράγοντα και είναι υπεύθυνα για τον μεταβολισμό περισσότερου από το 90% των κλινικών φαρμάκων. Ένα από τα σημαντικότερα ένζυμα αυτής της οικογένειας, το κυτόχρωμα P450 2D6 (CYP2D6), μεταβολίζει περίπου 25% των κλινικά χρησιμοποιούμενων φαρμάκων, συμπεριλαμβανομένων κρίσιμων και συχνά χορηγούμενων φαρμάκων όπως τα αντικαταθλιπτικά, τα χημειοθεραπευτικά, οι β-αναστολείς και τα οπιοειδή. Οι παραλλαγές του CYP2D6, ενός εξαιρετικά πολυμορφικού τόπου στο γονιδίωμα, είναι ικανές να αλλάξουν τη μεταβολική λειτουργικότητα του, επηρεάζοντας την αποτελεσματικότητα και την τοξικότητα πολλών φαρμάκων. Περισσότεροι από 100 απλότυποι του ενζύμου CYP2D6 έχουν ταυτοποιηθεί και καταχωρηθεί στη βάση δεδομένων του Pharmacogene Variation Consortium (PharmVar, www. pharmvar. org), παρουσιάζοντας μεγάλες διακυμάνσεις στη ικανότητα μεταβολισμού φαρμάκων και οδηγώντας σε μεταβολές της συγκέντρωσης των φαρμάκων στο πλάσμα. Ο πλήρης συσχετισμός μεταξύ των γενετικών παραλλαγών και της μεταβολικής ικανότητας εξακολουθεί να αποτελεί ένα ανοιχτό και δύσκολο ερώτημα. Ο κύριος στόχος μας ήταν να διερευνήσουμε τους παράγοντες που είναι καθοριστικοί για την μεταβολική δραστηριότητα του ενζύμου αξιοποιώντας και χρησιμοποιώντας κατάλληλα μεθόδους Μοριακής Δυναμικής. Η Μοριακή Δυναμική είναι μια εξελιγμένη υπολογιστική μέθοδος που επιτρέπει την πρόβλεψη της χρονικής εξέλιξης των θέσεων των ατόμων μέσα σε αλληλεπιδρώντα συστήματα μορίων. Για το σκοπό αυτό, εξετάσαμε τη δυναμική πολυάριθμων παραλλαγών του CYP2D6, ως μοντέλα λειτουργικών και μη λειτουργικών ενζύμων. Καταλήξαμε στο συμπέρασμα ότι οι μεταβολές στους b-factors των καταλοίπων και η ανάλυση Dynamic cross-correlation μπορούν να χρησιμοποιηθούν ως αναλύσεις για τη διάκριση των δύο κατηγοριών μεταβολικής δραστηριότητας. Η ανάλυση Molecular Docking μεταξύ των παραλλαγών του CYP2D6 και του BACE1 αναστολέα επιβεβαίωσε τα αποτελέσματα μας και ανέδειξε το ρόλο της έλικας I και της K-K' loop και της σχετική τους κίνησης στη δραστηριότητα του ενζύμου. Τα αποτελέσματα Μοριακής Δυναμικής του CYP2D6 *1 χρησιμοποιήθηκαν στον προσδιορισμό των σημαντικών περιοχών της πρωτεΐνης μέσω Markov State Modeling. Βασιζόμενοι σε αυτές τις περιοχές της πρωτεϊνης και χρησιμοποιώντας τα δεδομένα από την ανάλυση tICA/MSM, δημιουργήθηκε ένα dataset για κάθε αλληλόμορφο του CYP2D6, το οποίο στη συνέχεια χρησιμοποιήθηκε για τη δημιουργία ενός μοντέλου πρόβλεψης της μεταβολικής ικανότητας των διαφορετικών αλληλομόρφων. Είναι η πρώτη φορά που αναπτύσσεται ένα τέτοιο εργαλείο. Τα αποτελέσματα αυτής της εργασίας έχουν μεγάλη σημασία για τομείς όπως η Εξατομικευμένη Ιατρική, η πρόβλεψη ανεπιθύμητων ενεργειών φαρμάκων και η ανακάλυψη νέων φαρμάκων.

Summary

Cytochrome P450s enzyme belongs to the superfamily of heme-containing proteins, responsible for metabolizing more than 90% of clinical drugs. One of the most significant enzymes in this family, Cytochrome P450 2D6 (CYP2D6), metabolizes ~25% of the clinically used drugs including crucial and commonly administered drugs such as antidepressants, chemotherapeutics, beta-blockers and opioids. Variations in CYP2D6, a highly polymorphic loci in the genome, could alter its activity influencing the efficacy and toxicity of numerous drugs. More than 100 haplotypes (star alleles) of the drug metabolizing enzyme CYP2D6 have been reported in the Pharmacogene Variation Consortium (PharmVar, www.pharmvar.org), resulting in wide intraindividual variability in drug metabolism activity and changes of the drug plasma concentration. The complete connecting link between the genetic variants and the metabolizer phenotype is still an open and challenging question. Our main objective was to investigate the key factors that determine the metabolizer phenotype by exploiting and appropriately employing molecular dynamics (MD) methods. MD is an elaborate computational method that enables the prediction of the time evolution of atomic positions within interacting systems of molecules. To this end, we have probed the dynamics of numerous CYP2D6 variants, as enzyme models with normal and no function, at all-atom resolution. We concluded that changes in residue b-factors and Dynamical Cross-correlation analysis could be used as markers in the discrimination of the two classes of metabolizing activity. Molecular docking analysis between CYP2D6 variants and BACE1 inhibitor confirmed our observations and highlighted the role of helix I and of K-K' loop and their relative movement in the activity of the enzyme. Classical MD runs on the CYP2D6 *1 (wild-type) were used for identifying the important residues for the protein conformational space using Markov State Modeling. Based on these residues and using the data from the tICA/MSM analysis, a dataset for each variant has been produced which was then used to build a prediction model for the metabolizer phenotype. This is the first time such a tool has been developed. Results of this work are of great importance for areas like Personalized Medicine, Adverse Drug Reaction (ADR) prediction and drug discovery.

Keywords: Cytochrome P450, CYP2D6 genetic variants, drug metabolism, Molecular Dynamics.

1. Introduction

1.1 Cytochrome P450 Enzymes

Cytochrome P450 (CYPs) consists of one of the largest enzyme superfamilies that play a significant role in the metabolism of numerous endogenous compounds, drugs and other xenobiotics in almost all living organisms (Lynch and Price 2007). Cytochrome P450 enzymes are heme-containing monooxygenases that are located on the endoplasmic reticulum of cells or in mitochondrial membranes (Tsuneo Omura and Sato 1964). More than 50 different CYP genes have been identified in humans, classified into 18 families and 44 subfamilies (Korobkova 2015; “Human Cytochrome P450s” n.d.; Daniel W. Nebert, Wikvall, and Miller 2013). The first three CYP families, CYP1, CYP2 and CYP3, include the major enzymes involved in xenobiotic metabolism (Zanger and Schwab 2013).

In order to prevent incorrect assignments or duplications of P450 genes, CYP P450 Nomenclature Committee has established the nomenclature system for CYP 450 enzymes (Nelson et al. 1996). The term “P450” stands for the spectrophotometric peak at the maximum absorption wavelength of 450nm in the reduced state of the enzyme in the presence of carbon monoxide (T. Omura 1999). CYP enzymes are designated with the root symbol “CYP”, followed by a number which indicates the gene family, a letter for the subfamily and a final number for the individual gene as shown in Figure 1. Any two CYP enzymes with sequence identity greater than 40% belong to the same family and any two CYP enzymes with sequence identity greater than 55% belong to the same subfamily (D. W. Nebert et al. 1987).

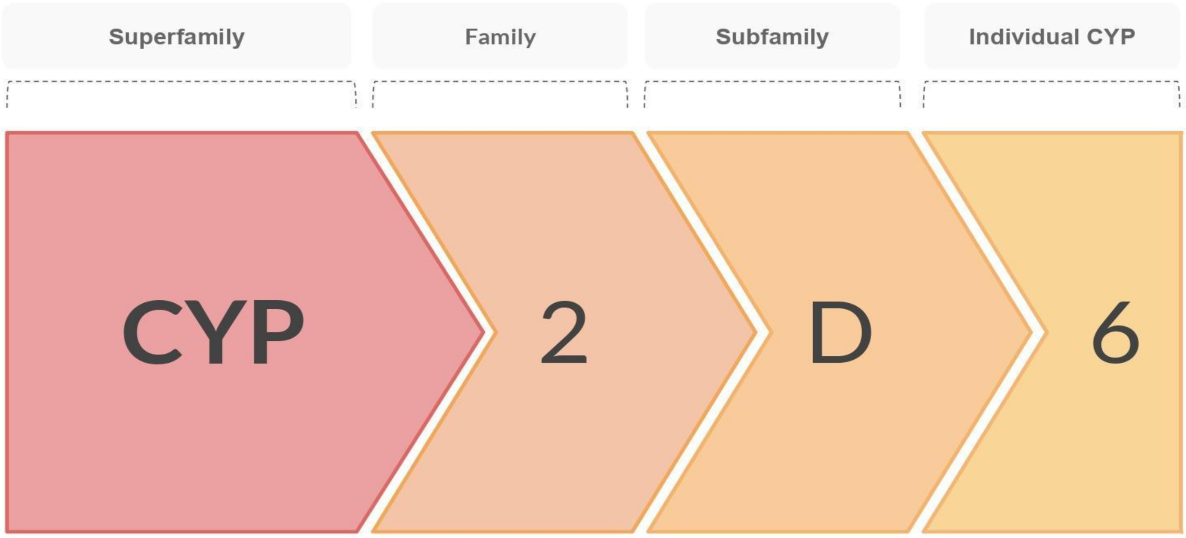


Figure 1 Cytochrome P450 enzyme nomenclature, the example of CYP2D6

1.2 The Structure of human Cytochrome P450 Enzymes

Cytochrome P450 enzymes have in general similar structures (Johnson and Stout 2013). They consist of about 500 amino acids and a heme factor buried in the active site (Manikandan and Nagini 2018). The three-dimensional structures of these hemoproteins, as shown in Figure 2, share a common overall fold and topology, being mostly α-helical with a small number of β-sheets (Denisov et al. 2005; Taylor et al. 2020; Hasemann et al. 1995). In all members of this enzyme superfamily the iron atom of the heme group is bound to the protein through the thiolate sulphur of a cysteine forming an anionic iron-cysteinate bond, a bond crucial for the function of the enzyme (Lamb and Waterman 2013). The cysteine residue which provides the sulfur atom to the heme iron along with the heme group, are the most conserved elements of the CYPs (J. Wu et al. 2021; Otyepka et al. 2007; Johnson and Stout 2005). The catalytic domain is connected to the transmembrane helix by a proline-rich region and the substrate-binding area is located on the distal side of the heme group (J. Wu et al. 2021; Otyepka et al. 2007). The conformational changes of structural domains that surround the heme complex play a key role in substrate accessibility of the active site.

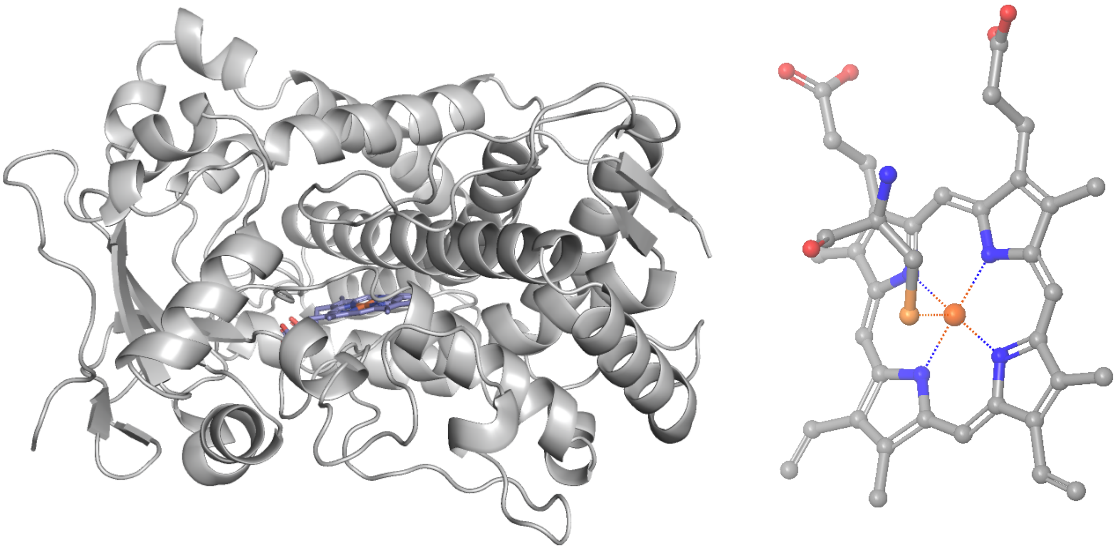


Figure 2 The general structure of human CYPs (left) and the crucial iron-cysteinate bond between the cysteine residue of the protein and the heme group (right)

1.3 The Mechanism of Cytochrome P450 Enzymes

Cytochrome P450 monooxygenase enzymes are present in all tissues but primarily in the liver and small intestine. In humans, cytochrome P450 enzymes are mostly membrane bound to endoplasmic reticulum and the inner membrane of mitochondria. However, they are present in many other tissues including the intestinal mucosa, brain, kidney, lung and skin. These proteins play a key role in bile acid biosynthesis, and metabolism of foreign compounds such as drugs, environmental pollutants, and carcinogens (Zanger and Schwab 2013). CYPs are also responsible for synthesis and degradation of endogenous steroid hormones and play a major role in vitamin metabolism, oxidation of unsaturated fatty acids, and cholesterol biosynthesis (Waring 2020; Manikandan and Nagini 2018; T. Omura 1999; Hasler et al. 1999). CYP enzymes participate in many different reactions including the hydrocarbon hydroxylation, epoxidation; O-, S-, and N-dealkylation, dehydrogenation, dehalogenation, oxidative deamination, decarboxylation, reductive dehalogenation, N-oxide, and epoxide reduction, isomerization and ring formation. However, oxidation is the most common reaction catalyzed by CYPs and includes the following steps (Figure 3) (J. Wu et al. 2021; Shaik et al. 2010; Isin and Guengerich 2007):

(1) Binding of the substrate to CYP450 enzyme. The binding of the substrates to CYP enzymes starts mainly when the iron atom of the heme complex is in ferric state (inactive resting state) and coordinated to a water molecule. The catalytic cycle starts with binding of the substrate (RH) to the ferric enzyme by displacing the water molecule from the heme iron. In several isoforms, such as CYP2D6, the water molecule is missing.

(2) Reduction of Fe III to Fe II. After the loss of the water molecule the pentacoordinated ferric-porphyrin becomes a better electron acceptor. Then, the first electron is provided from NADPH-CYP reductase (CPR) in order to reduce the ferric (Fe^{3+}) to ferrous (Fe^{2+}), a good dioxygen binder.

(3) Binding of an oxygen molecule to ferrous iron. An oxygen molecule binds to ferrous iron forming an oxy-ferrous complex which is a good electron acceptor.

(4) Reduction of the dioxygen complex. The formation of the oxy-ferrous complex triggers a second reduction of the system leading to the formation of the ferric-peroxo anion.

(5) Protonation of the ferric-peroxo anion. The ferric-peroxo anion acts as a good base and gets easily protonated leading to the formation of the ferric-hydroperdroperoxide species, Compound 0 (Cpd 0).

(6) Breakage of the O-O bond. In this step Cpd 0 takes another proton, the O-O bond breaks and a water molecule is released to form the iron-oxo species, the so-called Compound I (Cpd I).

(7) Monooxygenation of the substrate. The electron deficient complex withdraws an electron or a proton from the substrate.

(8) Heme iron returns in ferric state. Finally, substrate oxidation by this reactive complex produces the oxidized metabolite which exits the pocket and the CYP regenerates to its initial ferric state.

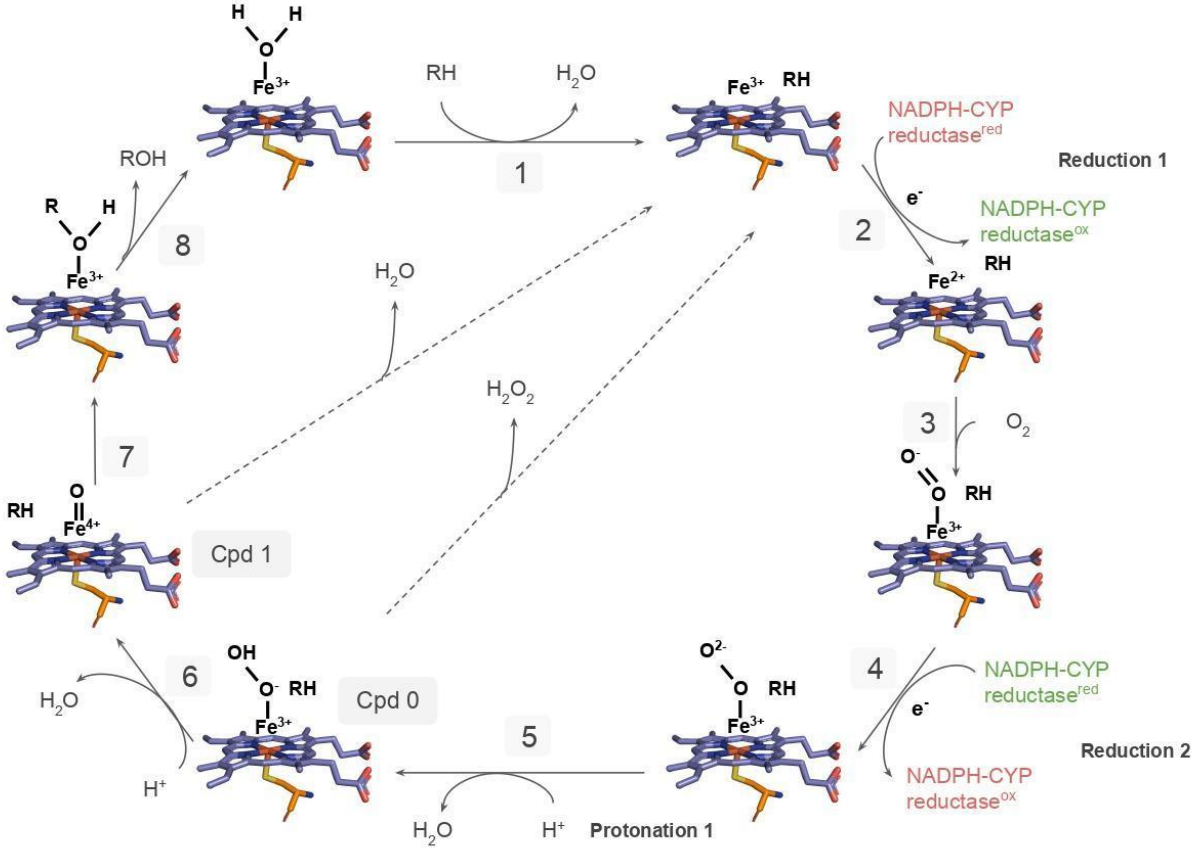


Figure 3 The general catalytic cycle of cytochrome P450 enzymes (for details see the text)

1.4 Cytochrome P450 Enzymes and drug metabolism

Drug metabolism refers to the process of the biotransformation of the drugs. During this process lipophilic compounds are converted into hydrophilic products that can be easily excreted from the body. Liver is the major site of metabolism, but other organs like kidney, placenta, adrenal gland, gastrointestinal tract and the skin also participate (Jaladanki et al. 2020).

There are two phases of drug metabolism; Phase I and Phase II (Figure 4) (Jancova, Anzenbacher, and Anzenbacherova 2010). Phase I consists of oxidation, reduction and hydrolysis reactions. At this stage, substrates cannot be easily excreted from the body. Phase II involves the so-called “conjugative reactions” with endogenous molecules such as glutathione, sulfate, glycine, glucuronic acid etc. After Phase II, metabolites are more polar and can be easily excreted. In some cases, reactive metabolites are formed before the Phase II stage which may lead to toxicity. Phase II reactions are catalyzed mainly by transferases such as UDP-glucuronosyltransferases, sulfotransferases, N-acetyltransferases, glutathione S-transferases and methyltransferases (Jaladanki et al. 2020; Jancova, Anzenbacher, and Anzenbacherova 2010; Kadlubar and Kadlubar 2010).

Cytochrome P450 enzymes and primarily CYP3A4, CYP2C9 and CYP2D6, are the major enzymes involved in the Phase I of drug metabolism (Lu and Xue 2019). During this stage, drugs either undergo deactivation by CYPs or bioactivation to form their active compounds. The majority of the members of the cytochrome P450 family present genetic polymorphism which influences their metabolizing phenotype and affects the efficacy and toxicity of the drug for patients who have very high or low metabolism rates. Based on different phenotypes within a population, individuals may be classified as "poor metabolizer" (PM), "intermediate metabolizer" (IM), "extensive metabolizer" (EM) or "ultra-rapid metabolizer" (UM) (Gaedigk et al. 2017). For example, rapid metabolizers clear the drug very quickly which can result in toxicity from reactive metabolites or in low drug efficacy due to the low drug plasma concentration. On the contrary, poor metabolizer phenotype can lead to higher drug concentrations and subsequently adverse side effects. Understanding and predicting the drug metabolism remains challenging but of great importance for drug discovery and development.

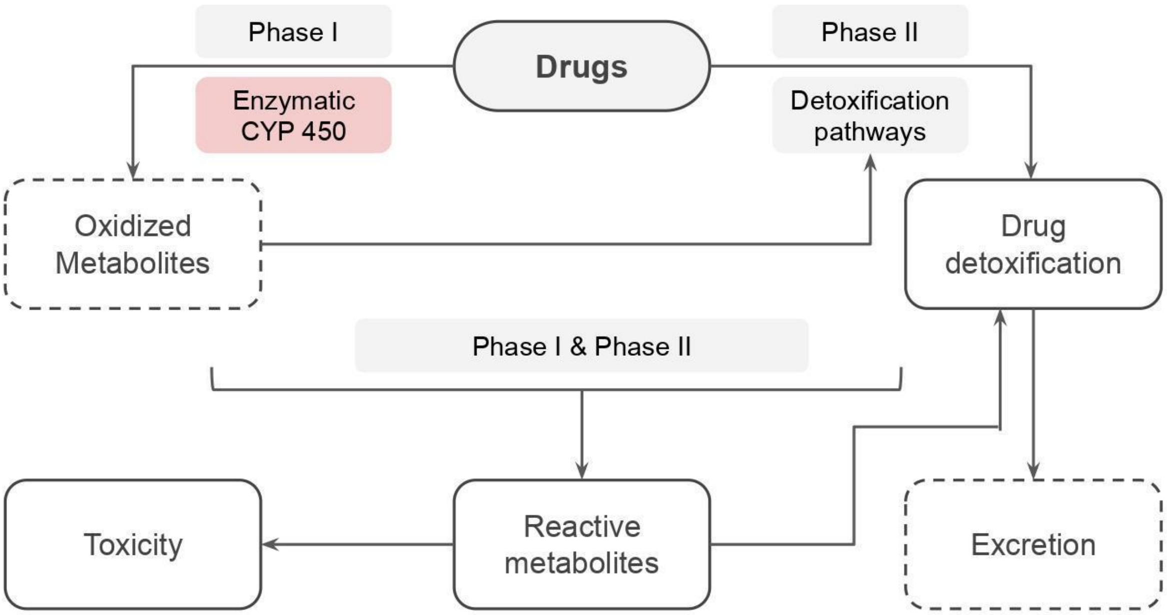
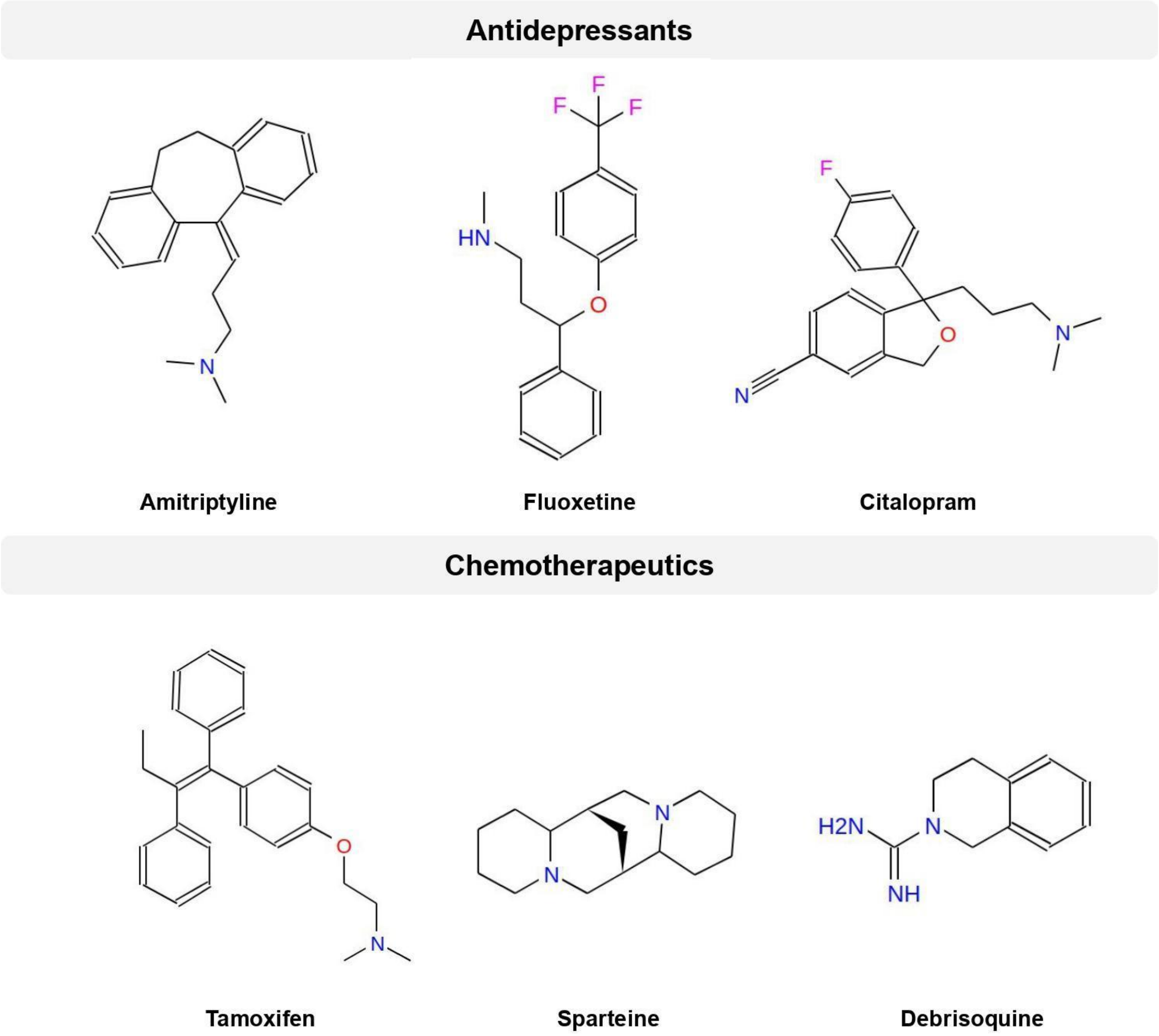


Figure 4 General pathways of drug metabolism

1.5 Cytochrome P450 2D6 (CYP2D6)

Cytochrome P450 2D6 (CYP2D6), one of the most significant enzymes in this superfamily, plays a fundamental role in the metabolism of many clinically important drugs. CYP2D6 metabolizes about 25% of the clinically used drugs (Petrović, Pešić, and Lauschke 2020; Ingelman-Sundberg et al. 2007) including crucial and commonly administered drugs such as antidepressants (i.e. amitriptyline and fluoxetine) (Brandl et al. 2014), chemotherapeutics (i.e. tamoxifen and irinotecan) (Algeciras-Schimnich, O'Kane, and Snozek 2008), beta-blockers (i.e. metoprolol) (Sharp et al. 2009) and opioids (i.e. codeine and tramadol) (Ruano and Kost 2018). Almost all known CYP2D6 substrates contain at least one aromatic ring which interacts with specific residues in the binding pocket (Figure 5) (J. Wang et al. 2010; Keizers et al. 2004).



