Molecular mechanisms of phase II metabolizing enzymes and ABC transporters, and their interactions with small molecules modeled through structure-based and machine learning methods

Balint Dudas

Université Paris Cité

ED 563 Médicament, Toxicologie, Chimie, Imageries

UMR 8038 Cibles Thérapeutiques et Conception de Médicaments
U1268 Chimie Médicinale et Recherche Translationnelle

Molecular mechanisms of phase II metabolizing enzymes and ABC transporters, and their interactions with small molecules modeled through structure-based and machine learning methods

Par **Bálint DUDÁS**

Thèse de doctorat de biologie, médecine et santé
Spécialité : Bioinformatique

Dirigée par Dr Maria MITEVA
Et par Dr David PERAHIA

Présentée et soutenue publiquement le 13 décembre 2022

Devant le jury composé de :

Dr Dominique DOUGUET	CR-HDR, Université Côte d'Azur	Rapporteur
Pr Matthieu MONTES	PU, Conservatoire National des Arts et Métiers	Rapporteur
Pr Ivet BAHAR	PU, University of Pittsburgh	Examinatrice
Pr Xavier DECLÈVES	PU-PH, Université Paris Cité	Examinateur
Dr Maria MITEVA	Directrice de recherche, Université Paris Cité	Directrice de thèse
Dr David PERAHIA	Directeur de recherche émérite, ENS Paris-Saclay	Directeur de thèse

Université Paris Cité

Doctoral School MTCI 563

UMR 8038 Therapeutic Targets and Drug Design
U1268 Medicinal Chemistry and Translational Research

Molecular mechanisms of phase II metabolizing enzymes and ABC transporters, and their interactions with small molecules modeled through structure-based and machine learning methods

by **Bálint DUDÁS**

Doctoral Thesis in Biology, Medicine, and Health
Specialty: Bioinformatics

Directed by Dr Maria MITEVA
and Dr David PERAHIA

Publicly defended on 13 December 2022

In front of the jury composed of:

Dr Dominique DOUGUET	CR-HDR, Université Côte d'Azur	Rapporteur
Pr Matthieu MONTES	PU, Conservatoire National des Arts et Métiers	Rapporteur
Pr Ivet BAHAR	PU, University of Pittsburgh	Examiner
Pr Xavier DECLÈVES	PU-PH, Université Paris Cité	Examiner
Dr Maria MITEVA	Research Director, Université Paris Cité	Thesis Director
Dr David PERAHIA	Emeritus Research Director, ENS Paris-Saclay	Thesis Director

Abstract

Title: *Molecular mechanisms of phase II metabolizing enzymes and ABC transporters, and their interactions with small molecules modeled through structure-based and machine learning methods*

Key words: *drug metabolizing enzymes, SULT, UGT, ABC transporters, BCRP, ABCG2, efflux mechanism, conformational exploration, molecular dynamics, normal modes, docking, machine learning*

The complex process of drug elimination is governed by drug metabolizing enzymes (DMEs) and transporters. Xenobiotics and endogenous compounds that should be eliminated from the human body can undergo phase I and/or phase II metabolism and then be excreted by efflux transporters. Phase I metabolism reactions primarily involve oxidation-reduction and are predominantly catalyzed by the cytochrome P450 enzymes, whereas phase II metabolism (conjugation) reactions include catalysis by, among others, sulfotransferases (SULTs) and UDP-glucuronosyltransferases (UGTs). Their conjugates are generally considered inactive and due to their reduced lipophilicity, they rely on transporters to cross cell membranes, such as ATP-binding cassette (ABC) transporters. Inhibition of DMEs and ABC transporters can lead to undesirable drug-drug interactions (DDI). Conformational changes are driving forces for the accommodation of the diverse ligands of the DMEs and for the substrate translocation of the different transporters due to their large promiscuity. Current machine learning (ML) models predicting the inhibition of DMEs and ABC transporters mostly neglect protein structure and dynamics, both being essential for the recognition of various substrates and inhibitors. To better understand their molecular mechanisms and their interactions with small molecules in all its complexity, we employed structure-based and ML approaches. In the present thesis work, the phase II DMEs SULT1A1 and UGT1A1, and the ABC transporter ABCG2 (BCRP) are studied in detail.

SULT1A1 catalyzes the sulfoconjugation from the cofactor 3′-Phosphoadenosine 5′-Phosphosulfate (PAPS). We performed molecular dynamics (MD) and the recently developed MD with excited Normal Modes (MDeNM) simulations which allowed an extended exploration of the conformational space of the PAPS-bound SULT1A1. The generated ensembles combined with the docking of SULT1A1 ligands shed new light on its substrate and inhibitor binding mechanism. Unexpectedly, our simulations demonstrated that large conformational changes of the PAPS-bound SULT1A1 could occur. Our results suggest that a wide range of drugs could be recognized by the PAPS-bound SULT1A1 independently of the cofactor presence and highlight the utility of including MDeNM in protein-ligand interactions studies where major rearrangements are expected.