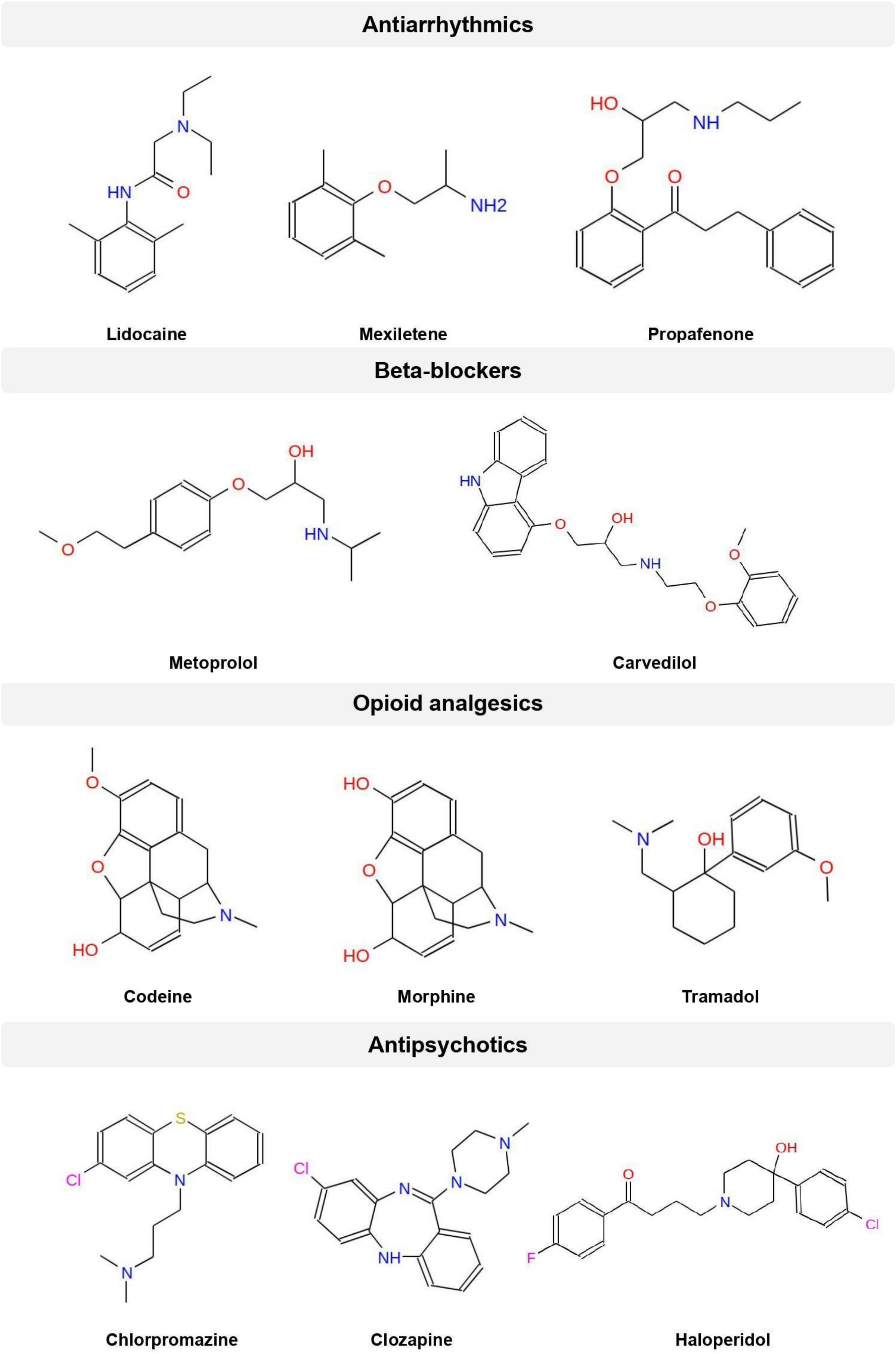
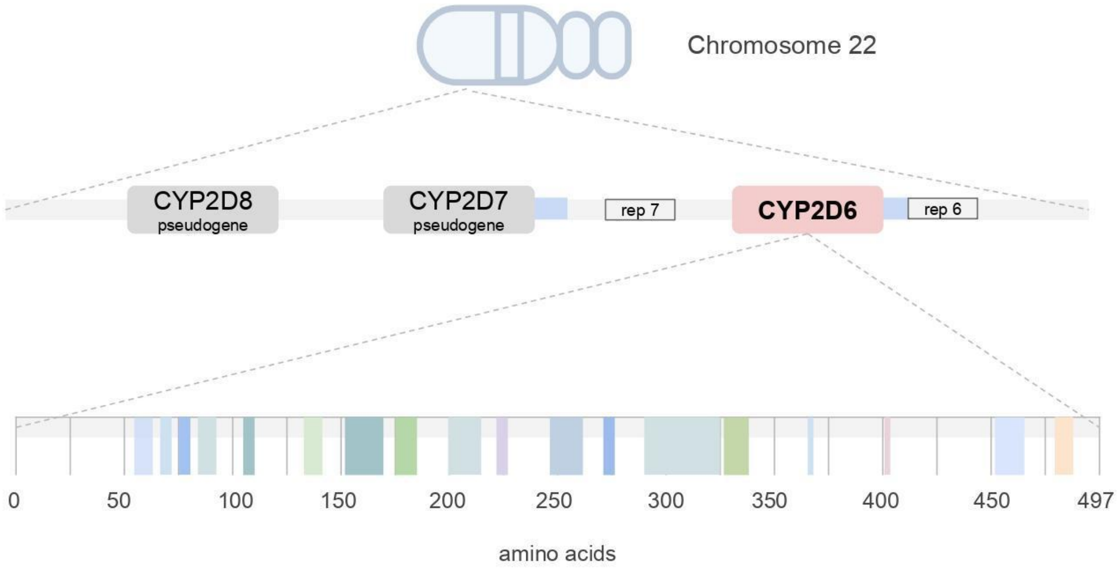


Figure 5 Examples of CYP2D6 substrates

The gene encoding CYP2D6 enzyme is located on chromosome 22q13.2 neighboring two pseudogenes, CYP2D7 and CYP2D8 (Figure 6) (Ruano and Kost 2018; Heim and Meyer 1992). Variations in CYP2D6, a highly polymorphic loci in the genome, affect the functionality of the enzyme influencing the efficacy and toxicity of numerous drugs (Ahmed et al. 2016). In comparison to the other drug-metabolizing CYPs, CYP2D6 is the only non-inducible enzyme which means that genetic alterations have a great influence on the interindividual variation in metabolizer phenotype (Ingelman-Sundberg et al. 2007). At present, more than 100 haplotypes (star alleles) of the drug metabolizing enzyme CYP2D6 with varying levels of evidence have been reported in the Pharmacogene Variation (PharmVar) Consortium ("PharmVar" n.d.) resulting in a wide intraindividual variability in drug metabolism activity and changes of the drug plasma concentration. The Clinical Pharmacogenetics Implementation Consortium (CPIC) ("Home Page" n.d.) along with PharmGKB ("PharmGKB" n.d.) has developed guidelines that enable the translation of genetic laboratory test results into actionable prescribing decisions for specific drugs. The current classification of CYP2D6 metabolizer status is based on activity scoring of known haplotypes [46]. Activity scores (AS) act as a tool to translate information regarding the function of individual haplotypes into an overall predicted metabolizer status for a given diplotype, and thus an individual (Taylor et al. 2020; Nofziger et al. 2020).



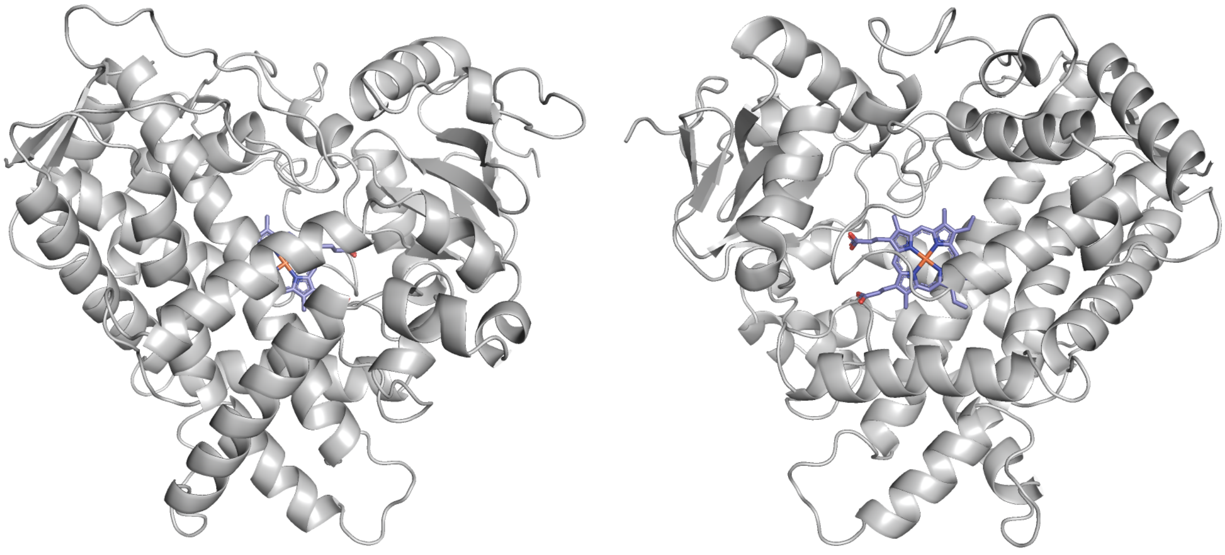


Figure 6 The relative position of CYP2D6 pharmacogene (red) to two non-functional pseudogenes (grey), *CYP2D7* and *CYP2D8* on the minus strand of Chromosome 22, the amino acid sequence encoded by this gene and the three-dimensional structure of the protein.

The *CYP2D6* polymorphism is historically one of the most representative examples of pharmacogenetics. Among all major drug metabolizing CYPs, CYP2D6 shows the greatest impact of genetic polymorphism due to its wide range of different allelic forms, comparably little influence by environmental and nongenetic factors, and its extraordinarily broad substrate selectivity. CYP2D6 has received much attention as far as cancer is concerned due to its role in the bioactivation of tamoxifen, 4-hydroxytamoxifen, and endoxifen used in the therapy of estrogen receptor-positive breast cancer (Lim et al. 2005; Kadlubar and Kadlubar 2010). It has been also associated with several diseases such as Parkinson disease due to the fact that CYP2D6 is expressed in the human brain and catalyzes the biosynthesis of dopamine from L-tyrosine via p-tyramine (Stefanović et al. 2000; X. Wang et al. 2014). Accurately detecting and predicting functional and non-functional star alleles, in clinically actionable pharmacogenes such as CYP2D6 is therefore crucial to the implementation of personalized medicine.

1.6 Molecular Dynamics

Molecular dynamics (MD) simulations are an important tool for predicting how every atom in a protein or other molecular system moves over time, based on a general model of the physics governing interatomic interactions. MD simulations can help us to understand the physical basis of the structure and function of biological macromolecules. They can be used to investigate the properties of a system, in some cases more easily than experiments on the actual system (Karplus and McCammon 2002; Hollingsworth and Dror 2018). Simulations can be used to answer a wide range of different questions. Some of the most common applications of MD Simulations include studies of conformational flexibility and stability, the investigation of how a biomolecular system responds to some perturbation (i.e. mutation or ligand binding) and the observation of a dynamic process over time (i.e. protein folding or membrane transport) (Hollingsworth and Dror 2018**).**

MD employs Molecular Mechanics (MM) to describe the interactions between atoms. The atomic forces that govern molecular movement can be classified into two major groups; Bonded and Non-bonded. Bonded interactions are referred to interactions between atoms that are chemically bonded, whereas non-bonded interactions are referred to interactions between atoms that are not bonded (Durrant and McCammon 2011). It uses Hooke's spring law for bonded interactions, Coulomb's law and the Lennard-Jones potential for non-bonded interactions, calculating the energy of the system E as a set of partial energies E_{bonded} and $E_{non\text{-}bonded}$.

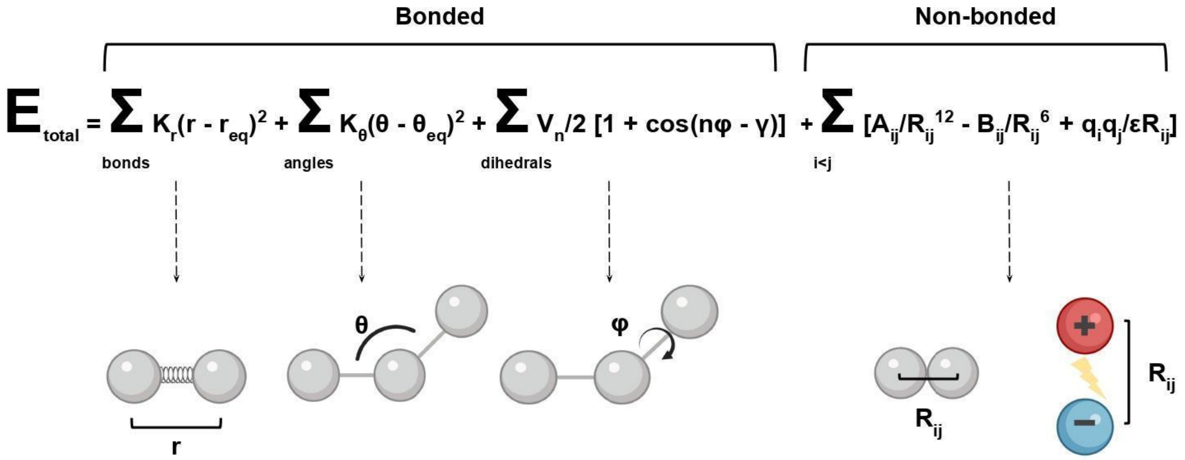

Bonded
Non-bonded
$E_{total} = \sum_{bonds} K_r(r - r_{eq})^2 + \sum_{angles} K_\theta(\theta - \theta_{eq})^2 + \sum_{dihedrals} V_n/2\,[1 + \cos(n\varphi - \gamma)] + \sum_{i<j} [A_{ij}/R_{ij}^{12} - B_{ij}/R_{ij}^{6} + q_iq_j/\varepsilon R_{ij}]$
r
θ
φ
R_{ij}
+
−
R_{ij}

Figure 7 Bonded and non-bonded interactions between the atoms of a system.

In MD, as in an experimental process, we can study the time evolution of a system, so that the system goes through all possible states. However, using simulations we can examine how systems evolve and go through different configurations, with adjustable variables (temperature, pressure, pH), at an atomic level, which is not always possible experimentally. To perform the simulation we need to follow specific steps (Figure 8). As in the experiment, we need to prepare the system, declare the initial positions and velocity distributions of the particles (to perform the simulation using specific temperature-pressure conditions) and then find an initial minimum energy structure relevant to the experimentally observed ensemble – average configurations. This is followed by the production phase of the simulation in which the energies, the forces that act on the atoms, their new speeds and positions are calculated for the next time step. After the end of the simulation, the data are collected and analyzed.

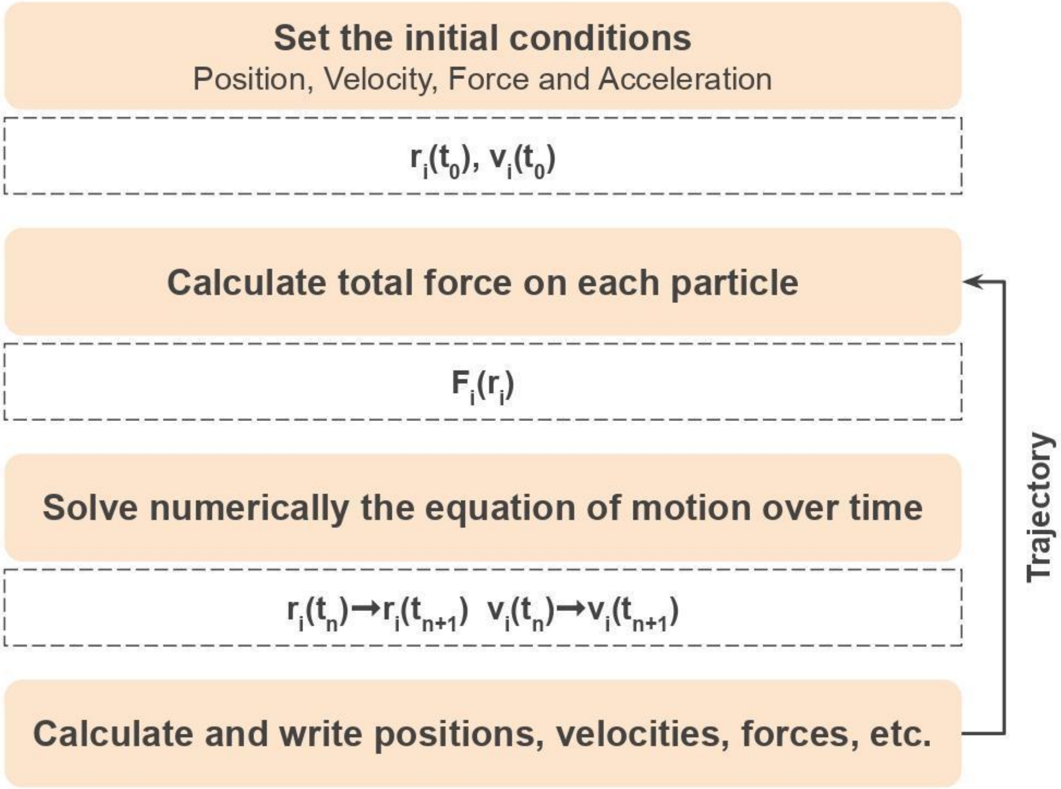
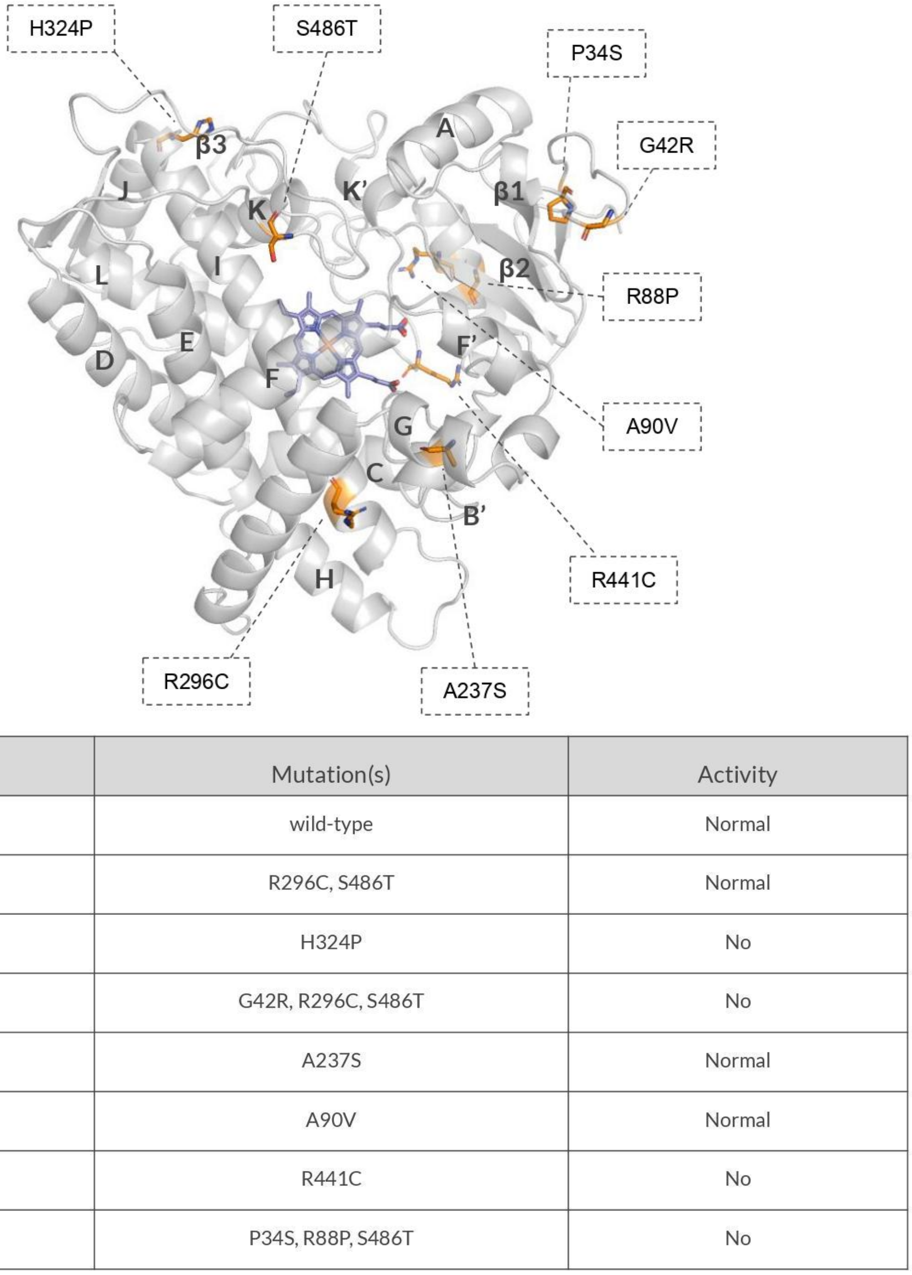

Set the initial conditions
Position, Velocity, Force and Acceleration
$r_i(t_0)$, $v_i(t_0)$
Calculate total force on each particle
$F_i(r_i)$
Solve numerically the equation of motion over time
$r_i(t_n) \rightarrow r_i(t_{n+1})$ $v_i(t_n) \rightarrow v_i(t_{n+1})$
Calculate and write positions, velocities, forces, etc.
Trajectory

Figure 8 Simplified Molecular Dynamics Simulation algorithm. All particles move according to Newton's second law or the equation of motion.

2. Aim of this work

The aim of the present study was to **advance scientific knowledge and investigate the key factors that determine the metabolizer phenotype of the different allelic forms of CYP2D6**. In order to do this, we have chosen to **probe the structural characteristics of several CYP2D6 star alleles** with definitive PharmVar level of evidence and known metabolic activity, as shown in Figure 9, by **exploiting and appropriately employing MD** methods and analytical tools. Results can thus be directly associated with applications in **Personalized Medicine** and can help to **improve the efficacy and safety of various prescription medicines**.



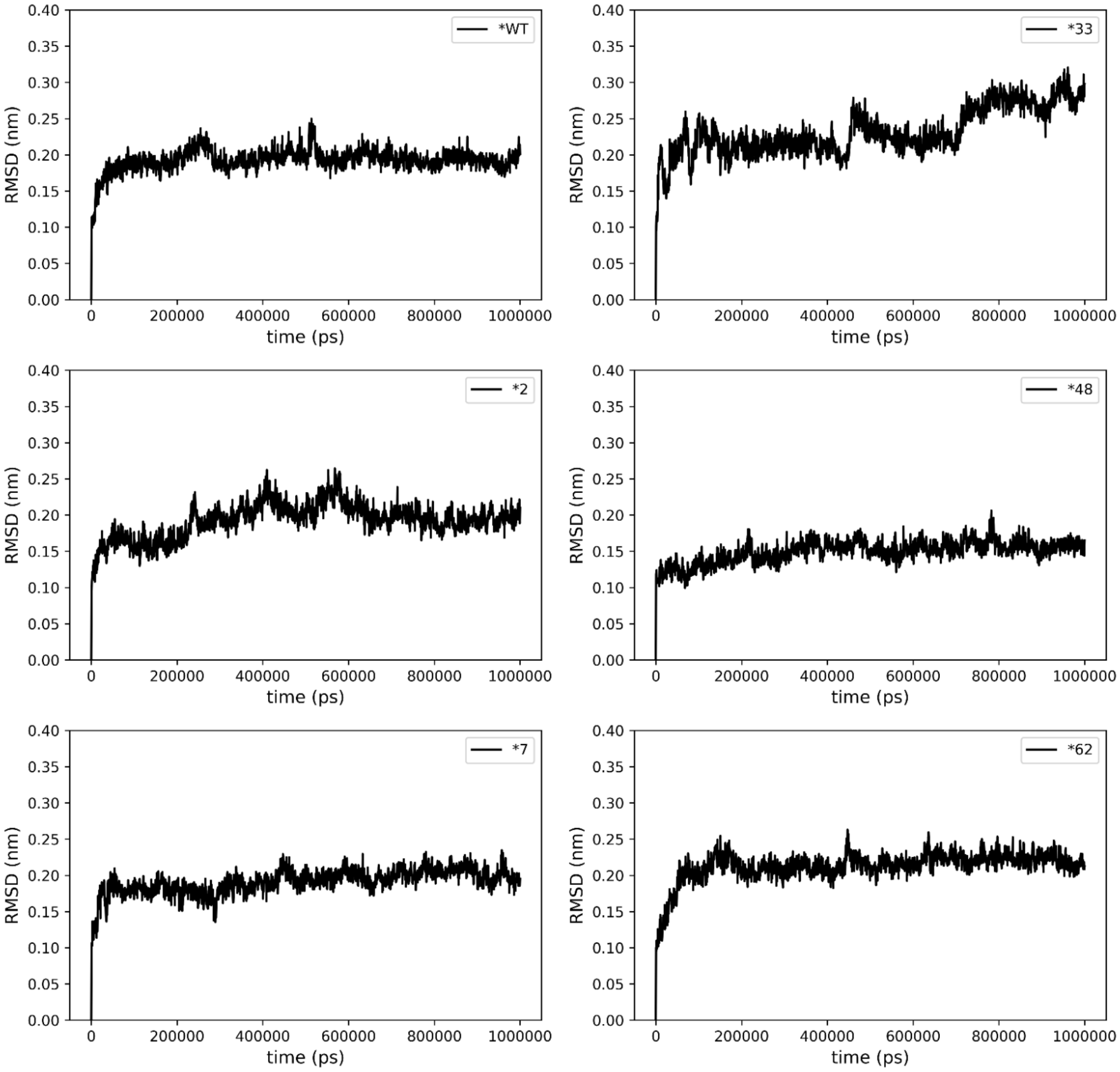
Allele	Mutation(s)	Activity
*1	wild-type	Normal
*2	R296C, S486T	Normal
*7	H324P	No
*12	G42R, R296C, S486T	No
*33	A237S	Normal
*48	A90V	Normal
*62	R441C	No
*99	P34S, R88P, S486T	No

Figure 9 CYP2D6 star alleles along with their mutations and their metabolic activity

3. Results

3.1 System convergence

The Root Mean Square Deviations (RMSDs) of the backbone atoms in functional and non-functional variants of CYP2D6 are shown in Figure 10 over the equilibrium trajectories. The initial structures for the MD trajectories were used as reference structures for the RMSD measurements. In general, simulations for the wild-type (WT) and all mutants (MT) converge before 150ns. The system with the CYP2D6 *33 variant shows rather high RMSD values compared to the other systems. RMSD of all backbone atoms, including oxygens in C-terminus, converged at 0.15 nm to 0.22 nm and the final 850 ns have been used for the analyses described below. The results of the RMSD calculations indicate that all systems were adequately equilibrated.



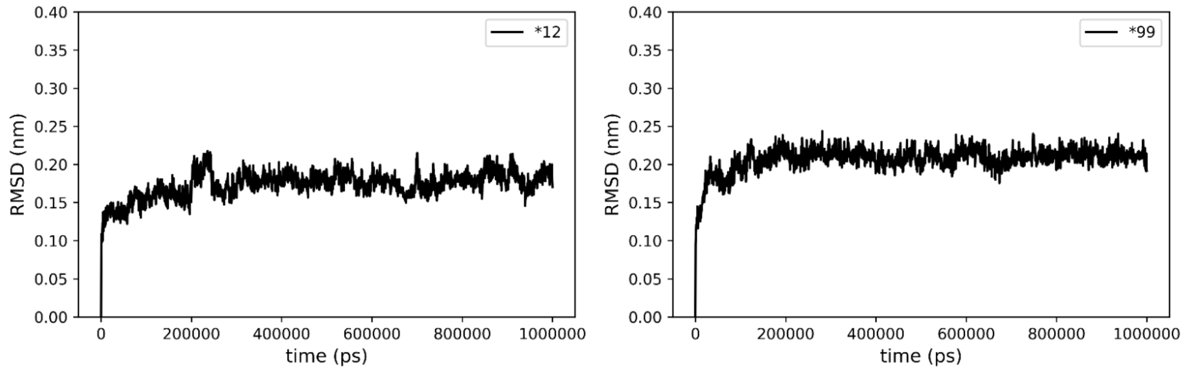
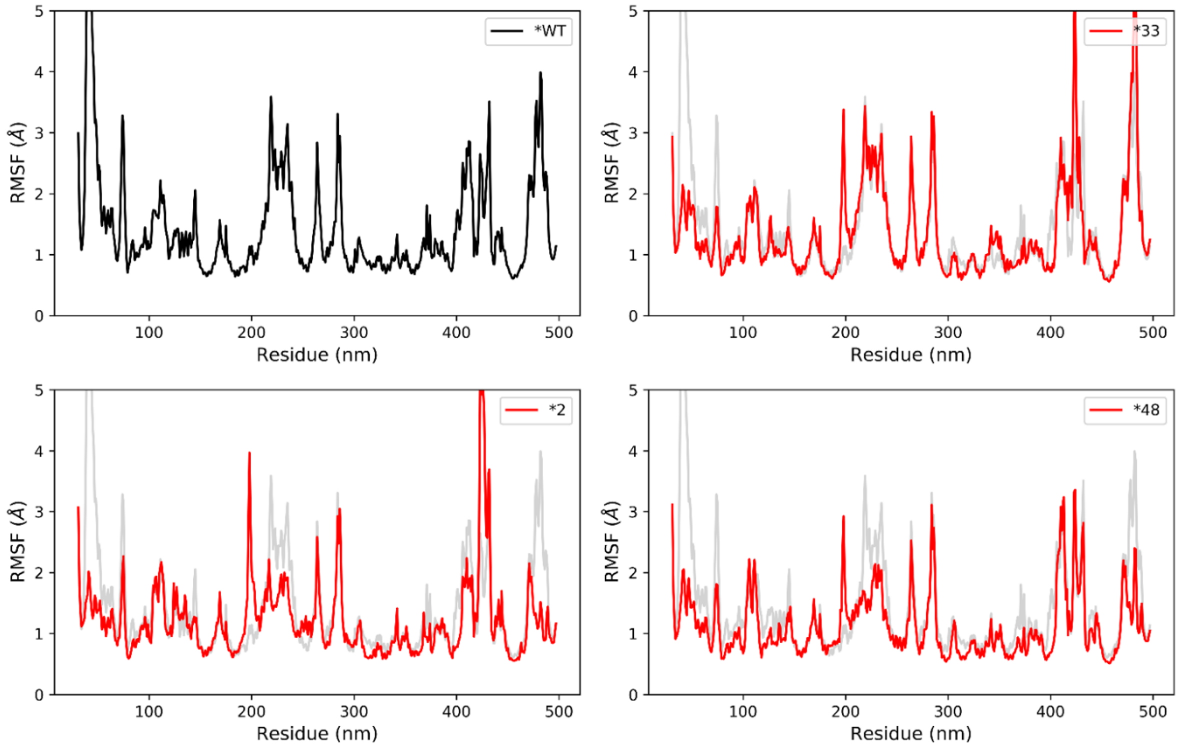


Figure 10 Root Mean Square Deviations for the functional and non-functional CYP2D6 variants.

3.2 Structural fluctuations

Root Mean Square Fluctuation (RMSF) of the residues within all the different allelic forms, functional and non-functional, are shown in Figure 11. The final 850ns of each equilibrium trajectory were used for the RMSF calculations. CYP2D6 is a membrane-anchored protein through its N-terminus, and therefore fluctuations around the N-terminus residues should be interpreted with care as simulations were performed without the membrane that would likely stabilize this site through non-bonded interactions.



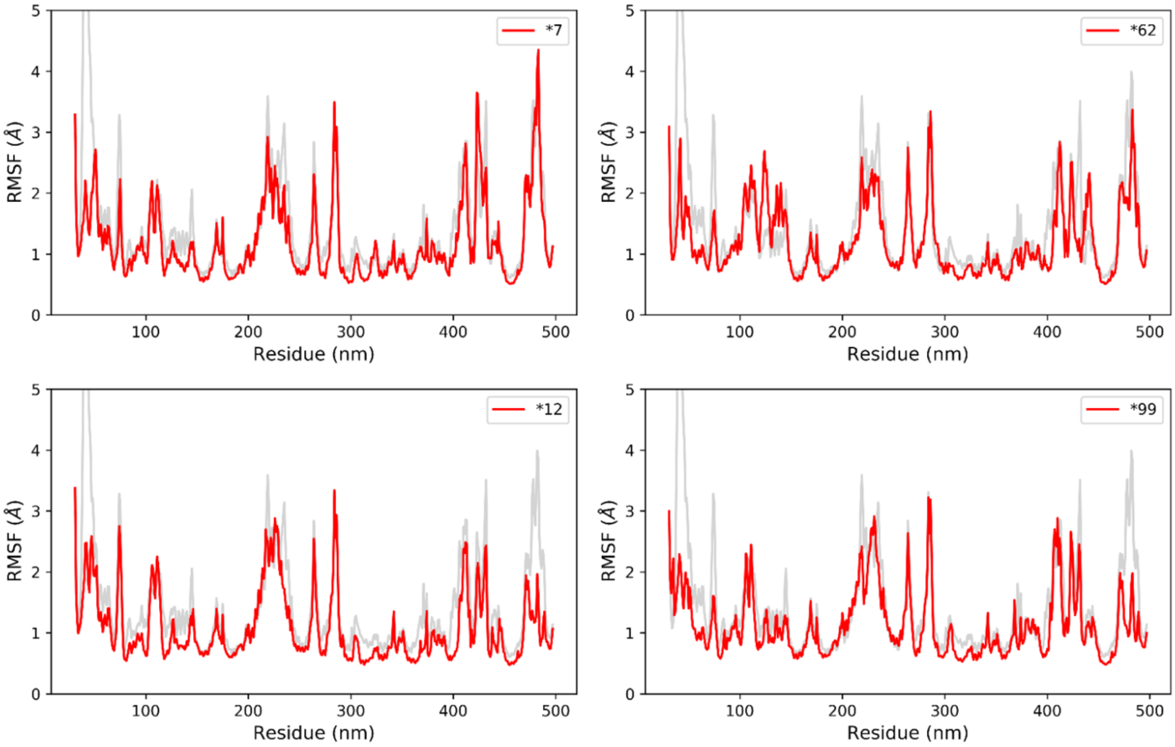


Figure 11 Root Mean Square Fluctuations of backbone atoms of functional and non-functional CYP2D6 variants.

In order to gain better insight into these fluctuations, changes between variants in Root Mean Square Fluctuations (ΔRMSF) relative to CYP2D6 *1 (wild-type) were examined as shown in Figure 12. ΔRMSFs were computed using the value per residue of the allelic form CYP2D6 *1 as a reference. Positive values of ΔRMSF correlate to atoms of decreased flexibility whereas negative values of ΔRMSF correspond to flexible atoms as compared with the reference. In general, helices D, E, I, J, K, L are of similar flexibility as the allelic form CYP2D6 *1. However, changes in plasticity were observed for the different variants compared to the wild-type. Non-functional variants seem to be more rigid than the functional ones.

For CYP2D6 *2 flexible sites were observed especially in helix F (residues 197-203) and K'-L loop (residues 422-428). Rigid sites were observed in β1 sheet (residues 37-52), loop connecting β1-1 β1-2 (residues 72-75), F-F' loop (residues 218-219), helix F' (residues 220-226), F'-G loop (residues 230-239), K'-L loop (residues 411-413) and β3 sheet(residues 476-488). CYP2D6 *33 seems to be a highly flexible variant especially in helix F (residues 195-202), K'-L loop (residues 417-428) and β3 sheet (residues 479-485). For CYP2D6 *33 rigid sites were observed in β1 sheet (residues 37-49), loop connecting β1-1 β1-2 (residues

73-76), K’ helix (residues 400-403) and K’-L loop (residues 430-433). CYP2D6 *48 flexible sites are located in helix F (residues 195-199) and especially in K’-L loop (residues 423-424). Rigid sites are located in β1 sheet (residues 37-52), loop connecting β1-1 β1-2 (residues 74-76), F-F’ loop (residues 218-219), helix F’ (residues 220-224), F’-G loop (residues 230-239), K’ helix (residues 400-401) and β3 sheet (residues 476-488).

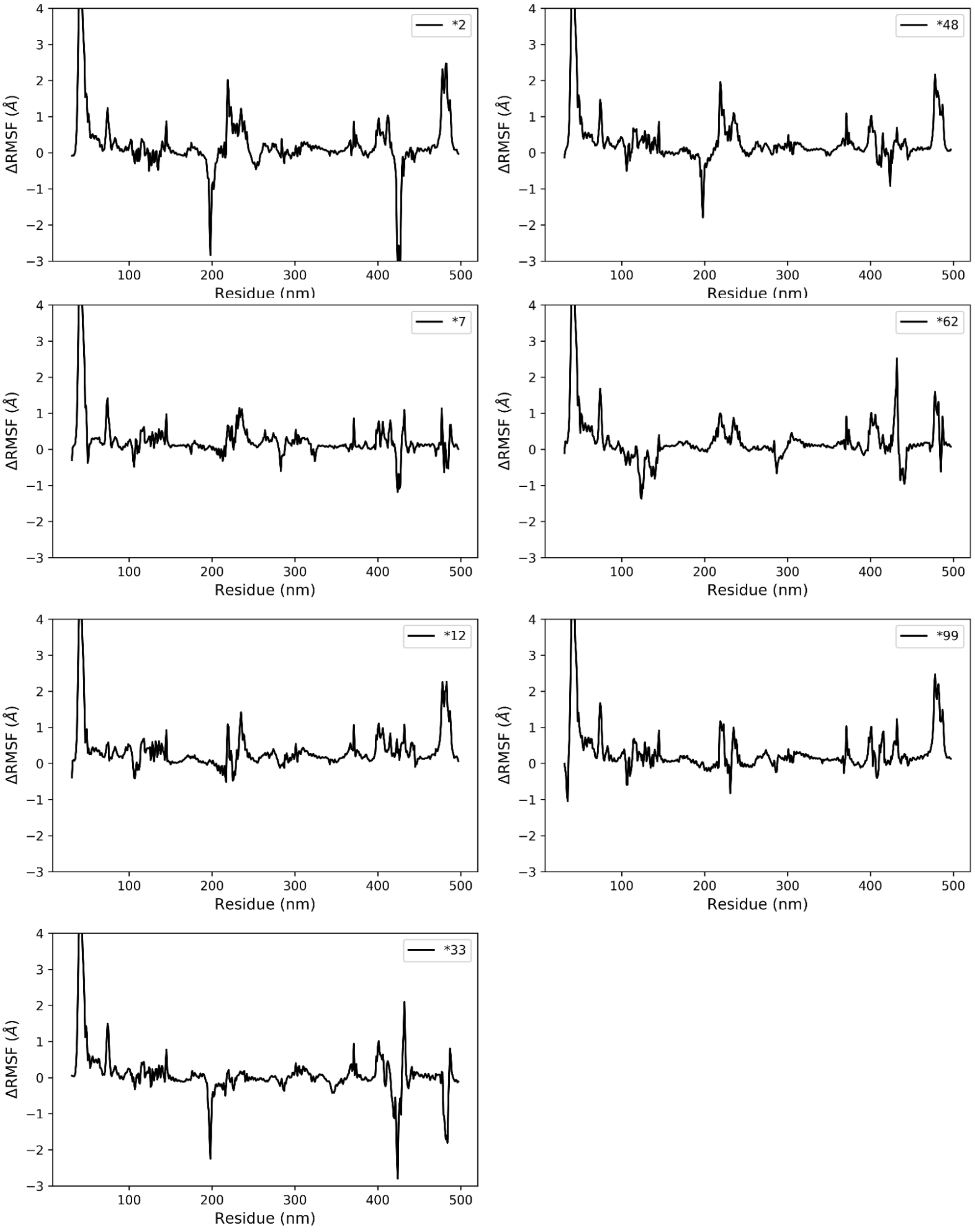


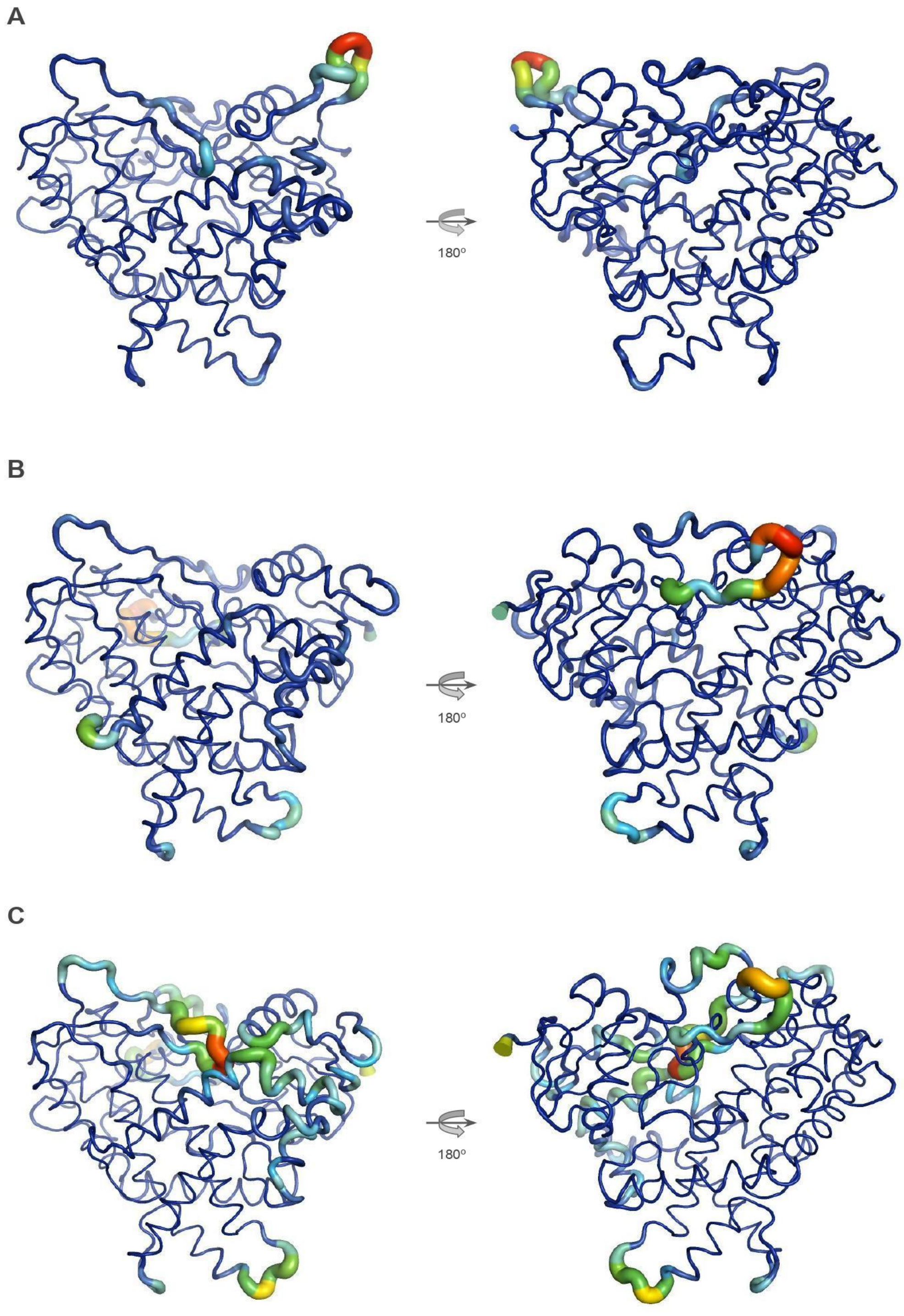
Figure 12 Changes in Root Mean Square Fluctuations (ΔRMSF) of backbone atoms of functional and non-functional CYP2D6 variants relative to CYP2D6 *1.

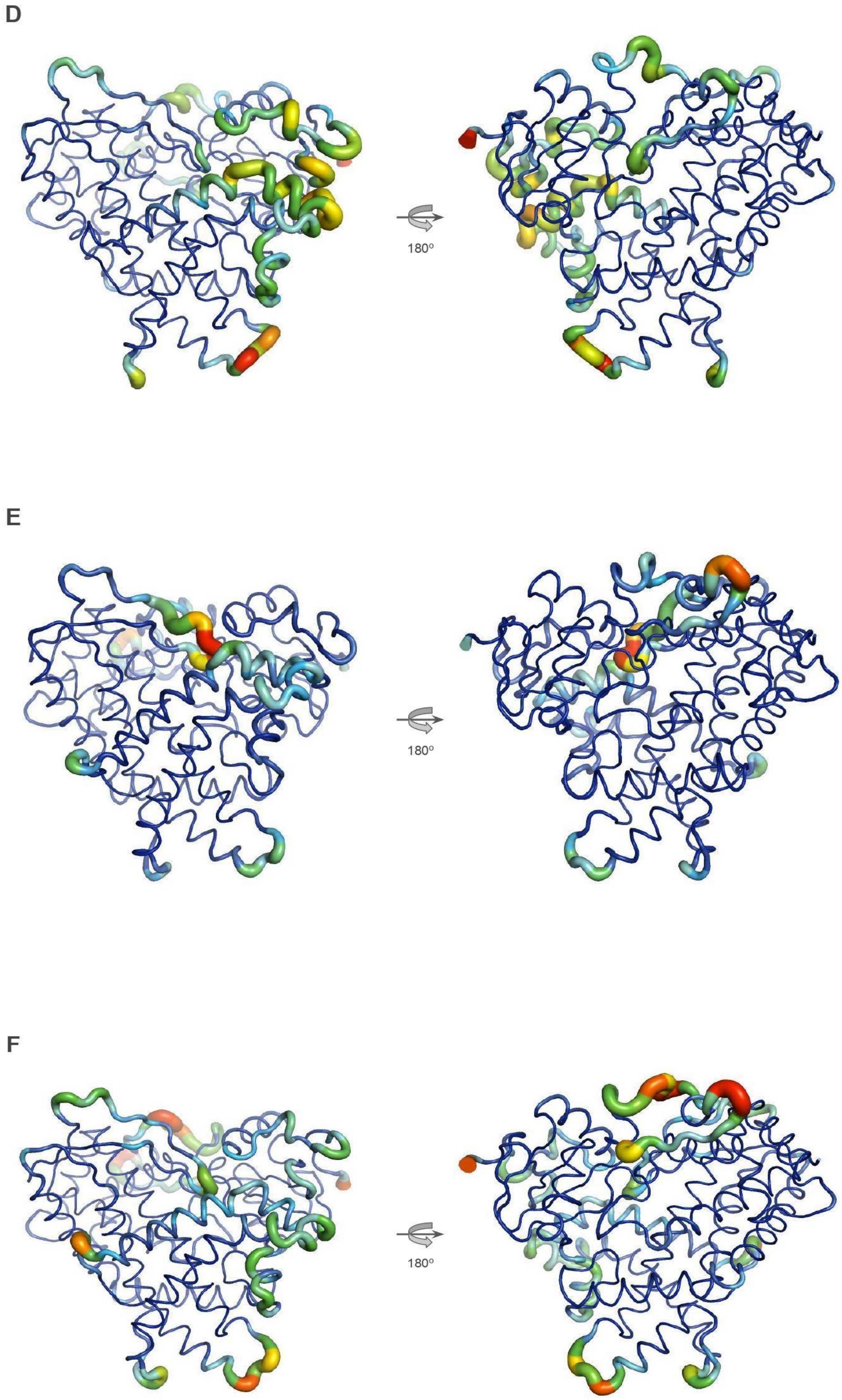
As far as the non-functional variants are concerned, for CYP2D6 *7 rigid sites were observed especially in β1 sheet (residues 37-49), loop connecting β1-1 β1-2 (residues 72-76), C-D loop (residue 144-145), F-F’ loop (residue 219), helix F’ (residues 220 and 224), F’-G loop (residues 229-238), K’ helix (residues 400-402), K’-L loop (residues 415-416 and 431-432) and β3 sheet (residues 487-488). Flexible sites in CYP2D6 *7 were observed in K’-L loop (residues 423-427) and in the β3 sheet (residues 483-485). For CYP2D6 *12 rigid sites were observed in β1 sheet (residues 37-49), B’-C loop (residues 116-118), F-F’ loop (residues 218-219), F’-G loop (residues 233-239), K-K’ loop (residues 398-400), helix K’ (residues 401-407), K’-L loop (residues 414-415) and β3 sheet (residues 475-488). No notably rigid sites were detected in CYP2D6 *12. For CYP2D6 *62 rigid sites were observed in β1 sheet (residues 37-52), loop connecting β1-1 β1-2 (residues 53-76), F-F’ loop (residues 217-219), helix F’ (residues 220-223), F’-G loop (residues 234-239), K’ helix (residues 400-407), K’-L loop (residues 428-433) and in β3 sheet (residues 477-482). Flexible sites for this allelic form were detected in B’-C loop (residues 120-126), helix C (residues 136-140) and in K’-L loop (residues 436-442). Last but not least, for CYP2D6 *99 rigid sites were observed mainly in β1 sheet (residues 37-52), loop connecting β1-1 β1-2 (residues 56-76), B’-C loop (residues 114-115), F-F’ loop (residues 218-219), helix F’ (residues 220-223), F’-G loop (residues 234-238), K-K’ loop (residues 398-400), helix K’ (residues 401-402), K’-L loop (residues 414-416), K’-L loop (residues 429-433) and in β3 sheet (residues 475-488). No notably rigid sites were identified in CYP2D6 *99.