UGT1A1 catalyzes the covalent addition of the glucuronic acid sugar moiety from the cofactor uridine-diphosphate glucuronic acid (UDPGA). Strong inhibition of UGT1A1 may trigger adverse drug interactions, or result in endobiotic metabolism disorders. We performed MD

simulations on a human UGT1A1 homology model and created, to the best of our knowledge, the first prediction models of UGT1A1 inhibition by integrating information on UGT1A1 structure and dynamics, interactions with diverse ligands, and ML methodologies. Our models can be helpful for the prediction of DDI of new drug candidates.

ABCG2 (BCRP) is involved in multidrug resistance (MDR), understanding its complex efflux mechanism is essential to prevent MDR and DDI. ABCG2 export is characterized by two major conformational transitions between inward- and outward-facing states, the structures of which have been resolved. We developed an innovative enhanced MD simulation approach, 'kinetically excited targeted MD', and successfully simulated the transitions between the inward- and outward-facing states in both directions and the transport of the endogenous substrate estrone 3-sulfate. We discovered an additional pocket between the two substrate-binding cavities and found that the presence of the substrate in the first cavity is essential to couple the movements between the nucleotide-binding and transmembrane domains. The generated transient conformations and the revealed translocation pathway can facilitate the identification of novel ABCG2 substrates and inhibitors, and the probing of new drug candidates for MDR and DDI.

Résumé

Titre : *Mécanismes moléculaires des enzymes de métabolisme de phase II et des transporteurs ABC, et leurs interactions avec de petites molécules modélisées par des méthodes structurales et d'apprentissage automatique*

Mots clés : *enzymes métabolisant des médicaments, SULT, UGT, transporteurs ABC, BCRP, ABCG2, mécanisme d'efflux, exploration conformationnelle, dynamique moléculaire, modes normaux, arrimage moléculaire, apprentissage automatique*

Le processus d'élimination des médicaments est régi par des enzymes métabolisant des médicaments (DME) et des transporteurs. Les xénobiotiques sont métabolisés dans le corps humain par des enzymes de phase I et/ou de phase II, et excrétés par des transporteurs. Les réactions de phase I impliquent principalement l'oxydation-réduction faisant intervenir les enzymes de cytochrome P450, tandis que les réactions de phase II (conjugaison) incluent le plus souvent les sulfotransférases (SULT) et les UDP-glucuronosyltransférases (UGT). Leurs conjugués sont généralement considérés comme inactifs et sont éliminés à travers la membrane par des transporteurs tels que les transporteurs ABC (ATP-binding cassette). L'inhibition des DME et transporteurs ABC peut entraîner des interactions médicamenteuses indésirables (DDI). Les changements conformationnels des DME et transporteurs sont essentiels pour la transformation et la translocation du substrat. Les modèles actuels d'apprentissage automatique (ML) prédisant l'inhibition des DME et des transporteurs ABC

négligent principalement leur structure et dynamique, toutes deux essentielles à la reconnaissance de divers substrats et inhibiteurs. Pour mieux comprendre leurs mécanismes moléculaires et leurs interactions avec les petites molécules, nous avons utilisé des approches structurales et ML. Dans cette thèse, les DME de phase II SULT1A1 et UGT1A1, et la protéine ABCG2 (BCRP) sont étudiés en détail.

SULT1A1 catalyse la sulfoconjugaison par le cofacteur 3′-Phosphoadénosine 5′-Phosphosulfate (PAPS). Nous avons effectué des simulations de dynamique moléculaire (MD) et de MDeNM (MD with excited Normal Modes), permettant une exploration étendue de l'espace conformationnel du SULT1A1 lié au PAPS. Les ensembles générés, combinés au criblage des ligands de SULT1A1, ont apporté un nouvel éclairage sur le mécanisme de fixations des composés. De manière inattendue, nos simulations ont montré que de grands changements conformationnels de SULT1A1 contenant le PAPS peuvent se produire. Elles suggèrent qu'une large gamme de médicaments peuvent être reconnus par SULT1A1 avec ou sans le cofacteur et soulignent l'utilité d'inclure MDeNM dans les études d'interaction protéine-ligand où des réarrangements majeurs sont attendus.

UGT1A1 catalyse l’inclusion de la partie sucre de l'acide glucuronique provenant du cofacteur acide uridine-diphosphate glucuronique (UDPGA). Une inhibition de l'UGT1A1 peut conduire à des DDI, ou à des troubles métaboliques endobiotiques. Nous avons effectué des simulations MD sur un modèle de l'UGT1A1 humain et créé les premiers modèles prédictifs de l'inhibition de cette protéine en intégrant des informations structurales et de dynamique, les interactions avec divers ligands et des techniques de ML. Nous montrons l’utilité de notre approche pour la prédiction DDI de nouveaux candidats médicaments.

ABCG2 est impliqué dans la multirésistance aux médicaments (MDR). La compréhension de son mécanisme pour la translocation de composés est essentielle pour prévenir la MDR et la DDI. Il est caractérisé par deux transitions conformationnelles majeures entre un état ouvert vers l'intérieur de la cellule et l’autre vers l'extérieur, dont les structures ont été résolues expérimentalement. Nous avons développé une approche de simulation MD innovante, 'kinetically excited targeted MD', pour simuler les transitions entre ces deux états et le transport du substrat endogène estrone 3-sulfate. Nous avons caractérisé une cavité supplémentaire entre les deux cavités de fixation du substrat, et mis en évidence que la présence du substrat dans la première cavité est essentielle pour coupler les mouvements entre le domaine nucléotidique et le domaine transmembranaire. Les structures transitoires dans le processus de translocation qui ont été générées seront utiles à l'identification de nouveaux substrats et inhibiteurs d'ABCG2, et la prédiction de la MDR et DDI de nouveaux médicaments candidats.

Acknowledgements

First and foremost I am extremely grateful to my supervisors, *Maria Miteva* and *David Perahia* for your invaluable advice and continuous support during my PhD research. By undertaking the task of supervising my work you have demonstrated to me on the one hand how to conduct research profoundly by willing to understand the different phenomena to the smallest details and on the other hand how to do it efficiently. Maria, you went above and beyond, thank you for introducing me to the world of computational drug design. David, it is always an honor to have a discussion with you over a coffee, whether it be about science or any other aspect of life.

I would like to say thank you to my present and past colleagues in the Miteva lab, in particular *Youcef Bagdad*, it was always very pleasant to discuss with you about our docking results and I appreciate that you tolerated my DJ skills, *Dani Toth*, you brought a hint of homeland to Paris during your stay with us, and *Elodie Goldwaser*, you really helped me arrive and integrate into the group. My appreciation also extends to the colleagues who helped me and guided me throughout these years at the Therapeutic Targets and Drug Design (CiTCoM) group in the Faculty of Pharmacy of Paris and at the Laboratory of Applied Biology and Pharmacology in ENS Paris-Saclay, with special thanks to you, *Marco Pasi*, for making me feel welcome at ENS.

One of the biggest professional challenges of my stay at the Université Paris Cité was my teaching assignment in Physical Chemistry, in particular the laboratory sessions. *Philippe Espeau*, thank you for your confidence and encouragement. *Yohann Corvis* and *Mathieu Lazerges* are welcomed for their guidance and practical advice, and *Monique Cadasse* and *José Nupert* are recognized for their continuous management of the laboratories.

I very much appreciated the helpful discussions and the continuous cooperation with *Erika Balog*, my former supervisor at Semmelweis University. You were there at the very first steps of my academic journey and I hope that our collaboration continues in the years to come. You helped me enormously to evolve along the path of becoming a true scientist.

My dear *Sister*, I appreciate that I have someone who kept repeating for the last one and a half years that our family housed one single Doctor exclusively, *you*. *Mama*, *Papa*, and *my whole family*, it is originally thanks to you that I became the person who would start this adventure in France in the first place. Your interest in my scientific activity is a constant motivation.

Finally, last but not least, *Gabor Zalanki*, it is your ten minutes coming. But those ten minutes would not be enough to express all my gratitude for being there for me throughout the PhD and for everything that you have done for me. Your vocabulary, grammar, and syntax insights together with your incredible creativity and graphical skills contributed enormously to the quality of my present and past works. Besides that, your continuous support and comfort made these amazing three years turn into a real adventure. Thank you!

Abbreviations

ABC transporter	*ATP-Binding Cassette transporter*
ADME-Tox	*Absorption, Distribution, Metabolism, Excretion, Toxicity*
ADR	*Adverse Drug Reactions*
AI	*Artificial Intelligence*
ATP	*Adenosine Triphosphate*
AUC	*Area Under the Curve*
BA	*Balanced Accuracy*
BBB	*Blood-Brain Barrier*
CNS	*Central Nervous System*
CV	*Cross Validation*
CYP	*Cytochrome P450 Enzyme*
DDI	*Drug-Drug Interactions*
DME	*Drug Metabolizing Enzyme*
E1S	*Estrone-3-Sulfate*
E2	*17β-Estradiol*
ER	*Endoplasmic Reticulum*
ER	*Endoplasmic Reticulum*
FAD	*Flavin Adenine Dinucleotide*
FMO	*Flavin-containing Monooxygenase*
GST	*Glutathione-S-transferase*
MAO	*Monoamine Oxidase*
MATE	*Multidrug And Toxin Extrusion*
MCC	*Matthews Correlation Coefficient*
MD	*Molecular Dynamics*
MDR	*Multidrug Resistance*
ML	*Machine Learning*
MLR	*Multiple Linear Regression*
MRP	*Multidrug Resistance-associated Protein*
NADPH	*Nicotinamide Adenine Dinucleotide Phosphate (reduced form)*
NBD	*Nucleotide-Binding Domain*
NMA	*Normal Mode Analysis*
NSAID	*Non-Steroidal Anti-Inflammatory Drug*