Overall, helices D and E and in particular residues 170-183 are more rigid in non-functional variants than in the wild-type and the functional variants. Also, residues in the meander loop, between helices K’ and L, are slightly more rigid in all non-functional variants than in functional. The different combinations in flexibility and rigidity may influence the function and metabolizer phenotype of the enzyme, as they can be directly associated with the effective migration of the drug towards the active site.

We cannot draw a general conclusion about mobility of the different allelic forms relative to CYP2D6 *1. For this reason, the previous diagrams were translated into B-factors and they were mapped on the structure of CYP2D6 *1. This was done by populating the B-Factor column on the pdb files for the eight different allelic forms in order to investigate the structural fluctuations of each CYP2D6 variant, as shown in Figure 13. The color scheme indicates the degree of fluctuation, going from blue/violet indicating little fluctuation to yellow/green indicating intermediate fluctuation and red indicating large fluctuation. Comparing the dynamics of CYP2D6 at different allelic forms, it can be observed that the

CYP2D6 functional variants are more stable than CYP2D6 non-functional variants. We could conclude that CYP2D6 variants with normal function and especially CYP2D6 *1 and *2 remain in a more stable configuration. More specifically, F’-G loop (around 230th residue) along with H-I loop (around 280th residue) and β3-sheet (around 280th residue) seem to be less stable in non-functional variants than in functional ones. This might be connected with the easiness of the drug accessibility to the active sites.




D
180º
E
180º
F
180º

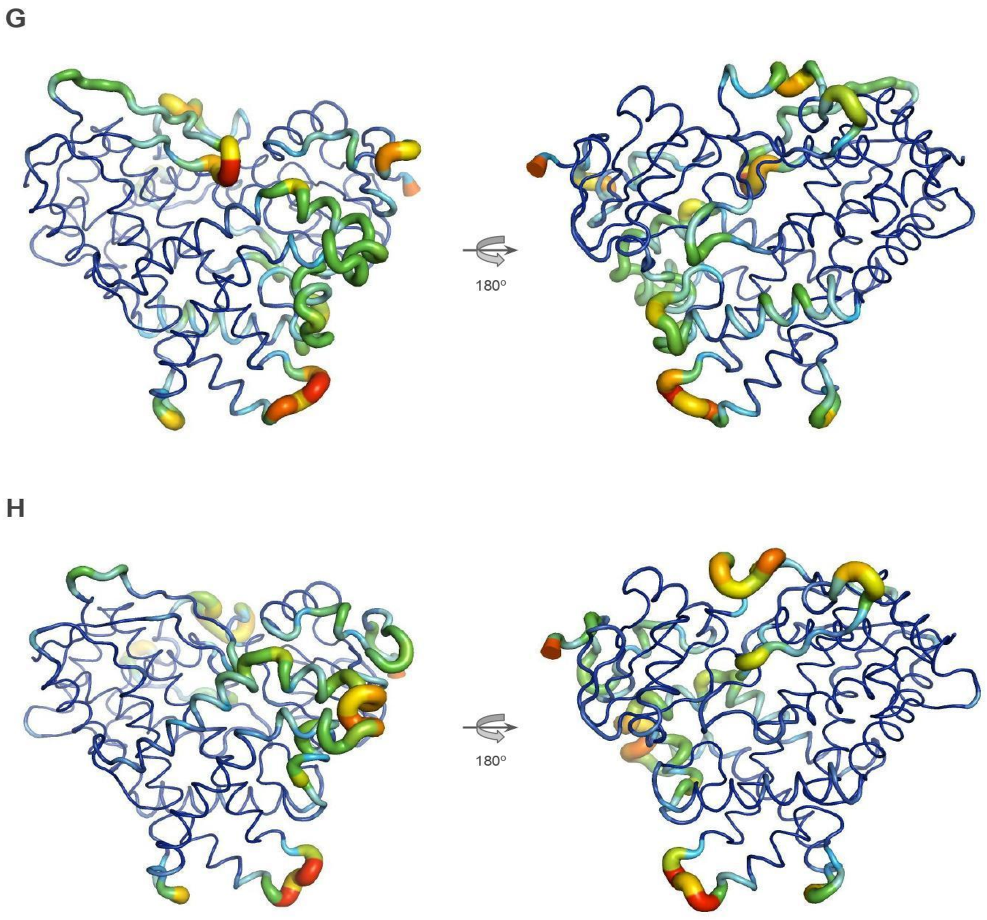
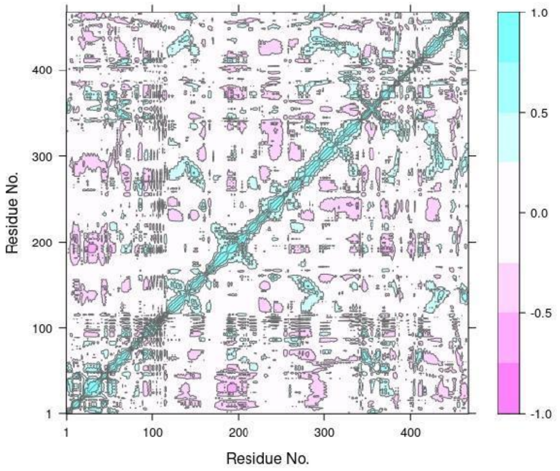


Figure 13 Structural fluctuations of the various CYP2D6 alleles (A) CYP2D6 *1, (B) CYP2D6 *2, (C) CYP2D6 *7, (D) CYP2D6 *12, (E) CYP2D6 *33, (F) CYP2D6 *48, (G) CYP2D6 *62, (H) CYP2D6 *99.

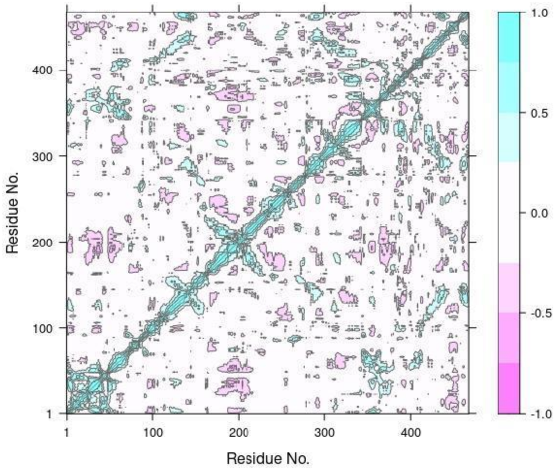
3.3 Dynamic cross correlation analysis

In order to understand the extent to which the atomic fluctuations of a system are correlated, a dynamic cross correlation analysis was performed using the Bio3D package as shown in Figure 14 and Figure 15 (Grant et al. 2006). Cross correlation between the i-th and j-th atoms is represented by Cij, which ranges from −1 to +1. The $C\alpha$ atoms for CYP2D6 *1, *2, *7, *12, *33, *62, *99 were used to compute the cross correlation ~~Cij~~ matrix (see Figure 14). A positive value represents the correlated motion, and a negative value represents the anti-correlated motion. In general, functional variants present more strong correlations between their residues compared to non-functional variants. The amount of correlated and especially the amount of anti-correlated motions for non-functional variants decreased significantly compared with that of the functional variants.

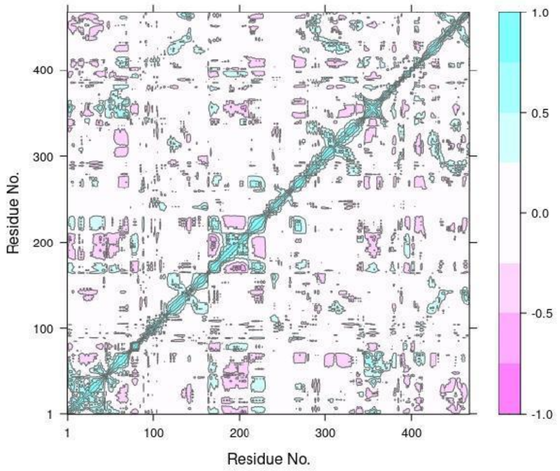
CYP2D6 *1



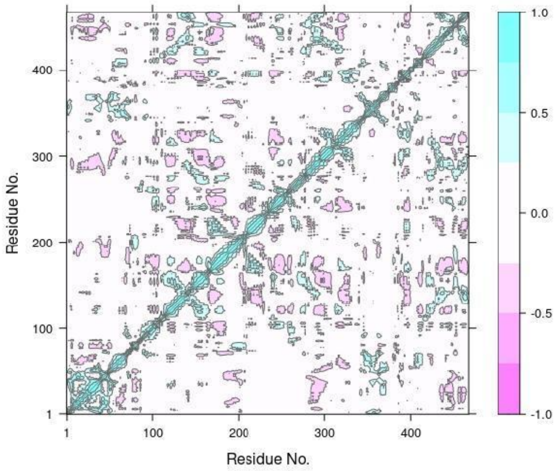
CYP2D6 *33



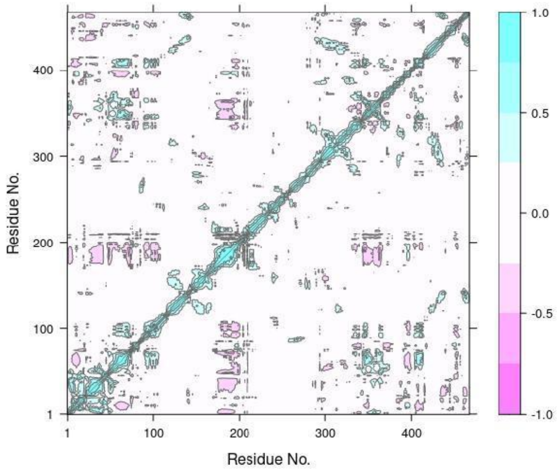
CYP2D6 *2



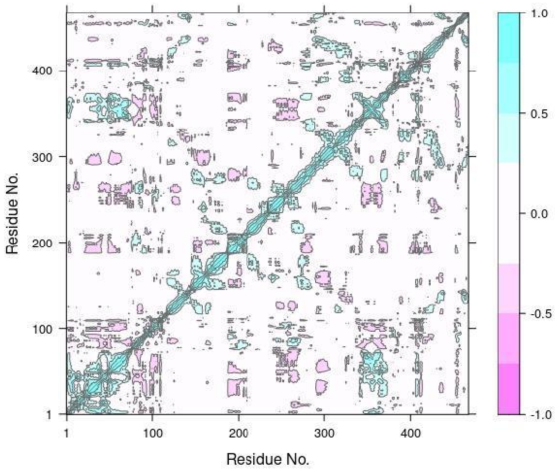
CYP2D6 *48



CYP2D6 *7



CYP2D6 *62



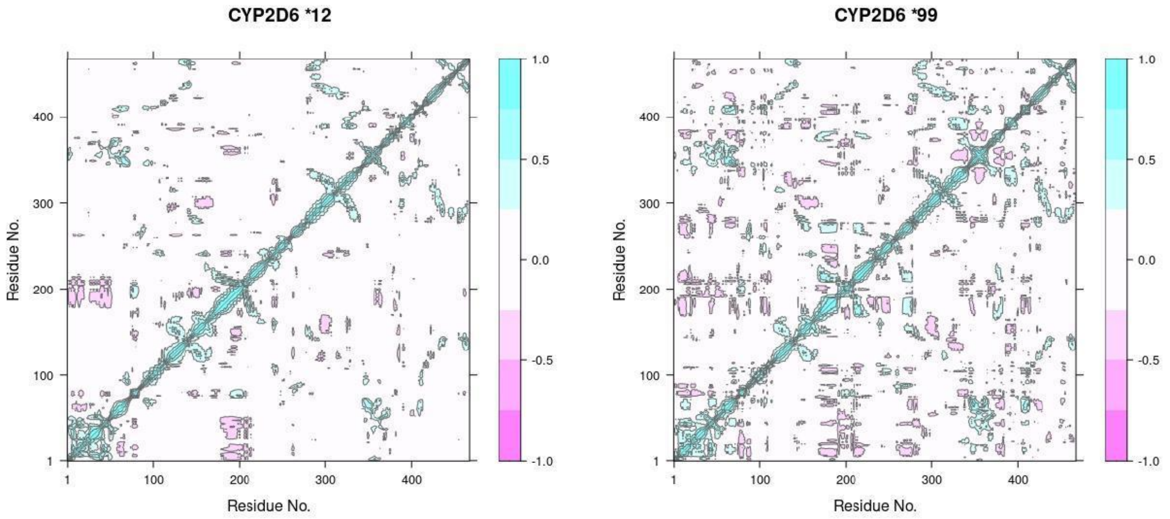
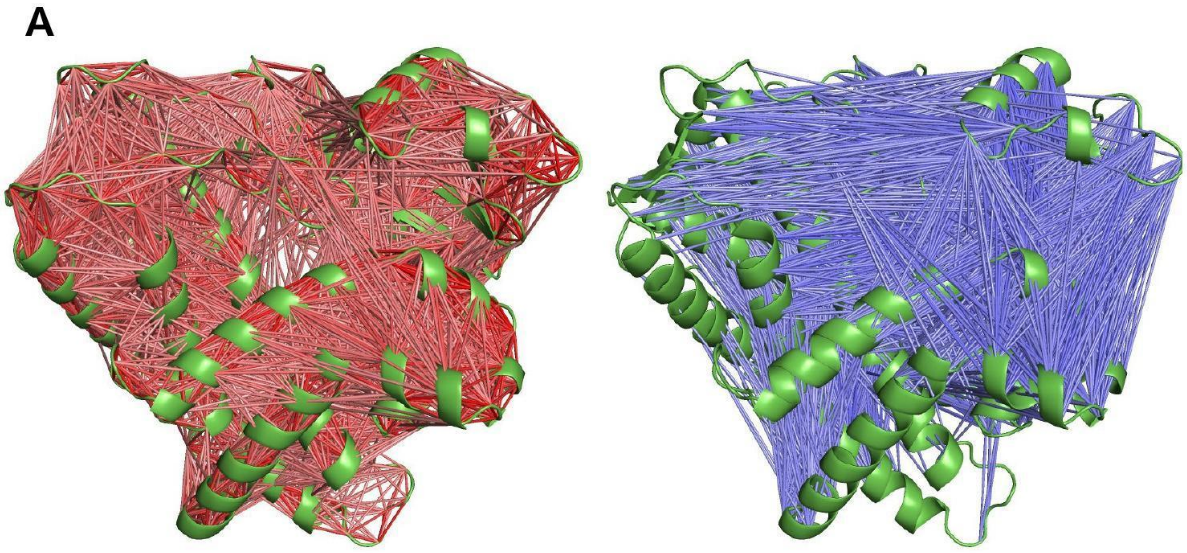
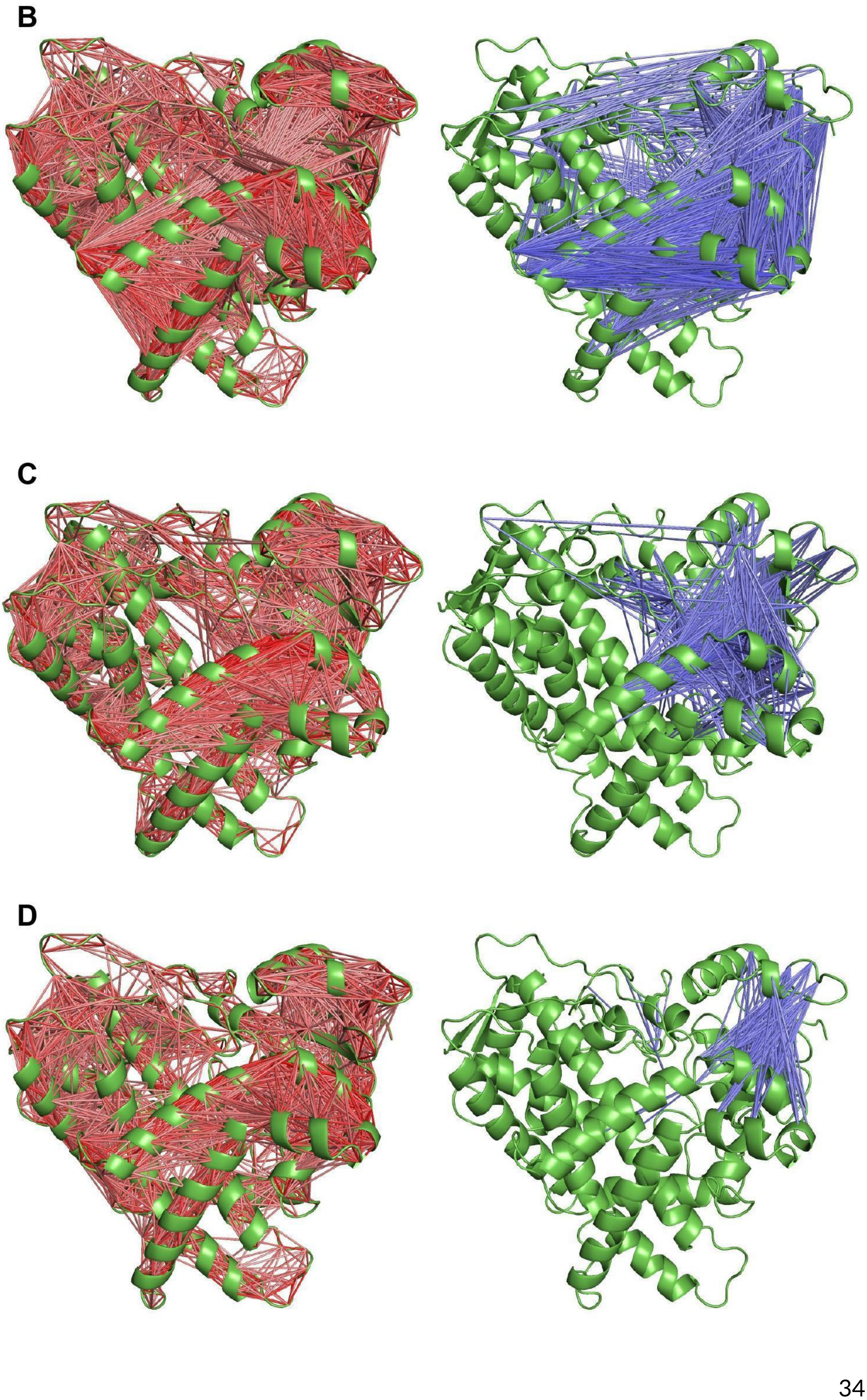
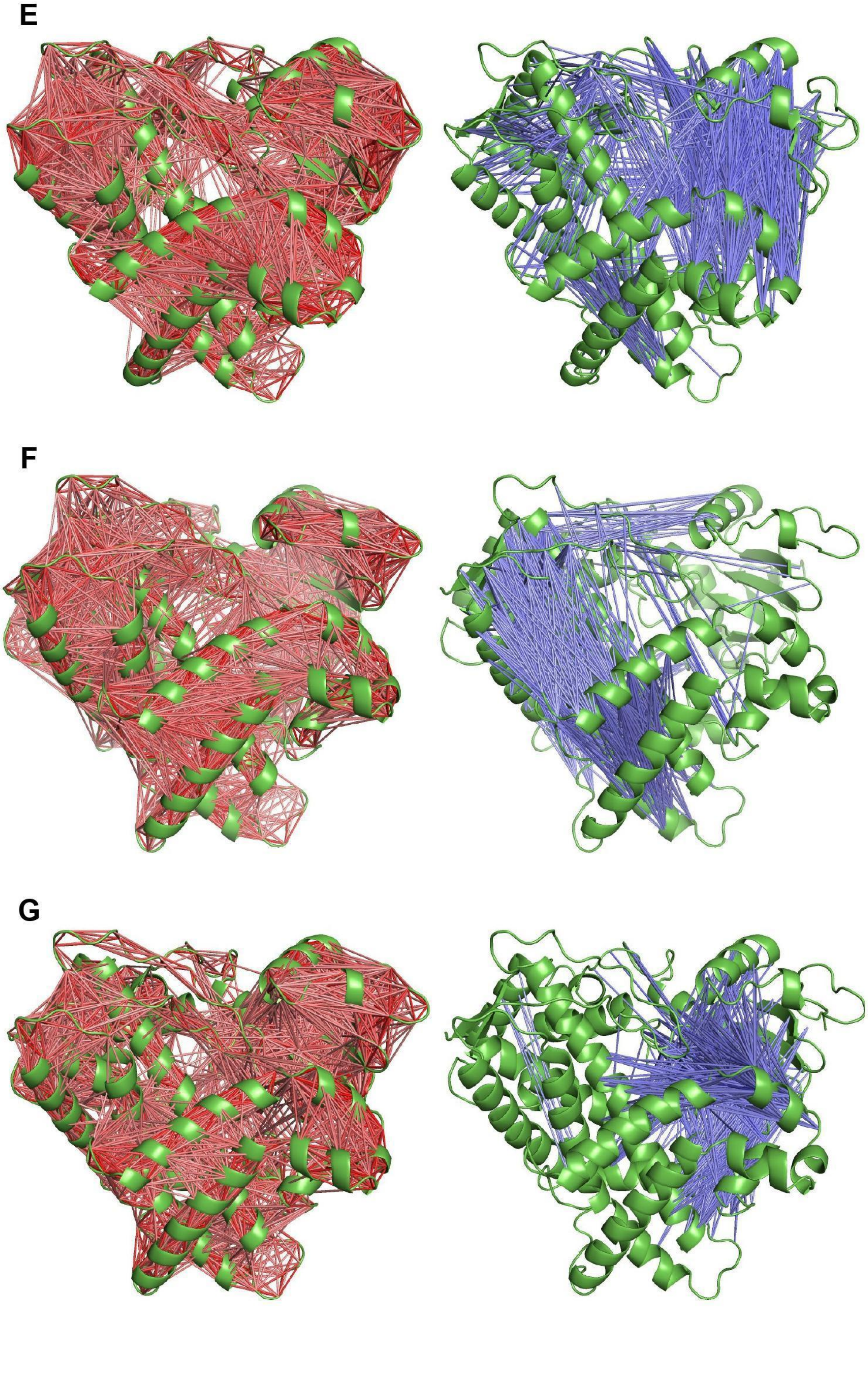


Figure 14 Residue-residue cross-correlations of the various CYP2D6 alleles. Red and blue lines indicate correlated and anti-correlated motions respectively.

In general, in functional variants and in particular in CYP2D6 *1 *2 and *33, correlated and anti-correlated motions were observed throughout the protein. In case of CYP2D6 *48, which also has normal function, these anticorrelated motions are more limited and are observed mostly between helix I and β1-sheet, C-D loop, helix D, E-F loop, helix F and helix G and between the beginning of helix F and B'-C loop, helix D, E-F loop, helix H and β3-sheet. In the non-functional variants, anticorrelated motions are even more limited and are observed mostly between F-F' loop/helix F and β1-sheet, B-B' loop, B'-C loop and K-K' loop.