OAT	*Organic Anion Transporter*
OATP	*Organic Anion-Transporting Polypeptides*
OCT	*Organic Cation Transporter*
OOB error	*Out-of-Bag error*
PAP	*3'-Phosphoadenosine 5'-Phosphate*
PAPS	*3'-Phosphoadenosine 5'-Phosphosulfate*
PCA	*Principal Component Analysis*
PEPT	*Peptide Transporter*
P-gp	*P-glycoprotein*
PLS	*Partial Least Squares*
QM/MM	*Quantum Mechanics/Molecular Mechanics*
(Q)SAR	*(Quantitative) Structure-Activity Relationship*
RF	*Random Forest*
Rgyr	*Radius of Gyration*
RMSD	*Root Mean Square Deviation*
RMSF	*Root Mean Square Fluctuation*
ROC	*Receiver Operator Characteristic*
SLC transporter	*Solute Carrier transporter*
SNP	*Single Nucleotide Polymorphism*
SOM	*Site of Metabolism*
SULT	*Cytosolic Sulfotransferase Enzyme*
SVM	*Support Vector Machine*
TMD	*Transmembrane Domain*
UDPGA	*Uridine-Diphosphate Glucuronic Acid*
UGT	*Uridine 5'-Diphosphate-Glucuronosyltransferase Enzyme*

Table of Contents

I. Introduction

„Nem én vagyok bonyolult, hanem a dolog, amiről beszélek.”

“It’s not me who is complicated, it’s the things I talk about.”

Karinthy Frigyes

A. Biological Background

Drug discovery and development is an expensive and slow process. According to the study of DiMasi et al. in 2016, development of a single drug cost $2.6 billion and took over 10 years on average [1]. A major challenge associated with the identification of promising drug candidates is to find a good balance between the required efficacy, selectivity, and affinity against their intended therapeutic target while at the same time showing appropriate absorption, distribution, metabolism, excretion, and toxicity (ADME-Tox) properties. ADME-Tox is a complex process that determines the pharmacokinetics of a drug molecule in the body which includes both transporters and drug metabolizing enzymes (DMEs) with physiological consequences on pharmacological and toxicological effects [2].

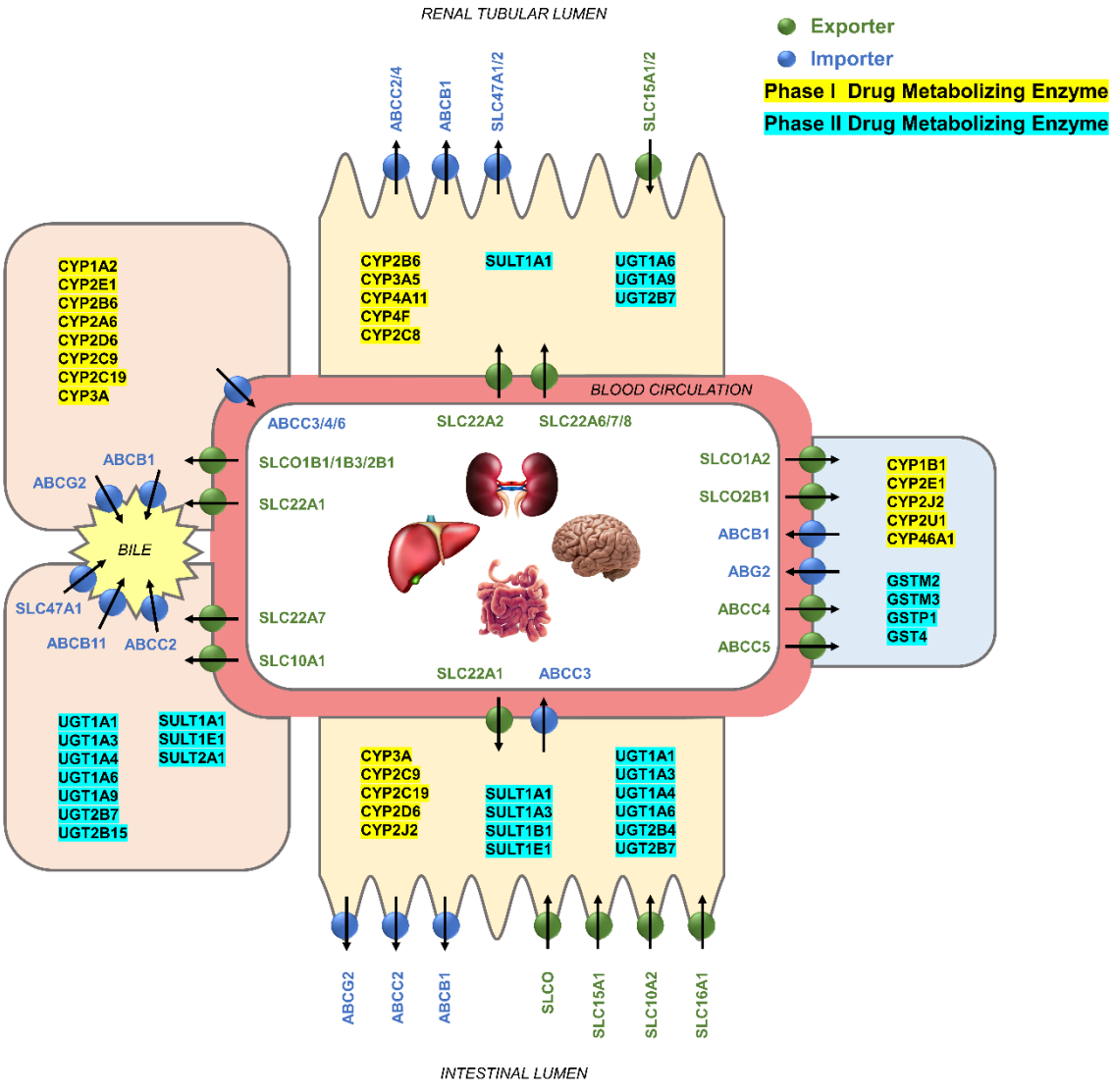

Figure A.1: Drug transport and metabolism at the most important drug metabolism sites (liver, kidneys, intestine) and at an important tissue barrier (blood brain barrier).

1. Drug elimination from the human body

Drug elimination through metabolism and excretion is a complex process that is governed by metabolizing enzymes and membrane transporters. Metabolism in the body is a

biotransformation process where compounds are structurally modified to different molecules (metabolites) by various metabolizing enzymes. Endogenous compounds as well as drugs and other xenobiotics that should be eliminated from the human body can undergo phase I and/or phase II metabolism catalyzed by drug metabolizing enzymes (DMEs). In general, DMEs contribute to the protection of the body against harmful compounds, both against xenobiotics (compounds from the environment, substances that are foreign to the given living organism; from the Greek *xenos* meaning 'stranger' and *biotic* 'related to living beings') and certain endogenous compounds that appear during different biological processes (compounds that originate within the given living organism). Such enzymes play a key role in the detoxification of organisms, and in some cases in the bioactivation of some so-called prodrugs, by chemically modifying toxic substances and that way they generally make them less active or even inactive. Following phase I and/or phase II metabolism, or even directly, compounds can be excreted to the extracellular space by transporters. Together with the metabolizing enzymes, efflux transporters modulate the intracellular bioavailability and pharmacokinetics of drugs and other xenobiotics. Drug metabolism influences drug pharmacodynamics and pharmacokinetics.

The rate of metabolism of a given drug, which can be highly influenced by different DMEs, as well as its rate of export are important factors in drug design. Both have a crucial impact on the intensity and the duration of the drug's effect in the body. The metabolic reprogramming and the resulting metabolic shift in certain tumorous cells (e.g. by the overexpression of DMEs and drug exporters) alters the rate of metabolism of several drugs and can contribute to the multidrug resistance (MDR) phenotype. MDR promotes resistance to drugs of different structures and mechanism of action [3]. When DMEs are overexpressed in tumorous tissues, they can cause resistance by metabolizing the drugs and rendering them inactive whereas the overexpression of efflux transporters and the resulting increase in the export activity can significantly impair treatment success by effluxing drugs from the cytosol [4]. Furthermore, the effects of some drugs as substrates (through the creation of toxic metabolites) or inhibitors of DMEs and drug transporters can cause high level of intracellular toxicity and are a common reason for hazardous adverse drug reactions (ADRs, unintended harmful effects) and drug-drug interactions (DDIs, alterations in the way a given drug acts when applied together with other drugs).

The inhibition of DMEs and drug transporters can directly increase intracellular toxicity while the formation of reactive or toxic metabolites are also a safety liability. Furthermore, high affinity substrates of a given efflux transporter or a given DME can have very low intracellular efficacy. As a consequence, the prediction of interactions with DMEs and drug transporters and associated drug side effects in early drug development stages can help reducing drug candidate failure. An efficient *in silico* prefiltering can also significantly decrease the experimental toxicity evaluation of a large number of poor drug candidates which are both costly and currently primarily rely on animal studies raising also ethical aspects.

2. Drug metabolism reactions

A large portion of drugs are small molecules that are mostly classified as xenobiotics. Nevertheless, several endogenous molecules, such as steroids and hormones, are also used for the treatment of certain disease conditions and are also referred to as drugs. The term *metabolism* refers to the process of biotransformation of chemicals catalyzed by an enzyme [5]. The primary objective of metabolism is the excretion of both endogenous and exogenous molecules from the body generally by converting hydrophobic compounds into more hydrophilic metabolites promoting their elimination. DMEs play a fundamental role through metabolic reactions in the detoxification and elimination of drugs and other xenobiotics introduced into the human body [6]. Most of the human tissues and organs express diverse and various DMEs either at their basal unstimulated level or at elevated levels in a response to exposure to xenobiotics [7]. The principal metabolic site can be found in the liver, more precisely in its hepatocytes [8], but other metabolic sites include the kidneys and the lungs. Within the cells, most DMEs are either anchored to the endoplasmic reticulum (ER) membrane (e.g. majority of cytochrome P450 enzymes, epoxide hydrolases, glucuronosyl transferases, or glutathione S-transferases) or are cytosolic enzymes (e.g. aldehyde oxidase, alcohol dehydrogenase, xanthine oxidase, or sulfotransferases) [9-11].