B
C
D


E
F
G

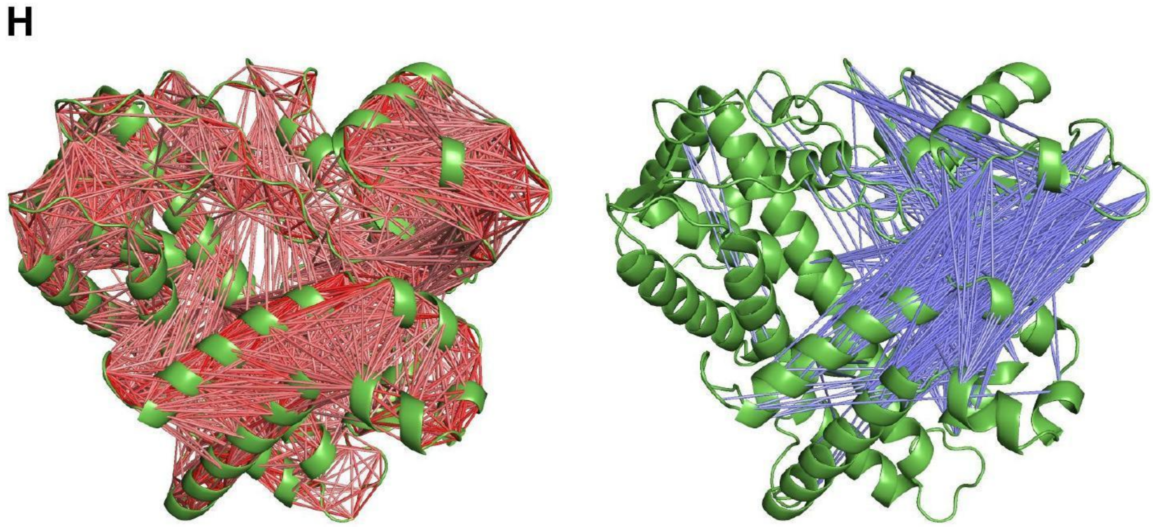


Figure 15 Cross-correlation matrix of Cα atoms of the various CYP2D6 alleles (A) CYP2D6 *1, (B) CYP2D6 *2, (C) CYP2D6 *7, (D) CYP2D6 *12, (E) CYP2D6 *33, (F) CYP2D6 *48, (G) CYP2D6 *62, (H) CYP2D6 *99. Red and blue lines indicate correlated and anti-correlated motions respectively.

According to the above results, the loss of the anti-correlation relationships developed by helix I in functional variants and their replacement by relationships developed by helix F’/F-F loop, in combination with the loss of correlation relationships can influence the stability and functionality of the enzyme. The crucial sites that participate in functional and non-functional anti-correlated motions are shown in Figure 16. Probably, these changes result in the closure of access and egress channels.

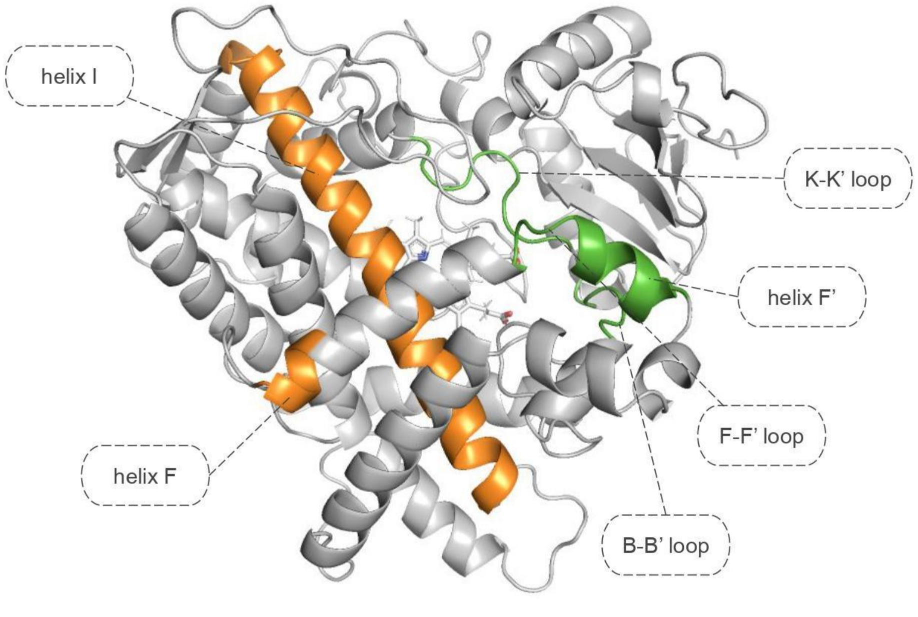


Figure 16 The sites that develop anti-correlation relationships and seem to play a key role in functional (orange) and non-functional (green) variants.

3.4 Molecular Docking

Molecular docking simulations have been used in order to probe the potential interactions between CYP2D6 and drug molecules. BACE1 inhibitor was selected for molecular docking as a representative drug. Docking was performed on key conformations (time-median configuration of populous ensembles) out of the classical MD trajectories of each variant produced by clustering.

This particular drug molecule, the BACE1 inhibitor, was chosen for the docking analyses because it has been co-crystalized in the CYP2D6 structure employed for the MD simulations. In the crystal structure the molecule is placed in the active site as shown in Figure 17, at a distance of 4.6 Å from the iron atom of the heme group. The selection of this molecule allows for a direct comparison of the docking results with that of the experimentally determined orientation of the drug in the structure.

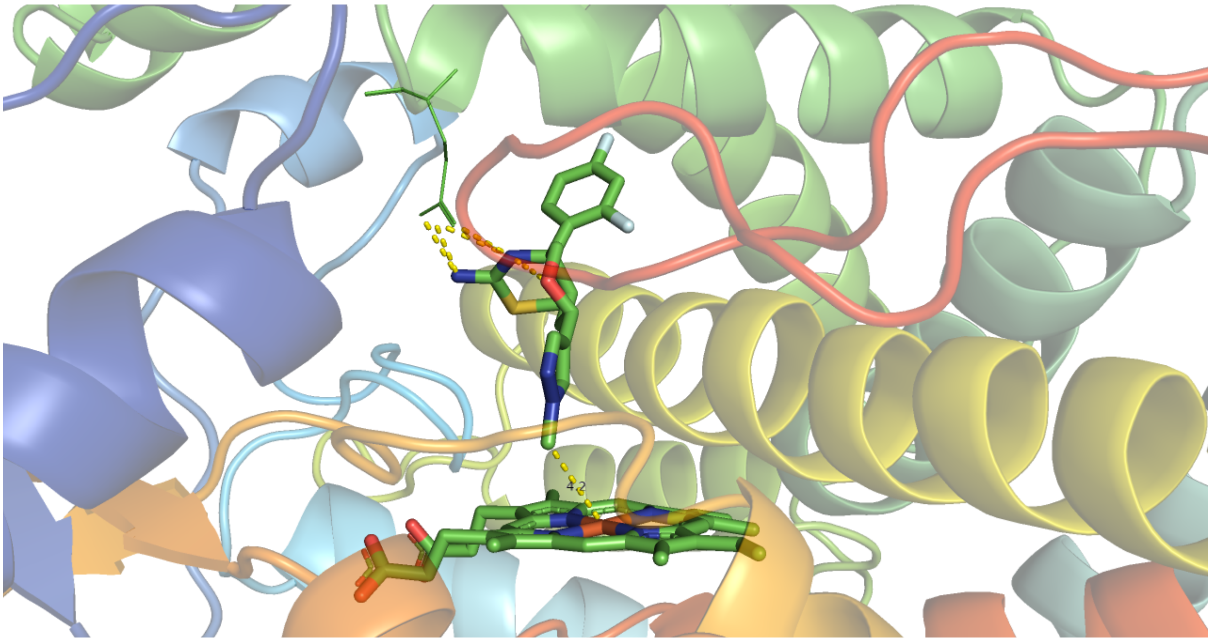


Figure 17 The heme-BACE1 inhibitor complex in crystal structure (PDB entry: 4XRY)

In CYP2D6 *1, six representative structures have been produced by clustering. In all representative structures, the ligand approaches the active site and the distance between ligand and the iron atom of the heme is similar to that of the crystal structure as shown in Figure 18.

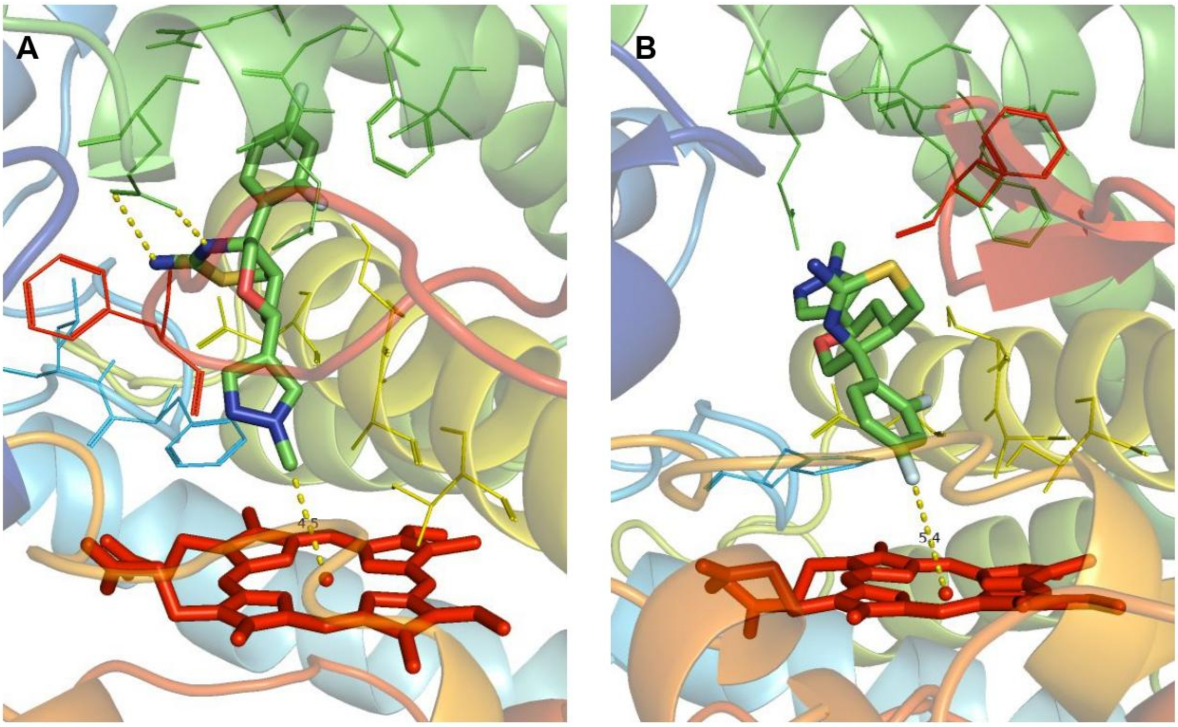


Figure 18 Results of Molecular Docking in the first (A) and in the second (B) representative structure of CYP2D6 *1

Similarly, in the case of CYP2D6 *2, the distance between the heme iron and the substrate is even smaller than that of the crystal structure. In addition, the substrate appears to interact with various residues within the active site. Two of the major representative structures along with the ligand are shown in Figure 19.

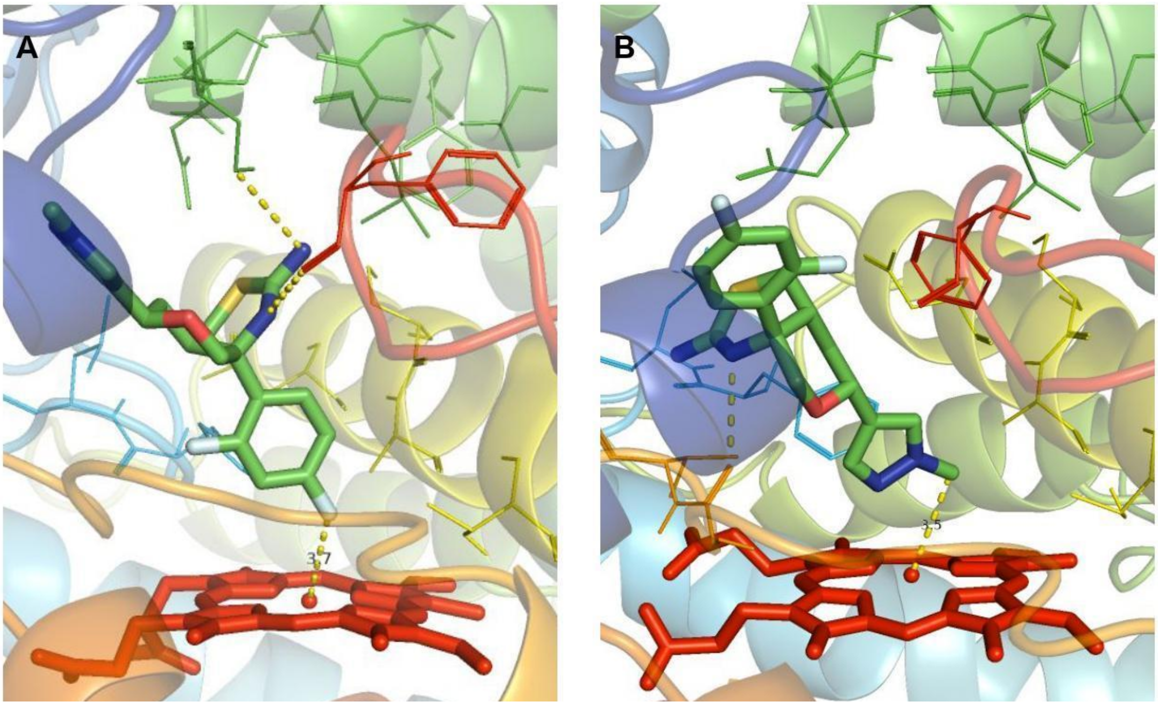


Figure 19 Results of Molecular Docking in the first (A) and in the second (B) representative structure of CYP2D6 *2

To the contrary, a change in the ligand binding orientation was observed in the case of the non-functional CYP2D6 * 7, compared to the functional variants. In this case, the heme and the substrate are way apart making their interaction impossible. This is shown in Figure 20. The catalytic site seems to be non-accessible by the ligand.

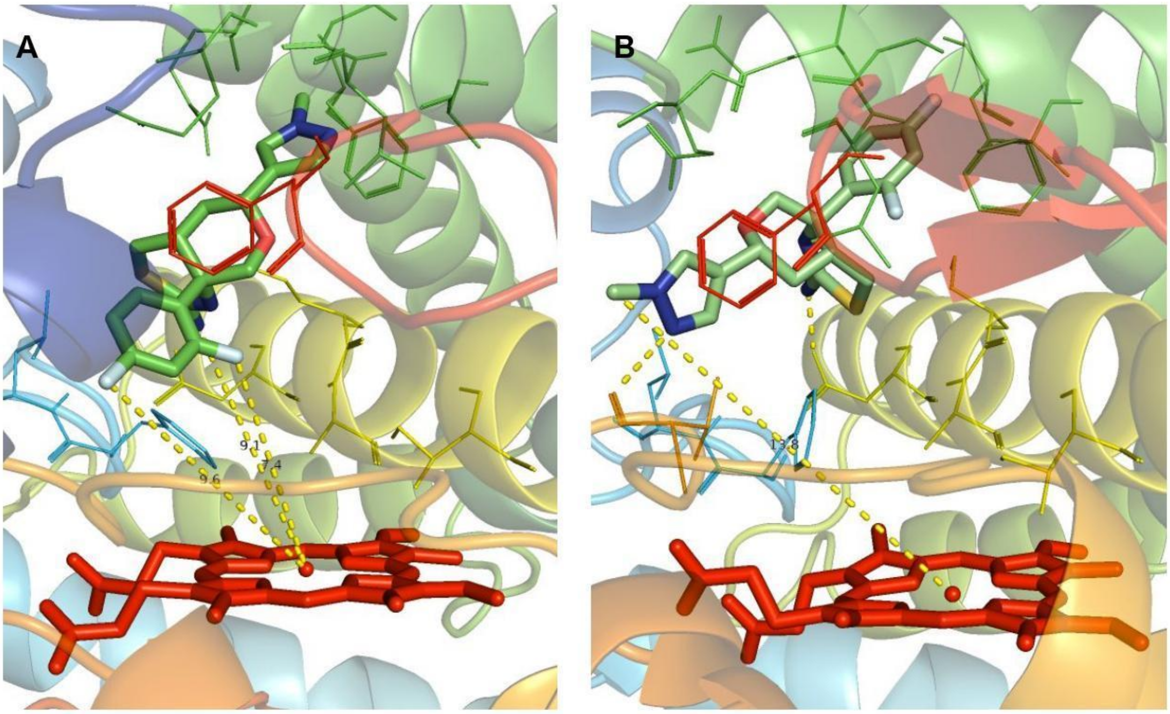


Figure 20 Results of Molecular Docking in the first (A) and in the second (B) representative structure of CYP2D6 *7

Similarly, to CYP2D6 *7, in case of the non-functional CYP2D6 *12, the substrate does not approach the catalytic site of the enzyme. Interestingly, the heme group interacts with helix I forming a hydrogen bond with Ala305 as shown in Figure 21C. This probably impairs the accessibility to the active site.

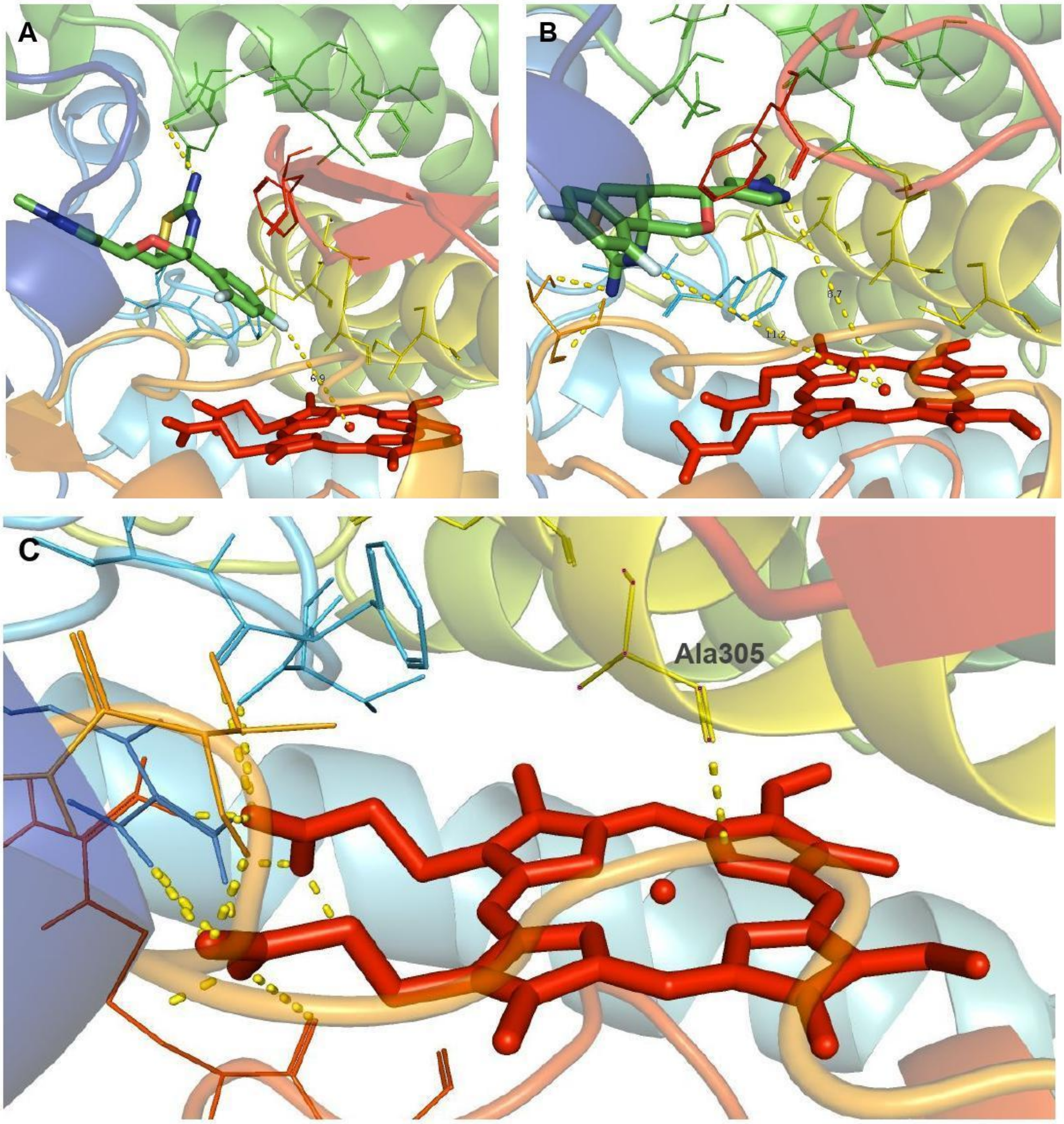


Figure 21 Results of Molecular Docking in the first (A) and in the second (B) representative structure of CYP2D6 *12 and (C) interaction between the heme group and the amino acid Ala305

For the functional variant CYP2D6 *33, seven representative structures have been produced by clustering. Figure 22 shows the distance between substrate and iron atom of the first two representative structures.

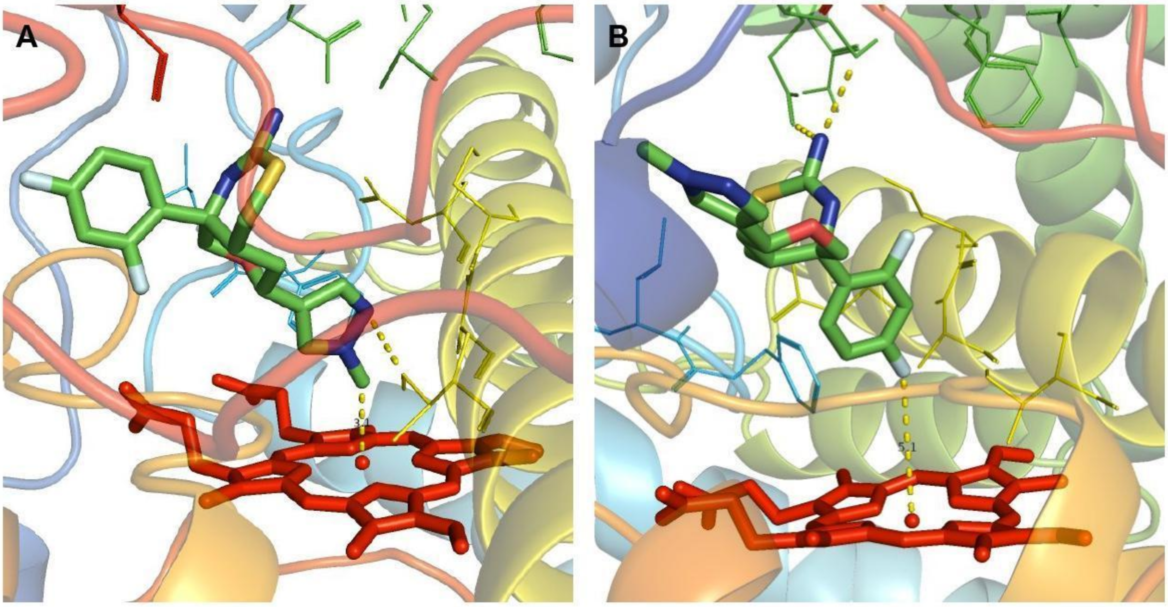


Figure 22 Results of Molecular Docking in the first (A) and in the second (B) representative structure of CYP2D6 *33

Although the results of CYP2D6 *48 for the first representative structure do not agree with the results of the other functional variants, in the second structure the substrate is close to the heme (Figure 23). We have to note that the different representative structures employed herein refer to the median structures of populous ensembles of structures. This indicates that in vivo we might have different configurations of the same CYP2D6 variant, that can interchange between functional and non-functional states. The percentage of each configuration present in vivo might relate to factors not probed herein that can shift the equilibrium towards the functional, or non-functional forms. Thus, the identification of even a single variant configuration that efficiently binds the reference drug could justify a functional variant, as it is the case for the reverse trend (ineffective binding, non-functional form). However, to draw solid conclusions, a more elaborate technique has to be employed (e.g. binding free energy simulations) to weigh the population of the different configurations of the variant and their importance in the vivo functionality.

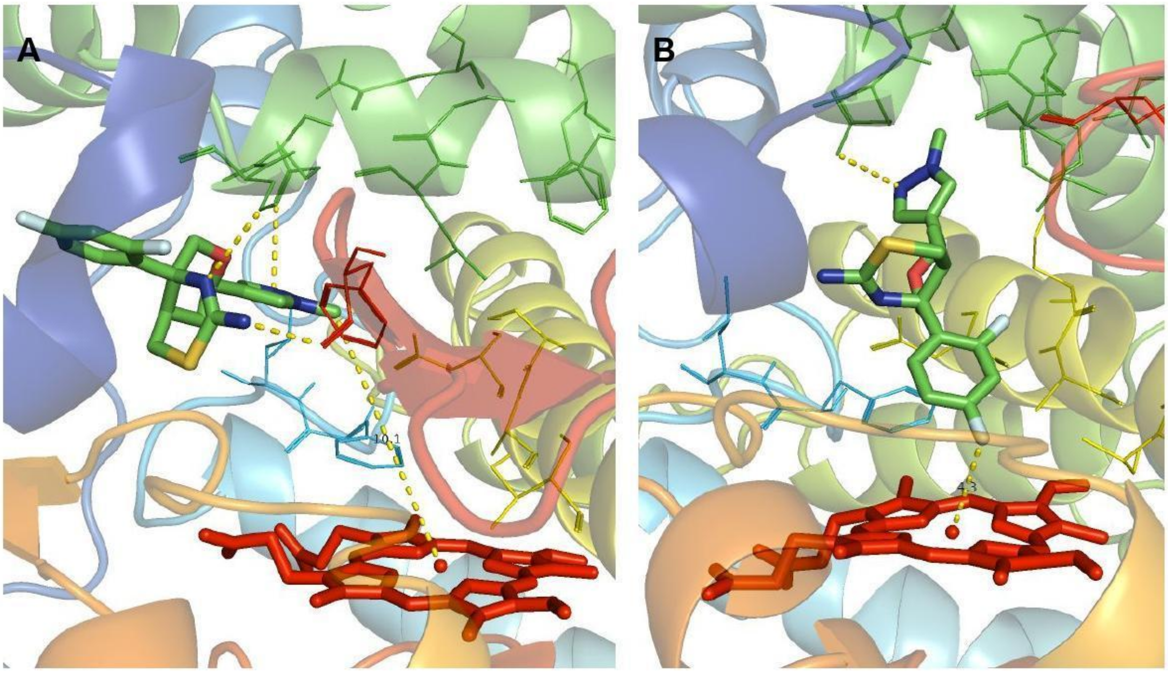


Figure 23 Results of Molecular Docking in the first (A) and in the second (B) representative structure of CYP2D6 *48

As far as the non-functional variant CYP2D6 *62 is concerned, on the one hand, in the first representative structure the ligand is very close to the heme which would favor the interaction between them. On the other hand, in the second representative structure the ligand is located at the plane of the heme group, which normally would not allow the drug molecule to interact with the iron atom. These results are shown in Figure 24.