The rate of metabolism of a given drug is an important aspect upon drug design and development. It is necessary to find a good balance so that the drug can carry out its effects in the body without being rapidly eliminated and reaching its desired target; but at the same time it should also not accumulate to toxic levels in the body over time. Also, in some cases, drug metabolism can generate more active metabolites than the corresponding parent drug. If this mechanism is exploited intentionally (approximately 5 % of all drugs are design in this way), the parent drugs are called prodrugs [12]. An example for prodrugs is acetylsalicylate, better known as aspirin, which is a synthetic prodrug of the metabolite salicylate which is responsible for most of the anti-inflammatory and analgesic effects of aspirin. (Besides aspirin being a prodrug, aspirin itself can be an active moiety with antiplatelet-aggregating effects) [13]. Similarly, lovastatin, a drug administered to treat high blood cholesterol and reduce the risk of cardiovascular diseases, is a prodrug which is converted to its active (open ring) form only at the site of action by *in vivo* hydrolysis [14]. In such cases the administrated drug has a lower activity and usually is better absorbed from the gastrointestinal tract [15, 16]. At the same time, reactive metabolites also arise from adverse or unintended drug reactions and can cause serious side effects, even interacting with macromolecules like the DNA and induce mutagenesis and cancer [17].

Drug biotransformation reactions are classified as either phase I (functionalization) reactions, or phase II (conjugation) reactions [18, 19]. In many cases phase I reactions produce a reactive functional group on a given molecule so that it can be attacked by phase II DMEs. However, numerous compounds can be directly conjugated by phase II DMEs, without a preceding phase I reaction (e.g. bilirubin, steroids, or paracetamol [20]).

2.1. Phase I metabolism

Phase I metabolism (functionalization, conversion of functional groups) includes oxidation, reduction, hydrolysis, and hydration reactions as well as some rarer reactions like isomerization, dimerization, or decarboxylation. Many pharmacologically active molecules are of hydrophobic nature and are unionized or only partially ionized at physiological pH. Without biotransformation they could be reabsorbed after glomerular filtration in the kidney and remain in the body [6]. Phase I DMEs aim to unmask a polar functional group on their substrates. Phase I metabolites may be directly eliminated or further metabolized by a consecutive conjugation reaction step catalyzed by phase II DMEs which render their substrates even more hydrophilic.

2.1.1. Cytochrome P450 superfamily

The most important phase I DMEs belong to the cytochrome P450 (CYP; EC1.14.14.1) superfamily that are capable of catalyzing the oxidative biotransformation of most drugs and other lipophilic xenobiotics [21]. CYP enzymes catalyze several reactions, including oxidation, aromatic hydroxylation, aliphatic hydroxylation, N-dealkylation, and O-dealkylation[5], oxidation being the primary reaction which results in the addition of one or more oxygen atoms to the parent drug [19]. The human CYP superfamily contains 57 functional genes (and 58 pseudogenes which are nonfunctional segments of the DNA that resemble functional genes). Cytochrome P-450 is the terminal oxidase component of an electron transfer system present in the endoplasmic reticulum (imbedded in the membrane) and is classified as a (non-covalently bound) haem-containing enzyme superfamily (haemoproteins) [18]. The P450 enzymes dominate the metabolism of drugs, the different CYP isoforms metabolize almost three-quarters of all clinical drugs [22, 23]. Based on protein sequence homology CYPs are classified into 18 families and 44 subfamilies. Most of the human CYP isoforms have specific endogenous functions e.g. the biosynthesis of steroid hormones, prostaglandins, or bile acids [24]. Even though the CYP isoforms exhibit broad and overlapping substrate specificities, most drugs are metabolized by one or few enzymes only [25]. Almost exclusively members belonging to three specific families (CYP1, CYP2, and CYP3) catalyze drug metabolism, they are accountable for the oxidative metabolism of more than 90 % of all drugs catalyzed by the CYP superfamily (46 % is catalyzed by the members of CYP3A, 16 % by CYP2C9, 12 % by both CYP2C19 and CYP2D6, 9% by members of the CYP1A family, and 2-2 % by both CYP2B6 and CYP2E1) [22, 26].

The CYP1 family members have overlapping catalytic activities and include hydroxylation and other oxidative transformations of many polycyclic aromatic hydrocarbons and other aromatic substances. In the CYP2 family, members of the CYP2C subfamily are of particular interest, which consist of the four highly homologous genes (>82 %) CYP2C18, CYP2C19, CYP2C9, and CYP2C8, yet each of them has a very unique substrate specificity and role in drug metabolism [27]; while CYP2D6 is the only protein-coding gene in the CYP2D subfamily, and its hepatic protein content varies dramatically from person to person mainly due to its genetic polymorphism [28, 29]. The human CYP3 family has one single subfamily, CYP3A, consisting of four genes, 3A4, 3A5, 3A7, and 3A43 [25].

CYP3A4 is the most abundant isoform in the liver and small intestine and plays a role in the metabolism of over 30 % of all drugs on the market and in development in almost all therapeutic categories [30], its substrates are large and lipophilic molecules of very diverse structures. Other important DMEs of the CYP1 and CYP2 families are CYP1A2, CYP2C9, CYP2C19, and CYP2D6. The isoform CYP1A2 typically metabolizes smaller and planar, aromatic, polyaromatic, and heterocyclic amides and amines, its typical biotransformations include *7*-ethoxyresorufin *O*-deethylation (breaking of a covalent bond between a substrate and its -CH_2CH_3 group), phenacetin *O*-deethylation, and caffeine *N3*-demethylation (breaking of a covalent bond between a substrate and its -CH_3 group) to paraxanthine [8, 31]. The major enzyme CYP2C9 of the CYP2C subfamily metabolizes more than 15 % of clinically administrated drugs [32] and typically interacts with weakly acidic substances having a hydrogen bond acceptor (including most nonsteroidal anti-inflammatory drugs) [33]. The isoform CYP2C19 on the other hand accepts neutral or weakly basic molecules or amides of higher molecular weight, possessing 2 or 3 hydrogen bond acceptors (including most proton pump inhibitors), and also plays an important role in the metabolism of numerous first- and second-generation antidepressants. Finally, the isoform CYP2D6 typically transforms basic molecules with a protonatable nitrogen atom 4-7 Å from the metabolism site (including many plant alkaloids and antidepressants, and other nervous system drugs) [25, 34-36].

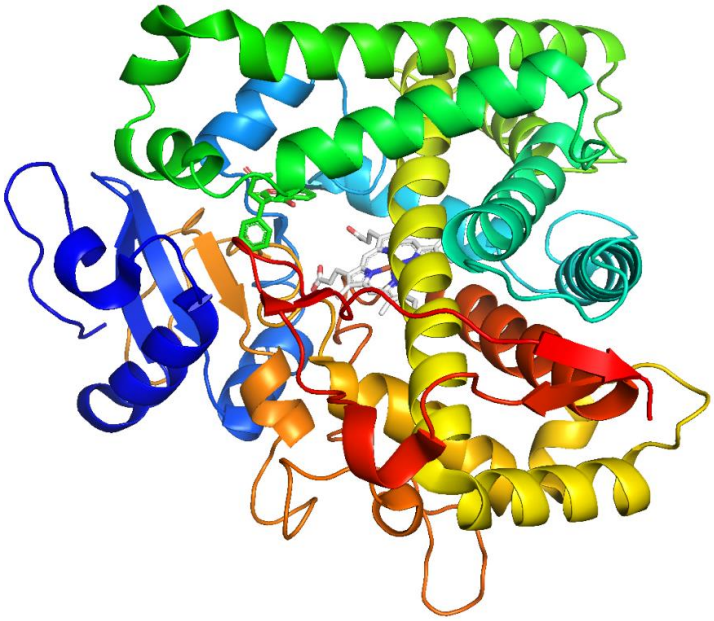

Figure A.2: Crystal structure (PDB 1OG5) of CYP2C9 in the presence of heme (white sticks) and warfarin (green sticks).

2.1.2. Other phase I DMEs

Apart from the CYP superfamily, several other phase I DMEs can contribute to the elimination of drugs. Flavin-containing monooxygenases (FMOs) oxygenate nucleophilic O, N, S, and Se atoms of a wide range of substrates, such as amines, amides, thiols, sulfides, and phosphites [37]. The enzymatic mechanism of FMOs differ from other monooxygenases as they only require the cofactor NADPH for the enzymatic activity on their substrates, as the prosthetic

group FAD (flavin adenine dinucleotide, a redox-active coenzyme associated with various proteins) is an integral part of the protein, FMOs are flavoproteins [38]. Monoamine oxidases (MAOs) were identified to catalyze the oxidative degradation of a number of neurologically important amine substrates like dopamine or serotonin. MAOs are also flavoenzymes containing a single covalent FAD cofactor per monomer [39]. Molybdenum-containing hydroxylases, unlike monooxygenases, use oxygen derived ultimately from water to catalyze the hydroxylation of carbon centers as the source of the oxygen atom incorporated into the product, rather than dioxygen [40]. Further examples of non-CYP450 phase I DMEs are alcohol and aldehyde dehydrogenases (that catalyze the general oxidation reaction of the hydroxyl or formyl group (aldehyde) via an electron acceptor [41, 42]), aldo-keto reductases (that catalyze redox transformations; substrates of the family include glucose, steroids, glycosylation end-products, and lipid peroxidation products [43]), NADPH:quinone reductases (that catalyze the two electron reduction of quinones and a wide range of other organic compounds [44]), and hydrolytic enzymes (alternatively referred to as hydrolases, they split different groups of biomolecules such as esters, peptides, and glycosides breaking them down into their simplest units [45, 46]).