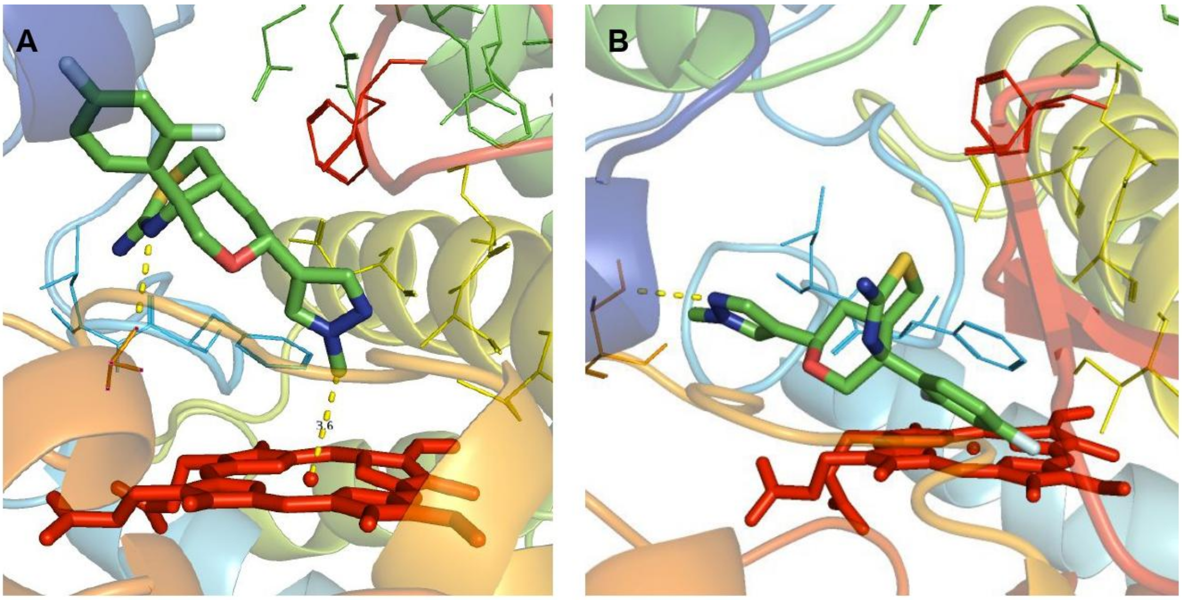


Figure 24 Results of Molecular Docking in the first (A) and in the second (B) representative structure of CYP2D6 *62

Last but not least, in the case of CYP2D6 *99, the distance between the ligand is large enough not to allow the approach to the catalytic site in all of the representative structures (Figure 25). The K-K' loop seems to obstruct the substrate from entering the active site.

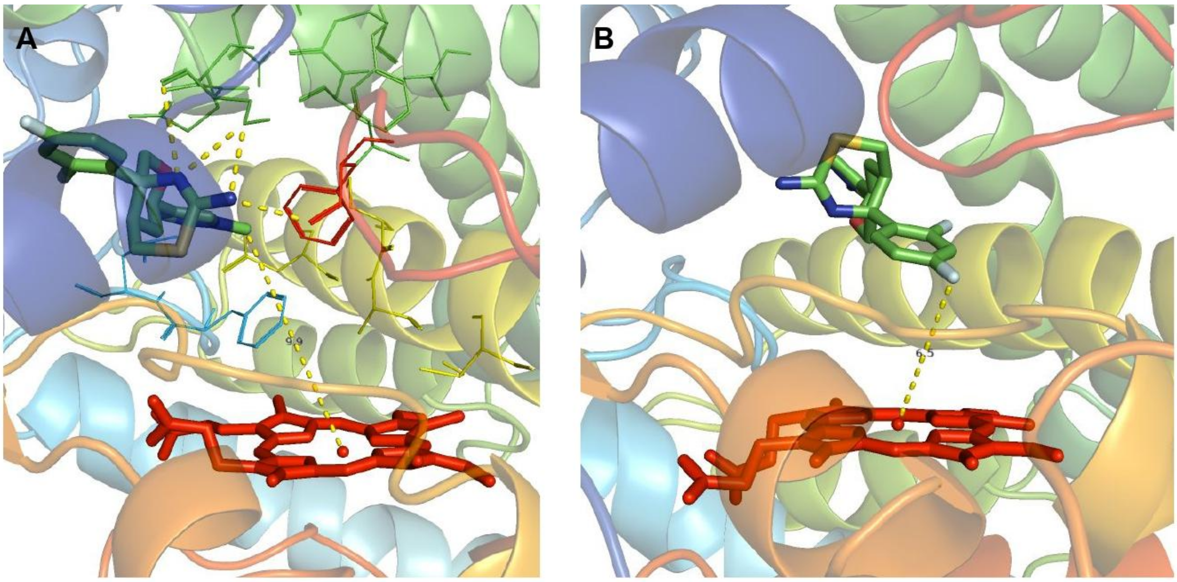
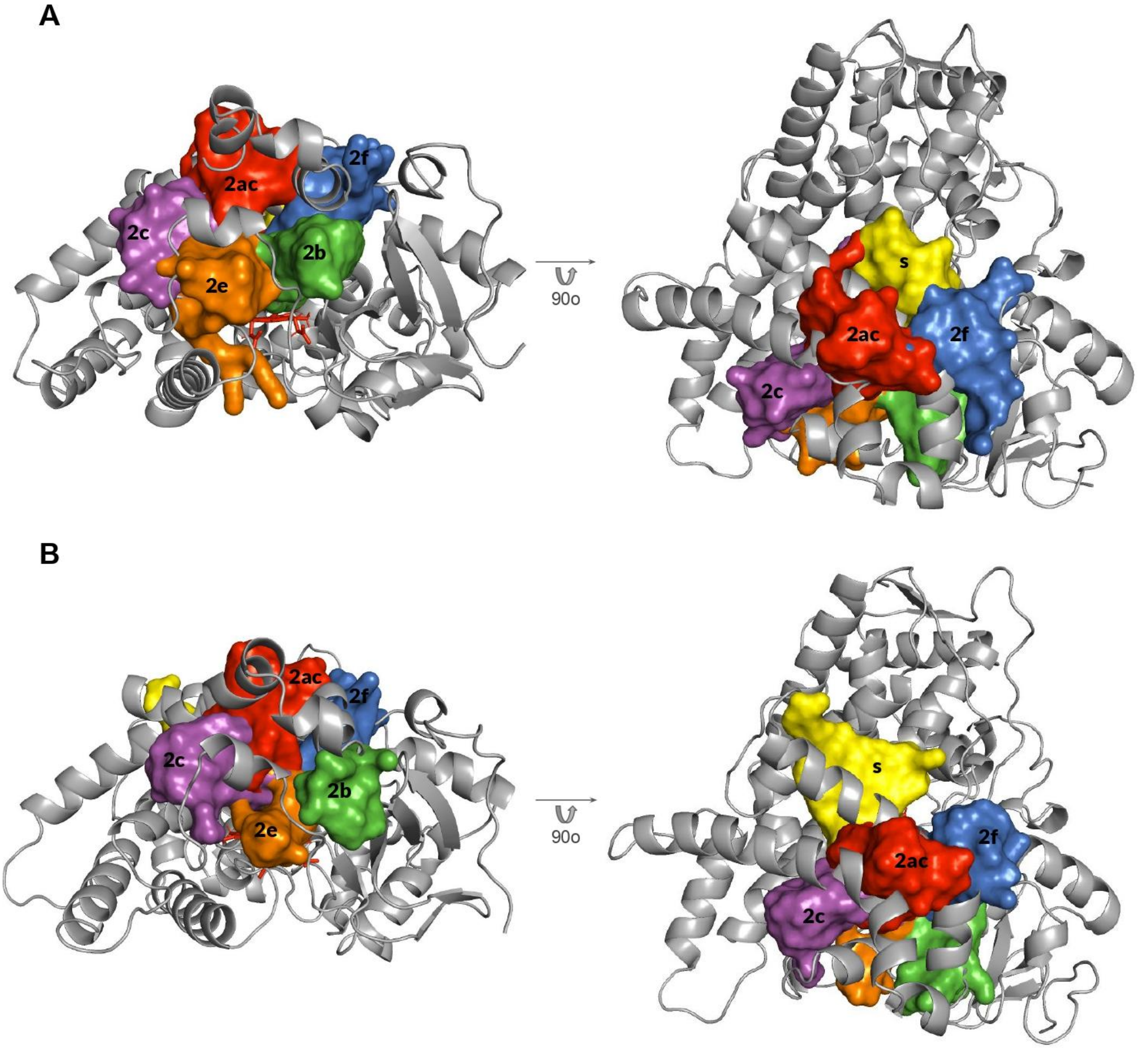


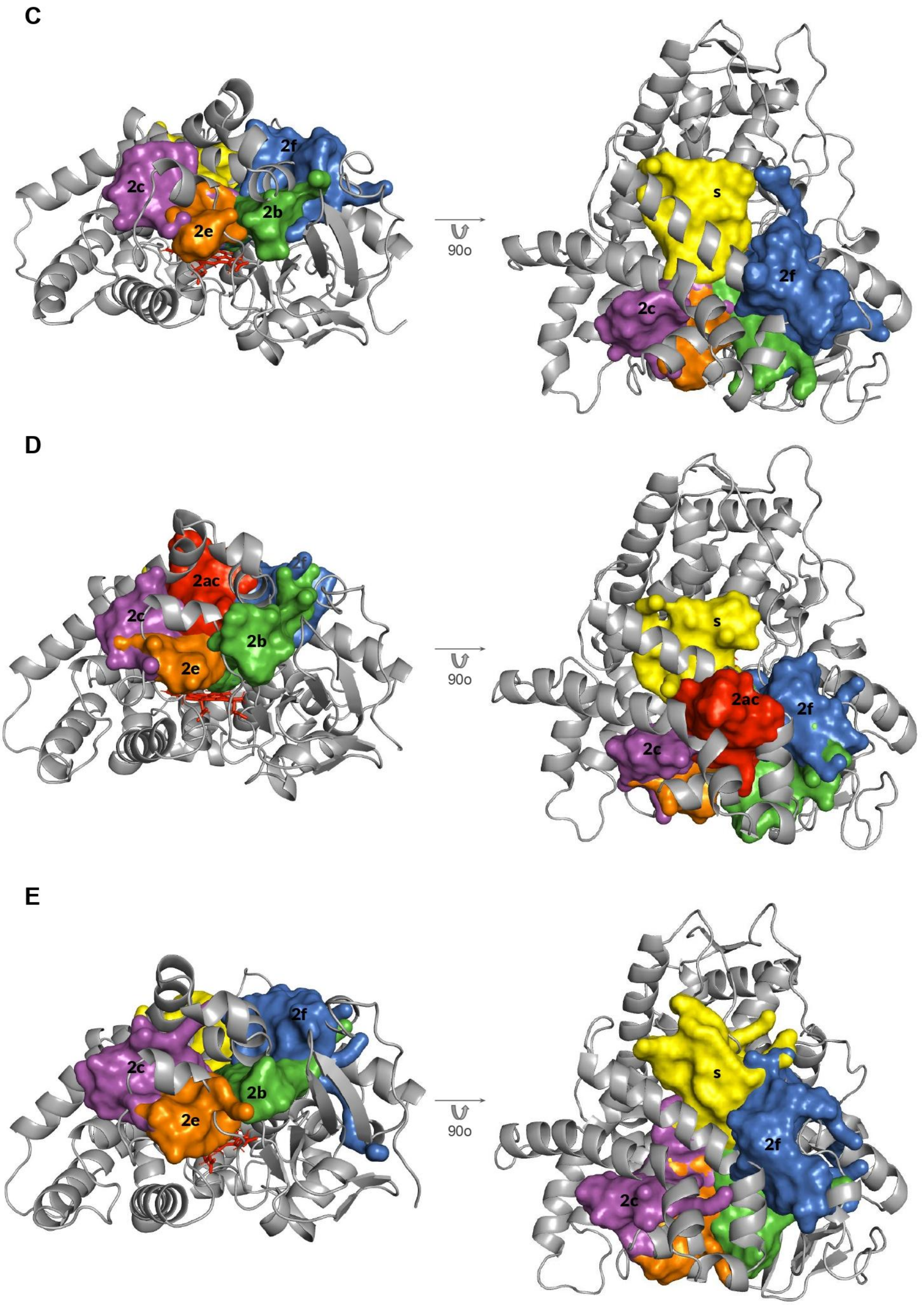
Figure 25 Results of Molecular Docking in the first (A) and in the second (B) representative structure of CYP2D6 *99

In general, in the majority of non-functional variants, the substrate appears to be unable to approach the iron atom of the heme and the distance between substrate and the heme group is greater than that observed in the crystal structure. This is not observed in the case of the functional ones. In allele *7 and *12, the role of helix I and K-K' loop respectively in the accessibility of the enzyme was observed. The I helix was positioned directly above the heme, not allowing its interaction with the ligand, while the K-K' loop prevented the substrate from entering the active site. The previous analyses have also highlighted the importance of these sites.

3.5 Analysis of access and egress tunnels

The active site of the enzyme is located deep in the protein next to the heme. Polymorphisms can affect the formation of access and egress channels between the surface and protein surface. In order to study how the overall motion of the protein affects the opening and the dimensions of these channels, we analysed the presence and the bottleneck radius of channels in each of the selected allelic forms over the course of the produced simulations. For the eight variants, six classes of tunnels were identified (2b, 2f, 2e, 2c, 2ac and s). The channel nomenclature we followed is based on a series of published papers (Lüdemann, Lounnas, and Wade 2000; Schleinkofer et al. 2005; Cojocaru, Winn, and Wade 2007). Tunnels 2b and 2e were found near to the B-B′ loop, with 2b opening on the side close to β-sheet and 2e in the middle of the B-C loop region. Tunnel 2c was found between the B-C loop and helix I. Tunnel s (solvent tunnel) was found between helices F and I; tunnel 2ac was located between the B′ helix and F′-G loop, and tunnel 2f was between F-F′ loop and A helix. The highest ranked pathways in the allelic variants were subclasses of channel 2. All of the above tunnels are present in all the models except for 2ac which is absent in CYP2D6 *7, *33 and *62, as shown in Figure 26, and for this reason no comparison was made between the variants for this tunnel.




C
2f
2c
2b
2e
90o
s
2f
2c
D
2ac
2c
2b
2e
90o
s
2ac
2f
2c
E
2f
2c
2b
2e
90o
s
2f
2c

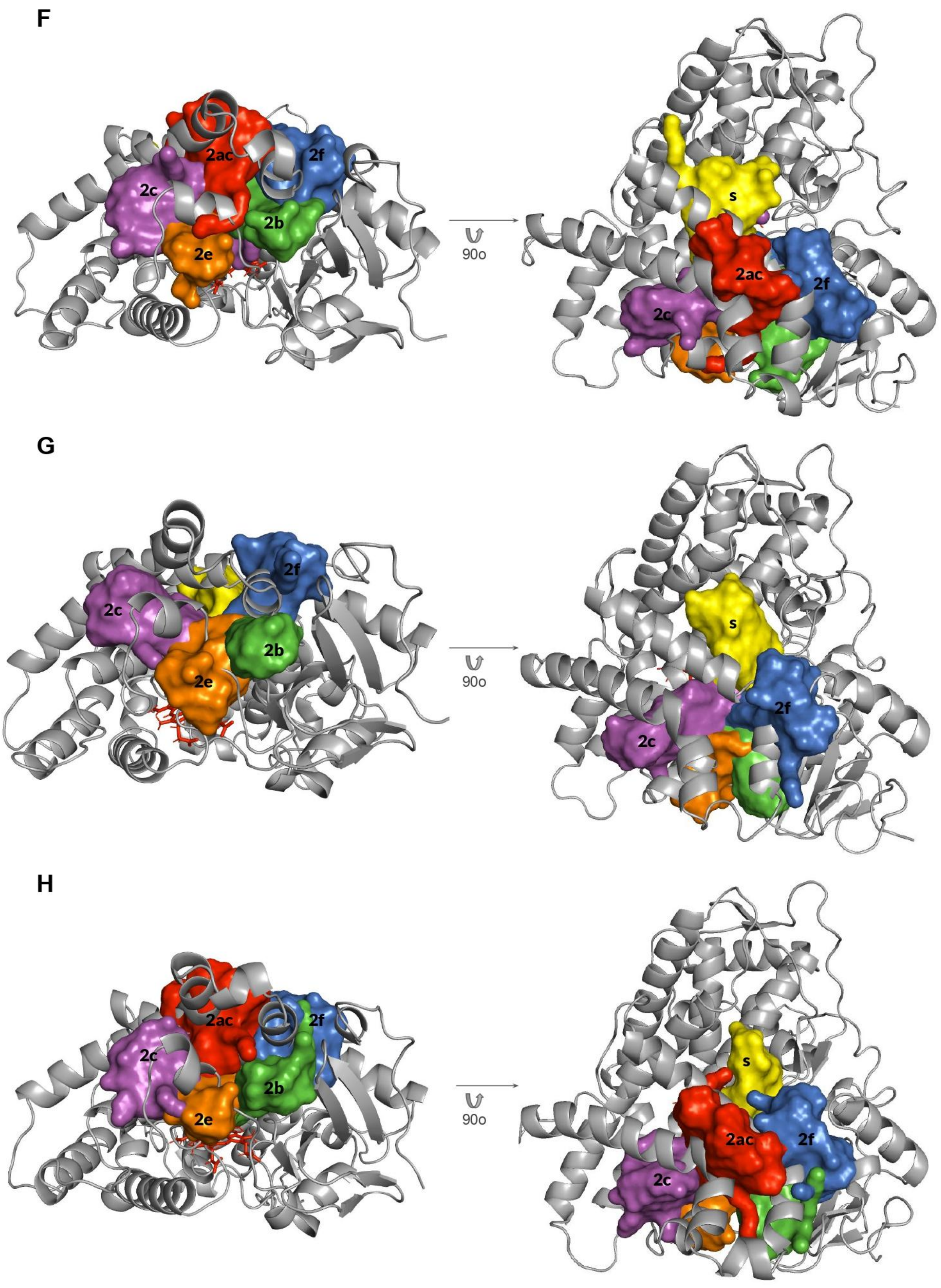


Figure 26 Major tunnels identified in the various CYP2D6 alleles (A) CYP2D6 *1, (B) CYP2D6 *2, (C) CYP2D6 *7, (D) CYP2D6 *12, (E) CYP2D6 *33, (F) CYP2D6 *48, (G) CYP2D6 *62, (H) CYP2D6 *99. The channels shown are 2b (green), 2f (blue), 2e (orange), 2c (magenta), 2ac (red) and s (yellow). Heme is shown in red sticks

Although each channel has different rates of occurrence in each variant, it was observed that channels 2b and 2f are the major pathways for all allelic forms. The top ranked channels for each variant were: 2b in *1 *7 *12 *48 *62 and *99, 2f in *2 and s in *33. Study of the time evolution of the bottleneck radius of channels 2b, 2c, 2e, 2f and solvent in each CYP2D6 allelic form during simulations was performed as shown in Figure 27. Channel 2b which is the major channel for the majority of the variants, is more open and for longer time intervals in case of CYP2D6 *1 than in CYP2D6 *7 and *12 where time intervals with low bottleneck radius were observed. The 2c channel is the dominant tunnel of CYP2D6 *2 and seems to be most open in CYP2D6 *2 as compared to the other variants. The solvent channel, which is not one of the major channels in the majority of the variants, is almost absent in the non-functional variant CYP2D6 *99. CYP2D6 *62 shows no major changes in relation to functional variants. The loss of activity in this variant may be due to the fact that the mutation occurs in an amino acid that interacts and forms a hydrogen bond with heme. With this mutation the orientation of the heme may changes and although the active site is accessible, the ligand may not be able to bind.

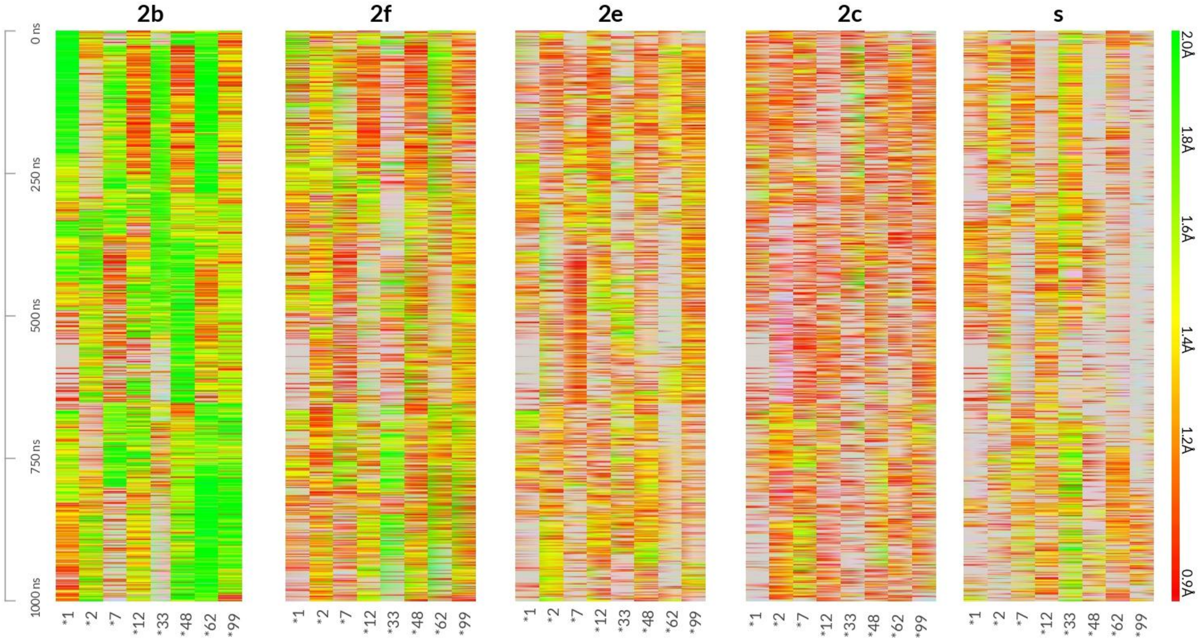
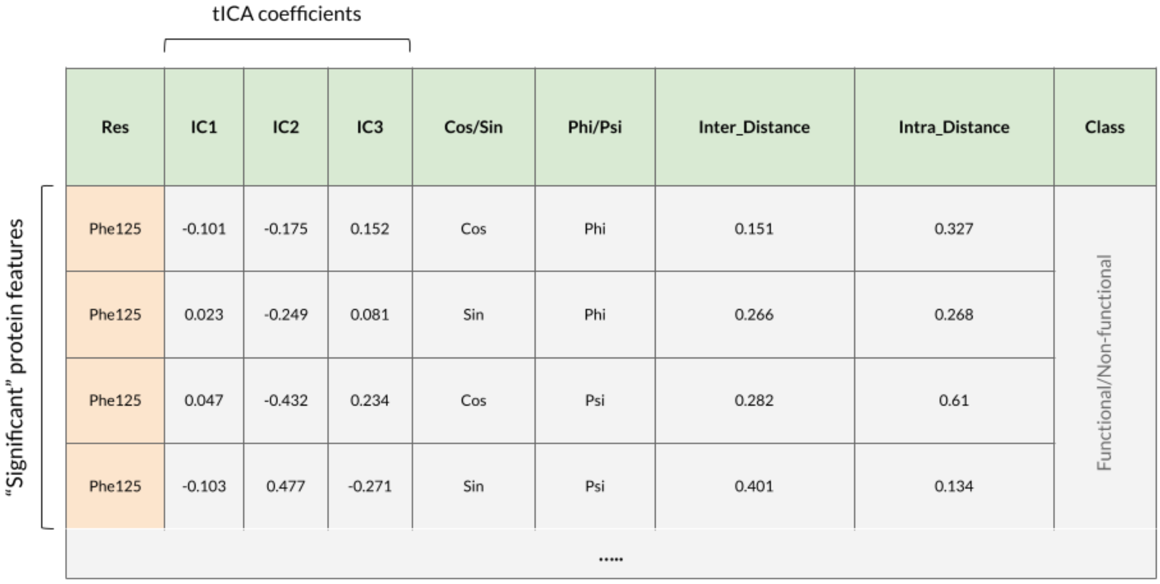


Figure 27 Time evolution of the bottleneck radius of channels 2b, 2f, 2e, 2c and s in each CYP2D6 variant during simulations. The color of each element expresses the bottleneck radius of a given tunnel cluster in x_{th} snapshot. The gray color indicates that no tunnel from a given cluster was identified in a given snapshot.The color scheme indicates the bottleneck radius, going from red indicating narrow opening to green indicating wide bottleneck

4. A Prediction Model for the metabolizer phenotype of CYP2D6

The classical MD runs on the CYP2D6 *1 were used for identifying the important residues for the protein conformational space. Thus, retrieved protein trajectories of CYP2D6 *1 were further analyzed by Markov State Modeling (MSM) (Prinz et al. 2011). Time-structure independent components analysis (tICA) was used to reduce the dimensionality of our data, in PyEMMA (Scherer et al. 2015). The tICA method identified the torsional angles of the following CYP2D6 residues: **103, 352, 378, 406, 407, 409, 410, 411, 460, 462 and 464** as the most important features, by setting a threshold equal to 0.19 for the coefficients in the tICA vectors. For the selection of this threshold, we checked for different thresholds the VAMP2-score and the states projected onto the first two independent components. Based on the torsional angles of these residues and the coefficients of the tICA vectors, we were able to induce a reduced space that represent each MD-based simulated protein trajectory in the form shown in Figure 28 (columns Residue, IC1, IC2, IC3, Cosine/Sine and Phi/Psi).


tICA coefficients
“Significant” protein features

Res	IC1	IC2	IC3	Cos/Sin	Phi/Psi	Inter_Distance	Intra_Distance	Class
Phe125	-0.101	-0.175	0.152	Cos	Phi	0.151	0.327	Functional/Non-functional
Phe125	0.023	-0.249	0.081	Sin	Phi	0.266	0.268	
Phe125	0.047	-0.432	0.234	Cos	Psi	0.282	0.61	
Phe125	-0.103	0.477	-0.271	Sin	Psi	0.401	0.134	
.....								

Figure 28 tICA-based reduced representation of a proteins’ trajectory (IC values), enhanced with engineered features that capture the inter- and intra-distance of residues (see text)

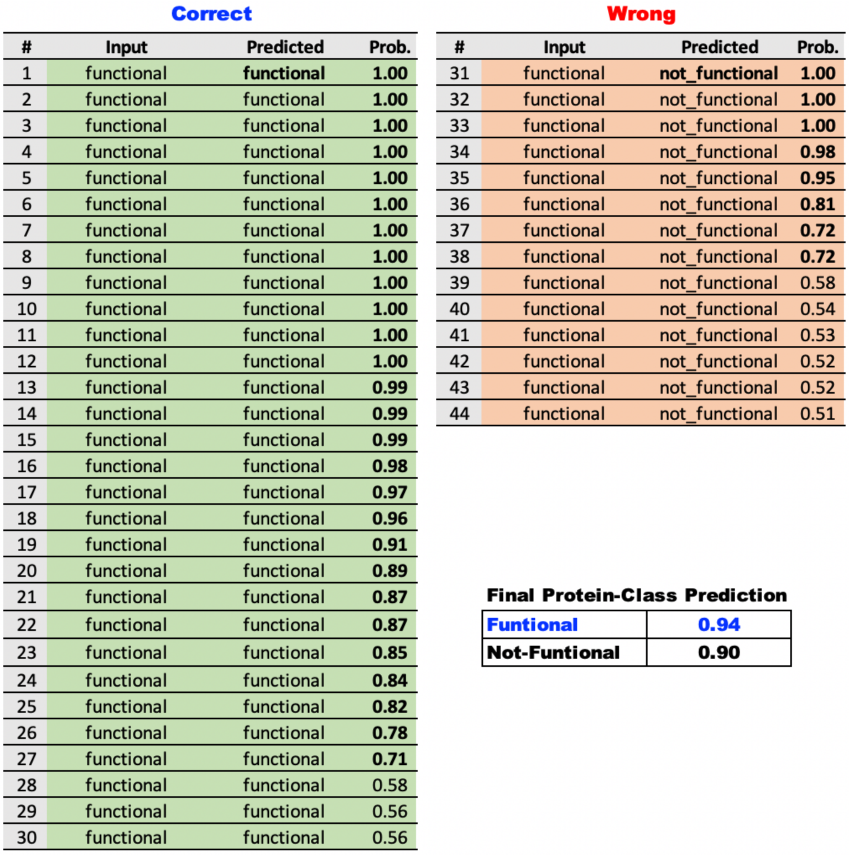
Representativeness of residues. In the representation shown in Figure 28 there are two specially ***engineered features***, added to capture “hidden” relations that underlie and putatively govern the functional status of the different allelic forms. For each residue entry (row) belonging to a specific functional class (functional, non-functional), ‘***Inter_Distance***’ is computed as the average over all distances between the tICA internal-components (IC1, IC2, IC3) of this residue entry and the corresponding internal-components of residue entries that belongs to the opposite class. In other words, Inter-Distance captures and contrasts the distance between corresponding entries (as defined by residue-position and cosine/sine between phi/psi dihedral angles) that belong to different classes. It could be considered as a measure of ***representativeness of the entry for the specific protein***. On the other hand, '***Intra_Distance***' is the distance of each entry to its corresponding ones for variants that belong to the same class. In other words, Intra_Distance seizes the ***representativeness of each entry for the class*** it belongs, capturing the ***cohesion*** of the functional class itself. In the conducted experiments the *Euclidean* distance was used. So, each reduced protein file (see Figure 28) comprises eleven – (11) positions for the respective residues identified as most influential. Each position refers to four – (4) different cosine/sine - phi/psi dihedral combinations, resulting into a total of forty-four – (44) entries, with each of them accompanied by the coefficients of the three tICA vectors and the corresponding Inter_Distance and Intra_Distance figures.

Functional class-prediction models from MD simulated proteins. As already mentioned, the work of this thesis comprises classical MD simulations of eight – (8) CYP2D6 variants including functional and non-functional ones. The wild-type, CYP2D6 *1 form, as well as three other variants, *2, *33 and *48 belong to the functional class, and forms *7, *12, *62 and *99 belong to the non-functional forms. For each variant we have at our disposal (after the respective classical MD runs and MSM analysis are performed) the corresponding matrix shown in Figure 28, and these files are our input.

- **Learning approach.** The induction and performance assessment of the devised prediction models was done with the utilization of various Machine Learning methods. In order to decide which of the learning approaches is the most promising for our task we conducted **cross-validation** experiments (10-fold and leave-one-out/LOOCV) using as input all eight proteins (functional and non-functional). After extensive experimentation, the **Multi-Layer Perceptron / MLP** learning algorithm shown the best results. MLP is a widely-utilized *feed-forward artificial neural network* (ANN) supervised learning approach based on *non-linear unit activation* functions and *backpropagation* in its core process (Hastie, Tibshirani, and Friedman 2001). The '***vanilla***'-like MLP architectures have just a single hidden layer. Our experiments were conducted with a vanilla-like MLP form. The Weka MLP implementation was used.
- **Data setup.** For each target protein the *training* input comprises the files of residue entries (as described above) from all the proteins except the file of the target protein, with the later used for *testing*. For example, the training input for the wt CYP2D6 form consists of seven concatenated files (for functional protein forms *2, *33, *48 and non-functional protein

forms *7, *12, *62 and *99). So, we train the MPL network on the wt-train input file and test it on the 'unseen' wt-test file in order to assess the accuracy of the prediction model, i.e., how well it differentiates between functional and non-functional protein classes.

- **Strength of predictions.** As with every learning algorithm, when an MLP trained model is applied on an unseen test case it outputs its prediction with a ***prediction probability***. For example, if we assume that we have trained the model with wt-train and test it on wt-test, the algorithm outputs its predictions in the form shown in Figure 29. Taking the **average over a specific range of these probabilities** (here the 0.7 threshold is used, figures in bold) we come up with a prediction for wt-test to be functional or non-functional with '**strength**' of 0.94 and 0.90, respectively.



Correct

#	Input	Predicted	Prob.
1	functional	**functional**	**1.00**
2	functional	functional	**1.00**
3	functional	functional	**1.00**
4	functional	functional	**1.00**
5	functional	functional	**1.00**
6	functional	functional	**1.00**
7	functional	functional	**1.00**
8	functional	functional	**1.00**
9	functional	functional	**1.00**
10	functional	functional	**1.00**
11	functional	functional	**1.00**
12	functional	functional	**1.00**
13	functional	functional	**0.99**
14	functional	functional	**0.99**
15	functional	functional	**0.99**
16	functional	functional	**0.98**
17	functional	functional	**0.97**
18	functional	functional	**0.96**
19	functional	functional	**0.91**
20	functional	functional	**0.89**
21	functional	functional	**0.87**
22	functional	functional	**0.87**
23	functional	functional	**0.85**
24	functional	functional	**0.84**
25	functional	functional	**0.82**
26	functional	functional	**0.78**
27	functional	functional	**0.71**
28	functional	functional	0.58
29	functional	functional	0.56
30	functional	functional	0.56

Wrong

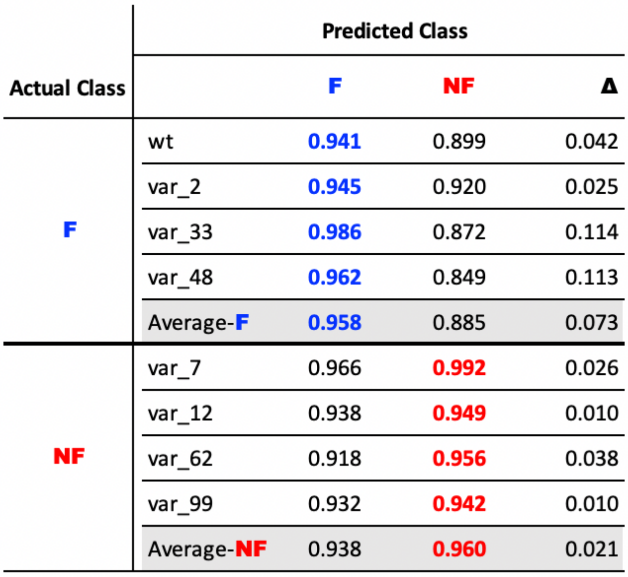
#	Input	Predicted	Prob.
31	functional	**not_functional**	**1.00**
32	functional	not_functional	**1.00**
33	functional	not_functional	**1.00**
34	functional	not_functional	**0.98**
35	functional	not_functional	**0.95**
36	functional	not_functional	**0.81**
37	functional	not_functional	**0.72**
38	functional	not_functional	**0.72**
39	functional	not_functional	0.58
40	functional	not_functional	0.54
41	functional	not_functional	0.53
42	functional	not_functional	0.52
43	functional	not_functional	0.52
44	functional	not_functional	0.51

Final Protein-Class Prediction

Funtional	**0.94**
Not-Funtional	**0.90**

Figure 29 Example of algorithm's predictions

Following the aforementioned prediction modeling process (with a threshold of 0.7) we came up with the prediction strengths for all CYP2D6 allelic forms shown in Table 30. The results show that **all variants are correctly predicted to belong to their correct functional classes**. We have performed experiments with higher prediction strength thresholds (e.g., 0.8) and the results are the same. Of course, the whole prediction modelling process should be tested and evaluated on other domains and with bigger populations of MD-simulated protein data in order to confirm its efficacy, efficiency and reliability.



Actual Class	Predicted Class			
		F	**NF**	**Δ**
F	wt	**0.941**	0.899	0.042
	var_2	**0.945**	0.920	0.025
	var_33	**0.986**	0.872	0.114
	var_48	**0.962**	0.849	0.113
	Average-**F**	**0.958**	0.885	0.073
NF	var_7	0.966	**0.992**	0.026
	var_12	0.938	**0.949**	0.010
	var_62	0.918	**0.956**	0.038
	var_99	0.932	**0.942**	0.010
	Average-**NF**	0.938	**0.960**	0.021

Figure 30 Predicting the functional status of CYP2D6 protein forms (blue: functional, red: non-functional) – Results

5. Discussion

In the present study, long time-scale Molecular Dynamics simulations was performed for a series of eight different CYP2D6 variants (1 microsecond per variant), including functional and non-functional allelic forms.

Molecular Dynamics analyses on CYP2D6 variants allowed an atomic-level description of the possible effects of specific mutations on their metabolizer phenotype. Based on the residue b-factors, several sites of non-functional variants seem to be less stable in comparison with the functional variants. More specifically, F'-G loop along with H-I loop K-K' loop and β3-sheet seem to be less stable in non-functional variants than in the functional ones. Also, the loss of anti-correlation relationships in combination with the loss of correlation relationships, in case of non-functional variants, can influence the functionality of CYP2D6. In particular, the loss of the anti-correlation relationships developed by helix I in functional variants and their replacement by relationships developed by helix F'/F-F loop, in combination with the loss of correlation relationships can influence the stability and functionality of the enzyme. Both changes could be used as markers in the discrimination of the two classes of metabolizing activity.

Results from Molecular docking using BACE1 inhibitor as a substrate, confirm the previous analyses and highlight the role of helix I and of K-K' loop and their relative movement in the activity of the enzyme. BACE1 inhibitor seems to be closer to the active site in functional than in the non-functional variants. Analysis of substrate channels revealed that there are several differences in the major pathways as well as in bottleneck radius and duration of openings.

Furthermore, retrieved protein trajectories of the CYP2D6 *1 (wild-type) were used for identifying the important residues for the protein conformational space using Markov State Modeling. Based on these residues and using the data from the tICA/MSM analysis, a dataset for each variant has been produced which was then used to build a prediction model for the metabolizer phenotype.

Various Molecular Dynamics simulation studies have been performed in the past on other CYP2D6 variants, however, to the best of our knowledge, this study is the first attempt in which a wide range of both functional and non-functional allelic forms are co-examined using Molecular dynamics and changes that are present in all non-functional enzymes and absent in all functional ones are identified. Additionally, it is the first time that a prediction model for the metabolizer phenotype has been developed.

6. Material and Methods

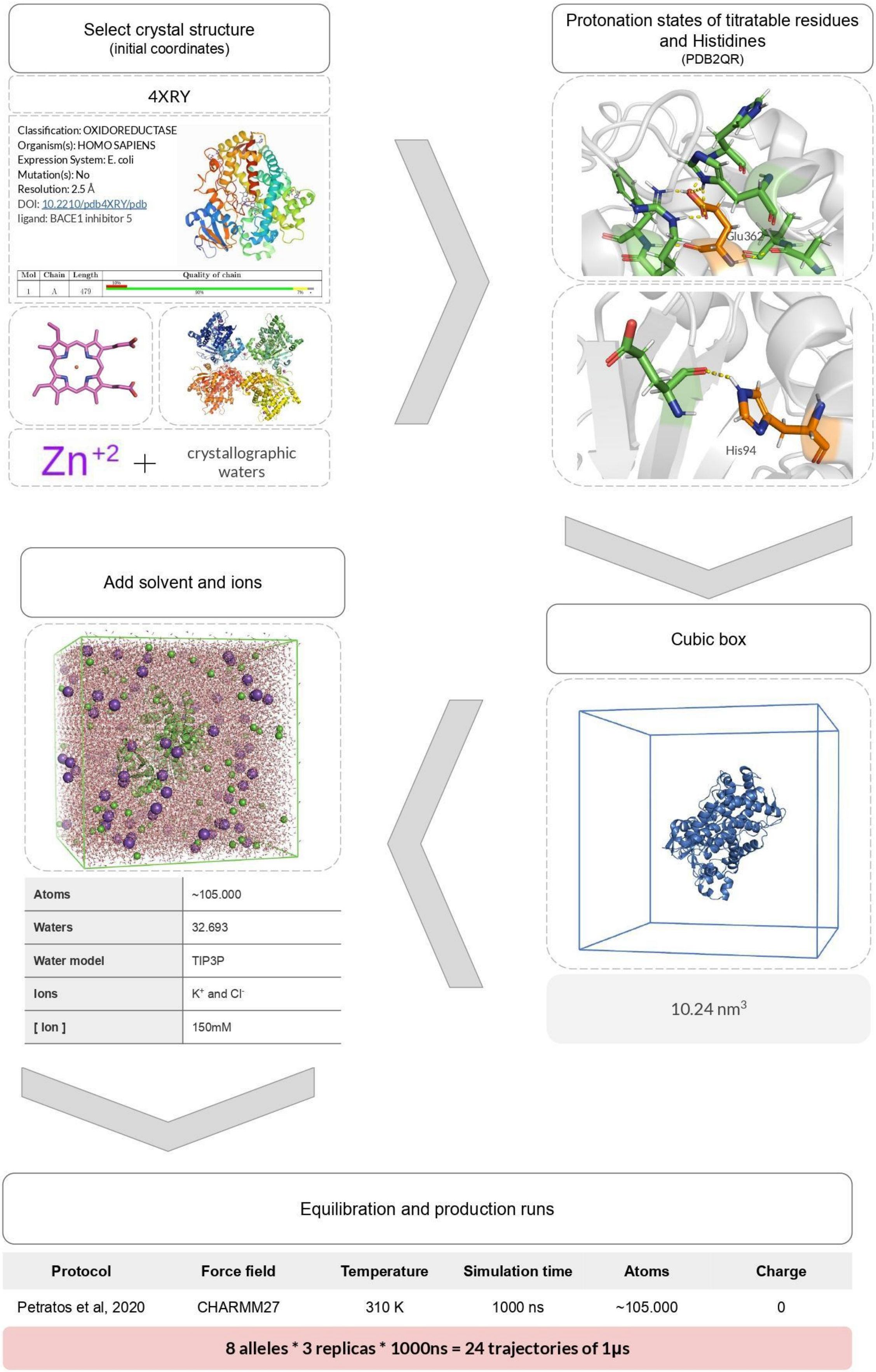
Model setup

There are various three-dimensional structures available in the literature for the Cytochrome P450 2D6 (CYP2D6). In the present study, the crystal structure of the Human Cytochrome P450 2D6 BACE1 Inhibitor 5 complex (PDB code 4XRY) was used for the initial coordinates to build the models. Chain A, its corresponding heme cofactor and crystallographic waters were retained whereas the inhibitor was removed. Our choice of coordinates was based on the completeness of the resolved CYP2D6 sequence and the quality of chain (at least 90%). The protonation states of titratable residues were simulated at neutral pH, thus all Glu, and Asp residues were left deprotonated, except Glu-362 which was protonated, in accordance with the PDB2PQR (propka 3.0 method, pH 7.3) predictions (Dolinsky et al. 2004). His-48, His-94, His-324, His-376, His-416, His-426, His-463 and His-477 were protonated only at the Nε site. The rest of His residues were protonated only at the Nδ sites, to maintain the hydrogen bonding network within the crystal structures. The all-atom models, as defined previously, were embedded in orthorhombic boxes of around 10.2nm x 10.2nm x 10.2 nm in the x, y and z dimensions. Up to around 32500 TIP3P water molecules (Mark and Nilsson 2001) were used to hydrate each protein. Ion (K^+, Cl^-) concentration was set at the value of 150 mM to mimic the physiological salt content. A surplus of Cl^- was also added to neutralize the protein charges in each sample, resulting in simulation unit boxes of around 105000 atoms. The CHARMM27 (MacKerell et al. 1998) protein force field was used for the residues and ions.

Equilibration-Production Molecular Dynamics setup

The equilibration-relaxation for the all-atom systems is employed based on a published protocol for water-soluble proteins (Petratos et al. 2020). This contains a steepest descend energy minimization with a tolerance of 0.5 kJ mol^{-1} for 1000 steps, and a sequence of isothermal (nVT), isothermal-isobaric (nPT) runs with the gradual relaxation of the constraints on protein heavy atoms (from 10^4 in steps 1-2 to 10^3 kJ mol^{-1} nm^{-2} in step-4) and Cα atoms (from 10^3 in step-5, to 10^2 in step-6, 10 in step-7, 1 in step-8 and 0 kJ mol^{-1} nm^{-2} in step-9) for around 30 ns, with a time step of 1.0 fs (steps 1-4) and 2.0 fs (steps 5-9). In detail: (step-1) Constant density and temperature (nVT) Brownian dynamics (BD) at 100 K for 50 ps that employs the Berendsen thermostat (Berendsen et al. 1984), with a temperature coupling constant at 1.0 fs. (steps 2-3) Two short constant density (nVT) and constant pressure (nPT) runs for 100 ps each, with a weak coupling Berendsen thermostat and barostat (Berendsen et al. 1984) at 100 K employing time coupling constants of 0.1 ps for the temperature and isotropic 50.0 ps coupling for the pressure with a compressibility of $4.6x10^{-5}$. (step-4) Heating from 100 to 250 K in a constant density ensemble (nVT) for 3 ns employing the v-rescale thermostat (Bussi, Donadio, and Parrinello 2007), with a time coupling constant of 0.1 ps. (step-5) Heating from 250 to 310K in a constant pressure ensemble (nPT) for 2 ns, employing the v-rescale thermostat (Bussi, Donadio, and Parrinello 2007) and Berendsen barostat (Berendsen et al. 1984), with time coupling constants of 0.1 ps for the temperature and 2.0 ps for the pressure, removing also all but the Cα-atom protein position restraints. (step-6) Equilibration at 310K (0.1 ps temperature coupling constant) for 5 ns (nPT, 1 atm, 2.0 ps coupling constant for pressure. (steps 7-8) Equilibration at 310K (0.5 ps temperature coupling constant) for 5 ns (nPT, 1 atm, 2.0 ps coupling constant for pressure). (step-9) Equilibration at 310K (0.5 ps temperature coupling constant) for 10 ns (nPT, 1 atm, 2.0 ps coupling constant for pressure). The barostats-thermostats employed for steps 6-9 were the same as in the production trajectories that follow.

For the production trajectories within the all-atom MD methodology, the Newton's equations of motion are integrated with a time step of 2.0 fs at 310K. All production simulations are run with the leap-frog integrator in GROMACS 2020 (Berendsen, van der Spoel, and van Drunen 1995) for 1.0 μs each. They were performed at the constant pressure nPT ensemble, with isotropic coupling (compressibility at $4.5x10^{-5}$) employing the v-rescale thermostat (Bussi, Donadio, and Parrinello 2007) (310K, temperature coupling constant 0.5) and the *Parrinello-Rahman* barostat (Parrinello and Rahman 1981) (1 atm, pressure coupling constant 2.0). Details for parameters can be found in an earlier work (Petratos et al. 2020). The first 150 ns were considered further equilibration from each independent trajectory per sample, and were disregarded in the analysis, based also on the RMSD fluctuations (a plateau is reached roughly beyond 100ns depending on the trajectory). Van der Waals interactions were smoothly switched to zero between 1.0-1.2 nm with the VERLET cut-off scheme. Electrostatic interactions were truncated at 1.2 nm (short-range) and long-range contributions were computed within the PME approximation (Darden, York, and Pedersen 1993; Yeh and Berkowitz 1999). Hydrogen bond lengths were constrained employing the LINCS algorithm (Hess et al. 1997).


Select crystal structure
(initial coordinates)
4XRY
Zn+2 + crystallographic waters
Protonation states of titratable residues and Histidines
(PDB2QR)
Glu362
His94
Cubic box
10.24 nm3
Add solvent and ions
Equilibration and production runs

Classification: OXIDOREDUCTASE
Organism(s): HOMO SAPIENS
Expression System: E. coli
Mutation(s): No
Resolution: 2.5 Å
DOI: 10.2210/pdb4XRY/pdb
ligand: BACE1 inhibitor 5

Mol	Chain	Length	Quality of chain
1	A	479	10% 90% 7%

Atoms	~105.000
Waters	32.693
Water model	TIP3P
Ions	K^+ and Cl^-
[Ion]	150mM

Protocol	Force field	Temperature	Simulation time	Atoms	Charge
Petratos et al, 2020	CHARMM27	310 K	1000 ns	~105.000	0

8 alleles * 3 replicas * 1000ns = 24 trajectories of 1μs

Figure 31 Molecular Dynamics simulations pipeline.

RMSD and RMSF

RMSD and RMSF were calculated over the 850 ns production simulation for all backbone atoms (Cα, C, N) using GROMACS 2020. Corresponding structures were prepared by populating the B-Factor column in the pdb of the corresponding genetic variant and were then visualized using PyMOL.

Root Mean Square Deviation (RMSD) and the Root Mean Square Fluctuations (RMSF) are two of the most common measures of structural fluctuations. RMSD, a useful measure for the analysis of time-independent motions of the structure, represents the average displacement of the atoms at an instant of the simulation relative to a reference structure. RMSF is the displacement of a particular atom relative to the reference structure averaged over the number of atoms and is a measure of individual residue flexibility

Dynamic cross correlation analysis

For each protein, we perform calculations for residue-level dynamic cross-correlations on the respective Cα trajectory using the *dccm* function in the Bio3D package with the following equation.

$$DCC_{MD}\left(i, j\right) = \frac{< \Delta r_i(t) . \Delta r_j(t) >_t}{\sqrt{< ||\Delta r_i(t)||^2 >_t}\sqrt{< ||\Delta r_j(t)||^2 >_t}}$$

with $r_i(t)$ and $r_j(t)$ refer to the coordinates of the ith and jth atoms as a function of time t, <> indicates the time ensemble average and $\Delta r_i(t) = r_i(t) - (< r_i(t) >)t$ and $\Delta r_j(t) = r_j(t) - (<r_j(t)>)t$.

Molecular Docking

AutoDock Vina was the docking program used in this study (Trott and Olson 2010). PDBQT file format was prepared, and the grid box size was determined using AutoDock Tools version 1.5.4 (Morris et al. 2009). Ligand was docked individually to the receptors with grid coordinates (grid center) and grid boxes of certain sizes for each receptor.

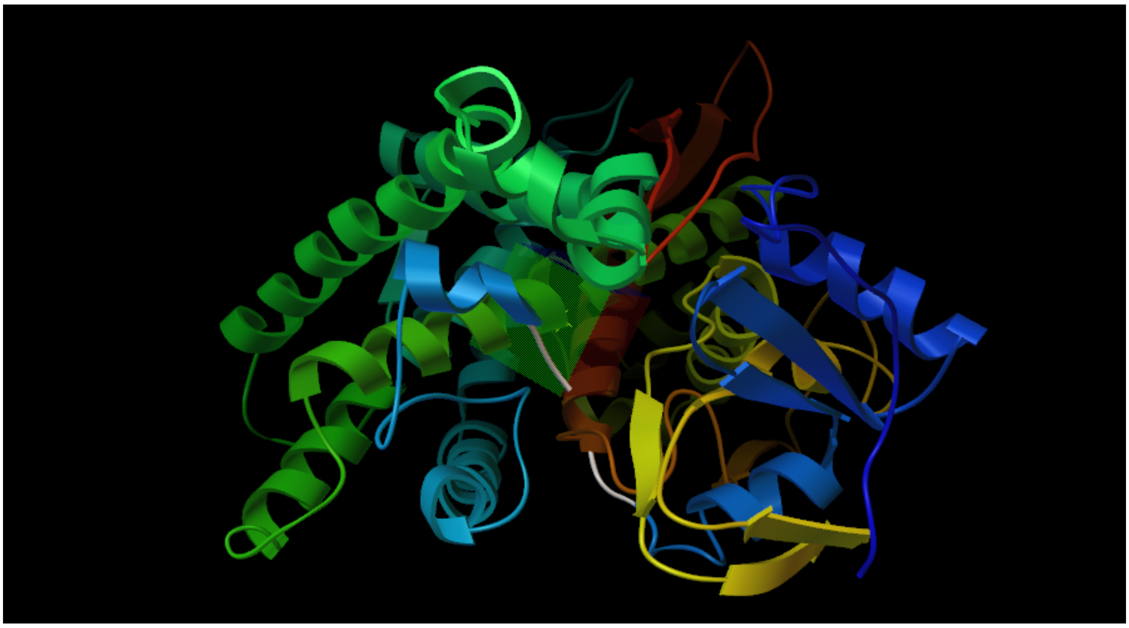


Figure 32 Screenshot from AutoDock Tools. Representative protein structure of CYP2D6 *1 along with the grid box.

Tunnel Analysis

Snapshots of each variant were taken over the 850 ns simulation run time generating a total of 1000 snapshots used in tunnel analysis for each allelic form. CAVER 3.0 software was used for the analysis of substrate accessibility and egress channels (Chovancova et al. 2012). The starting point for CAVER analysis was set at 4 Å above the iron atom of the heme group. The probe radius was set to 0.9 Å and clustering threshold was initially set at the default value of 4.0. The bottleneck heat map range was set at 0.9–2.5 Å and the profile heat map range was set at 0.9–2.0 Å. Seed was set to 1 to ensure consistent results. All other parameters were set to the default values as listed in the CAVER user guide version 3.0 and included: shell radius (3), weighting coefficient (1) for tunnel clustering, bottleneck contact distance (3), the number of approximating balls (12), max distance for the calculation starting point from the initial starting point (3), and desired radius (5) for the closest Voronoi vertex to the initial starting point. Resulting tunnels were identified and visualized in PyMOL.

Markov State Modeling of CYP2D6 *1 (wild-type)

We obtained a series of MD equilibrium trajectories of CYP2D6 *1 (3 x 1.0μs = 3μs). We combined the all-atom MD simulations with Markov state model (MSM) theory (Pande, Beauchamp, and Bowman 2010; Chodera and Noé 2014; Prinz et al. 2011) in order to enable the extraction of long-time-scale dynamics from rather short-time-scale MD trajectories of different states. The application and accuracy of the powerful MSM theory has been presented in many cases also by experiments that include protein−protein, or protein-drug binding kinetics, as well as protein folding rates and protein dynamics (Plattner and Noé 2015; Plattner et al. 2017; Voelz et al. 2010; Durrant et al. 2020). Our objective was to approximate the slow dynamics in a statistically efficient manner. Thus, a lower dimensional representation of our simulation data was necessary. In order to reduce the dimensionality of our feature space, we employed the time-structure independent components analysis (tICA) which yields a representation of our molecular simulation data with a reduced dimensionality and can greatly facilitate the decomposition of our system into the discrete Markovian states necessary for MSM estimation. The conformations of the system were projected on these slowest modes as defined by the tICA method, then the trajectory frames were clustered into 30 cluster-centers (microstates) by k-means clustering, as implemented in PyEMMA (Scherer et al. 2015). Conformational changes of a system can be simulated as a Markov chain, if the transitions between the different conformations are sampled at long enough time intervals so that each transition is Markovian. This means that a transition from one conformation to another is independent of the previous transitions. Therefore, an MSM is a memoryless model. The uncertainty bounds were computed using a Bayesian scheme (Noé 2008; Trendelkamp-Schroer et al. 2015). We found that the slowest implied timescales converge quickly and are constant within a 95% confidence interval for lag times above 50ns. The validation procedure is a standard approach in the MSM field. A lag time of 50 ns was selected for Bayesian model construction, and the resulting models were validated by the Chapman-Kolmogorov (CK) test. Subsequently, the resulting MSMs were further coarse grained into a smaller number of three metastable states or microstates, using PCCA++ as implemented in PyEMMA. The optimum number of microstates (three) was proposed based on the VAMP2-score (H. Wu and Noé 2020). Both the convergence of the implied timescales, as well as the CK test confirm the validity and convergence of the MSM. The CK test indicates that predictions from the built MSM agree well with MSMs estimated with longer lag times. Thus, the model can describe well the long-time-scale behavior of our system within error. The tICA method identified the torsional angles of the following CYP2D6 residues: 125, 274, 400, 428, 429, 431, 432, 433, 482, 484 and 486 as the most important features, by setting a threshold equal to 0.19 for the coefficients in the tICA vectors. For the selection of this threshold, we checked for different thresholds the VAMP2-score and the states projected onto the first two independent components. In MD approaches it is common to capture protein conformations via its internal atomic coordinates, with dihedral angles between specific atoms to represent one of the most utilized and successful representations (Sittel, Jain, and Stock 2014). Namely, two such angles are used, after their introduction by Ramachandran (Ramachandran, Ramakrishnan, and Sasisekharan 1963): φ (phi) - the angle in the atoms' backbone chain C' − N − Cα − C', and ψ (psi) – the angle in the atom's backbone chain N − Cα − C' – N. In particular the relation between the cosine (cos) and sine (sin) of these angles is utilized in order to capture the motion of the internal atomic coordinates during simulated conformations. Data from the tICA/MSM analysis of the different variants for these specific residues were used in order to build the prediction model for the metabolizer phenotype using WEKA software (Hall et al. 2009)

Computational resources

This work was supported by computational time granted from the National Infrastructures for Research and Technology S.A. (GRNET S.A.) in the National HPC facility - ARIS - under the projects ID pa201103 CYP2D6SIM and ID pr10010_thin CYP2D6-DYN.

References

Ahmed, Shabbir, Zhan Zhou, Jie Zhou, and Shu-Qing Chen. 2016. “Pharmacogenomics of Drug Metabolizing Enzymes and Transporters: Relevance to Precision Medicine.” *Genomics, Proteomics & Bioinformatics* 14 (5): 298–313.

Algeciras-Schimnich, Alicia, Dennis J. O’Kane, and Christine L. H. Snozek. 2008. “Pharmacogenomics of Tamoxifen and Irinotecan Therapies.” *Clinics in Laboratory Medicine* 28 (4): 553–67.

Berendsen, H. J. C., J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak. 1984. “Molecular Dynamics with Coupling to an External Bath.” *The Journal of Chemical Physics* 81 (8): 3684–90.

Berendsen, H. J. C., D. van der Spoel, and R. van Drunen. 1995. “GROMACS: A Message-Passing Parallel Molecular Dynamics Implementation.” *Computer Physics Communications* 91 (1-3): 43–56.

Brandl, E. J., A. K. Tiwari, X. Zhou, J. Deluce, J. L. Kennedy, D. J. Müller, and M. A. Richter. 2014. “Influence of CYP2D6 and CYP2C19 Gene Variants on Antidepressant Response in Obsessive-Compulsive Disorder.” *The Pharmacogenomics Journal* 14 (2): 176–81.

Bussi, Giovanni, Davide Donadio, and Michele Parrinello. 2007. “Canonical Sampling through Velocity Rescaling.” *The Journal of Chemical Physics* 126 (1): 014101.

Chodera, John D., and Frank Noé. 2014. “Markov State Models of Biomolecular Conformational Dynamics.” *Current Opinion in Structural Biology* 25 (April): 135–44.

Chovancova, Eva, Antonin Pavelka, Petr Benes, Ondrej Strnad, Jan Brezovsky, Barbora Kozlikova, Artur Gora, et al. 2012. “CAVER 3.0: A Tool for the Analysis of Transport Pathways in Dynamic Protein Structures.” *PLoS Computational Biology* 8 (10): e1002708.

Cojocaru, Vlad, Peter J. Winn, and Rebecca C. Wade. 2007. “The Ins and Outs of Cytochrome P450s.” *Biochimica et Biophysica Acta* 1770 (3): 390–401.

Darden, Tom, Darrin York, and Lee Pedersen. 1993. “Particle Mesh Ewald: An N⋅log(N) Method for Ewald Sums in Large Systems.” *The Journal of Chemical Physics* 98 (12): 10089–92.

Denisov, Ilia G., Thomas M. Makris, Stephen G. Sligar, and Ilme Schlichting. 2005. “Structure and Chemistry of Cytochrome P450.” *Chemical Reviews* 105 (6): 2253–77.

Dolinsky, Todd J., Jens E. Nielsen, J. Andrew McCammon, and Nathan A. Baker. 2004. “PDB2PQR: An Automated Pipeline for the Setup of Poisson-Boltzmann Electrostatics Calculations.” *Nucleic Acids Research* 32 (Web Server issue): W665–67.

Durrant, Jacob D., Sarah E. Kochanek, Lorenzo Casalino, Pek U. Ieong, Abigail C. Dommer, and Rommie E. Amaro. 2020. “Mesoscale All-Atom Influenza Virus Simulations Suggest New Substrate Binding Mechanism.” *ACS Central Science* 6 (2): 189–96.

Durrant, Jacob D., and J. Andrew McCammon. 2011. “Molecular Dynamics Simulations and Drug Discovery.” *BMC Biology* 9 (October): 71.

Gaedigk, Andrea, Katrin Sangkuhl, Michelle Whirl-Carrillo, Teri Klein, and J. Steven Leeder. 2017. “Prediction of CYP2D6 Phenotype from Genotype across World Populations.” *Genetics in Medicine: Official Journal of the American College of Medical Genetics* 19 (1): 69–76.

Grant, Barry J., Ana P. C. Rodrigues, Karim M. ElSawy, J. Andrew McCammon, and Leo S. D. Caves. 2006. “Bio3d: An R Package for the Comparative Analysis of Protein Structures.” *Bioinformatics* 22 (21): 2695–96.

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. “The WEKA Data Mining Software.” *ACM SIGKDD Explorations*

Newsletter. https://doi.org/10.1145/1656274.1656278.

Hasemann, C. A., R. G. Kurumbail, S. S. Boddupalli, J. A. Peterson, and J. Deisenhofer. 1995. “Structure and Function of Cytochromes P450: A Comparative Analysis of Three Crystal Structures.” *Structure* 3 (1): 41–62.

Hasler, Julia A., Ronald Estabrook, Michael Murray, Irina Pikuleva, Michael Waterman, Jorge Capdevila, Vijakumar Holla, et al. 1999. “Human Cytochromes P450.” *Molecular Aspects of Medicine* 20 (1-2): 1–137.

Hastie, Trevor, Robert Tibshirani, and Jerome H. Friedman. 2001. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media.

Heim, M. H., and U. A. Meyer. 1992. “Evolution of a Highly Polymorphic Human Cytochrome P450 Gene Cluster: CYP2D6.” *Genomics* 14 (1): 49–58.

Hess, Berk, Henk Bekker, Herman J. C. Berendsen, and Johannes G. E. Fraaije. 1997. “LINCS: A Linear Constraint Solver for Molecular Simulations.” *Journal of Computational Chemistry*. https://doi.org/10.1002/(sici)1096-987x(199709)18:12<1463::aid-jcc4>3.0.co;2-h.

Hollingsworth, Scott A., and Ron O. Dror. 2018. “Molecular Dynamics Simulation for All.” *Neuron* 99 (6): 1129–43.

“Home Page.” n.d. Accessed May 17, 2021. https://cpicpgx.org/.

“Human Cytochrome P450s.” n.d. Accessed May 14, 2021. http://drnelson.uthsc.edu/human.P450.table.html.

Ingelman-Sundberg, Magnus, Sarah C. Sim, Alvin Gomez, and Cristina Rodriguez-Antona. 2007. “Influence of Cytochrome P450 Polymorphisms on Drug Therapies: Pharmacogenetic, Pharmacoepigenetic and Clinical Aspects.” *Pharmacology & Therapeutics* 116 (3): 496–526.

Isin, Emre M., and F. Peter Guengerich. 2007. “Complex Reactions Catalyzed by Cytochrome P450 Enzymes.” *Biochimica et Biophysica Acta* 1770 (3): 314–29.

Jaladanki, Chaitanya K., Anuj Gahlawat, Gajanan Rathod, Hardeep Sandhu, Kousar Jahan, and Prasad V. Bharatam. 2020. “Mechanistic Studies on the Drug Metabolism and Toxicity Originating from Cytochromes P450.” *Drug Metabolism Reviews* 52 (3): 366–94.

Jancova, Petra, Pavel Anzenbacher, and Eva Anzenbacherova. 2010. “Phase II Drug Metabolizing Enzymes.” *Biomedical Papers of the Medical Faculty of the University Palacky, Olomouc, Czechoslovakia* 154 (2): 103–16.

Johnson, Eric F., and C. David Stout. 2005. “Structural Diversity of Human Xenobiotic-Metabolizing Cytochrome P450 Monooxygenases.” *Biochemical and Biophysical Research Communications* 338 (1): 331–36.

———. 2013. “Structural Diversity of Eukaryotic Membrane Cytochrome p450s.” *The Journal of Biological Chemistry* 288 (24): 17082–90.

Kadlubar, Susan, and Fred F. Kadlubar. 2010. “Enzymatic Basis of Phase I and Phase II Drug Metabolism.” In *Enzyme- and Transporter-Based Drug-Drug Interactions*, 3–25. New York, NY: Springer New York.

Karplus, Martin, and J. Andrew McCammon. 2002. “Molecular Dynamics Simulations of Biomolecules.” *Nature Structural Biology* 9 (9): 646–52.

Keizers, Peter H. J., Barbara M. A. Lussenburg, Chris de Graaf, Letty M. Mentink, Nico P. E. Vermeulen, and Jan N. M. Commandeur. 2004. “Influence of Phenylalanine 120 on Cytochrome P450 2D6 Catalytic Selectivity and Regiospecificity: Crucial Role in 7-Methoxy-4-(aminomethyl)-Coumarin Metabolism.” *Biochemical Pharmacology* 68 (11): 2263–71.

Korobkova, Ekaterina A. 2015. “Effect of Natural Polyphenols on CYP Metabolism: Implications for Diseases.” *Chemical Research in Toxicology* 28 (7): 1359–90.

Lamb, David C., and Michael R. Waterman. 2013. "Unusual Properties of the Cytochrome P450 Superfamily." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 368 (1612): 20120434.

Lim, Young Chai, Zeruesenay Desta, David A. Flockhart, and Todd C. Skaar. 2005. "Endoxifen (4-Hydroxy-N-Desmethyl-Tamoxifen) Has Anti-Estrogenic Effects in Breast Cancer Cells with Potency Similar to 4-Hydroxy-Tamoxifen." *Cancer Chemotherapy and Pharmacology* 55 (5): 471–78.

Lüdemann, S. K., V. Lounnas, and R. C. Wade. 2000. "How Do Substrates Enter and Products Exit the Buried Active Site of Cytochrome P450cam? 1. Random Expulsion Molecular Dynamics Investigation of Ligand Access Channels and Mechanisms." *Journal of Molecular Biology* 303 (5): 797–811.

Lu, Jian-Da, and Jun Xue. 2019. "Poisoning." In *Critical Care Nephrology*, 600–629.e7. Elsevier.

Lynch, Tom, and Amy Price. 2007. "The Effect of Cytochrome P450 Metabolism on Drug Response, Interactions, and Adverse Effects." *American Family Physician* 76 (3): 391–96.

MacKerell, A. D., D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, et al. 1998. "All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins." *The Journal of Physical Chemistry. B* 102 (18): 3586–3616.

Manikandan, Palrasu, and Siddavaram Nagini. 2018. "Cytochrome P450 Structure, Function and Clinical Significance: A Review." *Current Drug Targets* 19 (1): 38–54.

Mark, Pekka, and Lennart Nilsson. 2001. "Structure and Dynamics of the TIP3P, SPC, and SPC/E Water Models at 298 K." *The Journal of Physical Chemistry. A* 105 (43): 9954–60.

Morris, Garrett M., Ruth Huey, William Lindstrom, Michel F. Sanner, Richard K. Belew, David S. Goodsell, and Arthur J. Olson. 2009. "AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility." *Journal of Computational Chemistry* 30 (16): 2785–91.

Nebert, Daniel W., Kjell Wikvall, and Walter L. Miller. 2013. "Human Cytochromes P450 in Health and Disease." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 368 (1612): 20120431.

Nebert, D. W., M. Adesnik, M. J. Coon, R. W. Estabrook, F. J. Gonzalez, F. P. Guengerich, I. C. Gunsalus, E. F. Johnson, B. Kemper, and W. Levin. 1987. "The P450 Gene Superfamily: Recommended Nomenclature." *DNA* 6 (1): 1–11.

Nelson, D. R., L. Koymans, T. Kamataki, J. J. Stegeman, R. Feyereisen, D. J. Waxman, M. R. Waterman, et al. 1996. "P450 Superfamily: Update on New Sequences, Gene Mapping, Accession Numbers and Nomenclature." *Pharmacogenetics* 6 (1): 1–42.

Noé, Frank. 2008. "Probability Distributions of Molecular Observables Computed from Markov Models." *The Journal of Chemical Physics* 128 (24): 244103.

Nofziger, Charity, Amy J. Turner, Katrin Sangkuhl, Michelle Whirl-Carrillo, José A. G. Agúndez, John L. Black, Henry M. Dunnenberger, et al. 2020. "PharmVar GeneFocus: CYP2D6." *Clinical Pharmacology and Therapeutics* 107 (1): 154–70.

Omura, T. 1999. "Forty Years of Cytochrome P450." *Biochemical and Biophysical Research Communications* 266 (3): 690–98.

Omura, Tsuneo, and Ryo Sato. 1964. "The Carbon Monoxide-Binding Pigment of Liver Microsomes." *The Journal of Biological Chemistry* 239 (7): 2379–85.

Otyepka, Michal, Josef Skopalík, Eva Anzenbacherová, and Pavel Anzenbacher. 2007. "What Common Structural Features and Variations of Mammalian P450s Are Known to Date?" *Biochimica et Biophysica Acta* 1770 (3): 376–89.

Pande, Vijay S., Kyle Beauchamp, and Gregory R. Bowman. 2010. "Everything You Wanted to Know about Markov State Models but Were Afraid to Ask." *Methods* 52 (1): 99–105.

Parrinello, M., and A. Rahman. 1981. "Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method." *Journal of Applied Physics* 52 (12): 7182–90.

Petratos, Kyriacos, Renate Gessmann, Vangelis Daskalakis, Maria Papadovasilaki, Yannis Papanikolau, Iason Tsigos, and Vassilis Bouriotis. 2020. "Structure and Dynamics of a Thermostable Alcohol Dehydrogenase from the Antarctic Psychrophile Sp. TAE123." *ACS Omega* 5 (24): 14523–34.

Petrović, Jelena, Vesna Pešić, and Volker M. Lauschke. 2020. "Frequencies of Clinically Important CYP2C19 and CYP2D6 Alleles Are Graded across Europe." *European Journal of Human Genetics: EJHG* 28 (1): 88–94.

"PharmGKB." n.d. Accessed May 17, 2021. https://www.pharmgkb.org/.

"PharmVar." n.d. Accessed May 17, 2021. https://www.pharmvar.org/.

Plattner, Nuria, Stefan Doerr, Gianni De Fabritiis, and Frank Noé. 2017. "Complete Protein-Protein Association Kinetics in Atomic Detail Revealed by Molecular Dynamics Simulations and Markov Modelling." *Nature Chemistry* 9 (10): 1005–11.

Plattner, Nuria, and Frank Noé. 2015. "Protein Conformational Plasticity and Complex Ligand-Binding Kinetics Explored by Atomistic Simulations and Markov Models." *Nature Communications* 6 (July): 7653.

Prinz, Jan-Hendrik, Hao Wu, Marco Sarich, Bettina Keller, Martin Senne, Martin Held, John D. Chodera, Christof Schütte, and Frank Noé. 2011. "Markov Models of Molecular Kinetics: Generation and Validation." *The Journal of Chemical Physics* 134 (17): 174105.

Ramachandran, G. N., C. Ramakrishnan, and V. Sasisekharan. 1963. "Stereochemistry of Polypeptide Chain Configurations." *Journal of Molecular Biology* 7 (July): 95–99.

Ruano, Gualberto, and Jonathan A. Kost. 2018. "Fundamental Considerations for Genetically-Guided Pain Management with Opioids Based on CYP2D6 and OPRM1 Polymorphisms." *Pain Physician* 21 (6): E611–21.

Scherer, Martin K., Benjamin Trendelkamp-Schroer, Fabian Paul, Guillermo Pérez-Hernández, Moritz Hoffmann, Nuria Plattner, Christoph Wehmeyer, Jan-Hendrik Prinz, and Frank Noé. 2015. "PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models." *Journal of Chemical Theory and Computation* 11 (11): 5525–42.

Schleinkofer, Karin, Sudarko, Peter J. Winn, Susanne K. Lüdemann, and Rebecca C. Wade. 2005. "Do Mammalian Cytochrome P450s Show Multiple Ligand Access Pathways and Ligand Channelling?" *EMBO Reports* 6 (6): 584–89.

Shaik, Sason, Shimrit Cohen, Yong Wang, Hui Chen, Devesh Kumar, and Walter Thiel. 2010. "P450 Enzymes: Their Structure, Reactivity, and Selectivity-Modeled by QM/MM Calculations." *Chemical Reviews* 110 (2): 949–1017.

Sharp, C. F., S. J. Gardiner, B. P. Jensen, R. L. Roberts, R. W. Troughton, J. G. Lainchbury, and E. J. Begg. 2009. "CYP2D6 Genotype and Its Relationship with Metoprolol Dose, Concentrations and Effect in Patients with Systolic Heart Failure." *The Pharmacogenomics Journal* 9 (3): 175–84.

Sittel, Florian, Abhinav Jain, and Gerhard Stock. 2014. "Principal Component Analysis of Molecular Dynamics: On the Use of Cartesian vs. Internal Coordinates." *The Journal of Chemical Physics* 141 (1): 014111.

Stefanović, M., E. Topić, A. M. Ivanisević, M. Relja, and M. Korsić. 2000. "Genotyping of CYP2D6 in Parkinson's Disease." *Clinical Chemistry and Laboratory Medicine: CCLM / FESCC* 38 (9): 929–34.

Taylor, Christopher, Ian Crosby, Vincent Yip, Peter Maguire, Munir Pirmohamed, and

Richard M. Turner. 2020. “A Review of the Important Role of in Pharmacogenomics.” *Genes* 11 (11). https://doi.org/10.3390/genes11111295.

Trendelkamp-Schroer, Benjamin, Hao Wu, Fabian Paul, and Frank Noé. 2015. “Estimation and Uncertainty of Reversible Markov Models.” *The Journal of Chemical Physics* 143 (17): 174101.

Trott, Oleg, and Arthur J. Olson. 2010. “AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading.” *Journal of Computational Chemistry* 31 (2): 455–61.

Voelz, Vincent A., Gregory R. Bowman, Kyle Beauchamp, and Vijay S. Pande. 2010. “Molecular Simulation of Ab Initio Protein Folding for a Millisecond Folder NTL9(1-39).” *Journal of the American Chemical Society* 132 (5): 1526–28.

Wang, Jingfang, Chengcheng Zhang, Dongqing Wei, and Yixue Li. 2010. “Docking and Molecular Dynamics Studies on CYP2D6.” *Chinese Science Bulletin = Kexue Tongbao* 55 (18): 1877–80.

Wang, Xiaoshuang, Jie Li, Guicheng Dong, and Jiang Yue. 2014. “The Endogenous Substrates of Brain CYP2D.” *European Journal of Pharmacology* 724 (February): 211–18.

Waring, Rosemary H. 2020. “Cytochrome P450: Genotype to Phenotype.” *Xenobiotica; the Fate of Foreign Compounds in Biological Systems* 50 (1): 9–18.

Wu, Hao, and Frank Noé. 2020. “Variational Approach for Learning Markov Processes from Time Series Data.” *Journal of Nonlinear Science* 30 (1): 23–66.

Wu, Jingjing, Xiaoqing Guan, Ziru Dai, Rongjing He, Xinxin Ding, Ling Yang, and Guangbo Ge. 2021. “Molecular Probes for Human Cytochrome P450 Enzymes: Recent Progress and Future Perspectives.” *Coordination Chemistry Reviews* 427 (213600): 213600.

Yeh, In-Chul, and Max L. Berkowitz. 1999. “Ewald Summation for Systems with Slab Geometry.” *The Journal of Chemical Physics* 111 (7): 3155–62.

Zanger, Ulrich M., and Matthias Schwab. 2013. “Cytochrome P450 Enzymes in Drug Metabolism: Regulation of Gene Expression, Enzyme Activities, and Impact of Genetic Variation.” *Pharmacology & Therapeutics* 138 (1): 103–41.