2.2. Phase II metabolism

Phase II metabolism (conjugation, addition of functional groups) links a relatively large endogenous polar group to diverse types of compounds, generally creating water-soluble products with increased molecular weight which can be excreted in bile or urine [5]. Phase II metabolism includes sulfation, glucuronidation, glycosidation, methylation, acetylation, condensation, and amino acid-, glutathione-, and fatty acid-conjugation. The resulted metabolites are typically of reduced membrane permeability, consequently active transport is required for their eventual excretion. Most conjugation reactions are cytosolic with the exception of glucuronidation which takes place in the lumen of the ER. In most cases conjugation reactions terminate the biological activity of drugs. However, reactive conjugated metabolites have also been reported (e.g. many glucuronide conjugates of opioids, steroid sulfates, morphine-6-glucuronide, or conjugates of midazolam) [47-50]. Phase II metabolism can follow the unmasking of a polar functional group by phase I DMEs. Nevertheless, numerous compounds (e.g. bilirubin, steroids, or paracetamol) can be directly conjugated by phase II DMEs, without a preceding phase I reaction [51, 52]. The catalytic rates of phase II DMEs are generally significantly higher than the rates of CYPs, if phase II is preceded by such a phase I catalysis, the rate limiting step is usually the oxidation reaction [47, 53].

2.2.1. Sulfation

Sulfation is one of the major conjugating pathways responsible for the detoxification and subsequent elimination of xenobiotics and endogenous small molecules [54]. Sulfation is also important in the biosynthesis of steroid hormones and in modulating signaling pathways involving thyroid hormones, steroids, and sterols [55-57]. Sulfation (or sulfoconjugation) primarily involves phenol substrates, but can also occur for alcohols, amines, and thiols [18]. Sulfation reactions are catalyzed by sulfotransferases, two large groups of sulfotransferases

have been identified: membrane-bound enzymes involved in the metabolism of endogenous peptides, proteins, glycosaminoglycans, and lipids, and the cytosolic sulfotransferase enzymes (SULTs). SULTs are a supergene family catalyzing the transfer of the sulfonate (SO_3^-) group from the co-factor 3′-Phosphoadenosine 5′-Phosphosulfate (PAPS) to a hydroxyl or amino group of substrates [58-60]. The cofactor PAPS is synthesized in a two-step reaction from inorganic sulfate and ATP by PAPS synthetases enzymes. Human cytosolic SULTs are divided into 4 families (SULT1, SULT2, SULT4, and SULT6), they include 13 known enzymes [60]. Expression of the different SULT enzymes in human occurs almost in every organ and are most commonly present in liver, gut, breast, lung, adrenal glands, kidney, blood cells, brain, and placenta [5].

SULTs are promiscuous enzymes with only some degree of substrate selectivity [61, 62]. The major isoforms involved in drug metabolism are SULT1A1, SULT1A3/4 (the SULT1A3 and SULT1A4 genes arose from a gene duplication event, yet despite the slight sequence variations, they encode identical proteins), SULT1B1, SULT1E1, and SULT2A1.

The isoform SULT1A1 (also known as the thermostable phenol sulfotransferase) has the broadest substrate specificity within the SULT superfamily, and it displays an extensive tissue distribution [57]. It accounts for over 50 % of the total SULT protein content in the liver, it is also physiologically expressed in the kidneys, small intestine, and the lungs [63]. It catalyzes both endogenous and xenobiotic phenolic molecules (e.g. estradiol or isoflavones) with high affinity [64]. A large number of drugs approved by the United States Food and Drug Administration (FDA) were predicted as substrates of SULT1A1 by an *in silico* study [65], with experimentally validated examples like paracetamol, levodopa, opioid drugs, or fulvestrant [58, 66, 67]. Additionally, SULT1A1 catalyzes the metabolism of numerous environmental mutagens and carcinogens which, as a result, can be either detoxified or in some cases even activated [68].

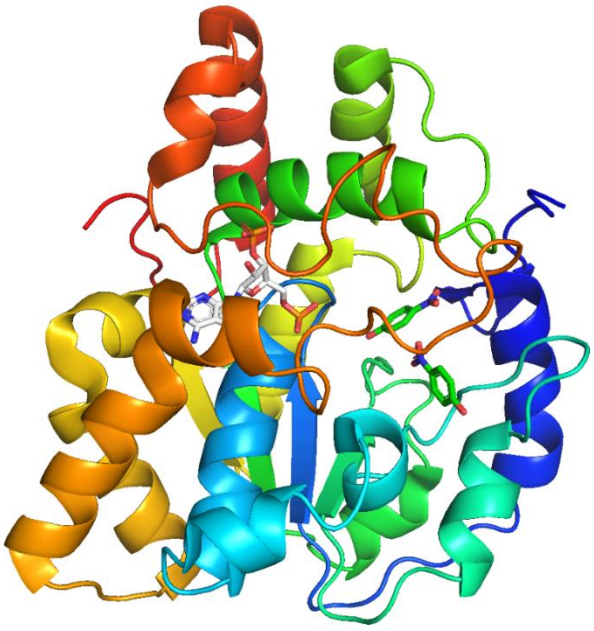

Figure A.3: Crystal structure of SULT1A1 (PDB 1LS6) in the presence of PAP (white sticks) and p-Nitrophenol (green sticks).

SULT1A3 (or SULT1A3/4) is encoded by both the SULT1A3 and SULT1A4 genes [69]. It possesses a unique glutamate residue (E146) which together with another carboxylic-group-containing residue (D86) ensures a high selectivity for catecholamines (monoamines that have a catechol (benzene with two neighboring hydroxyl side groups) and a side-chain amine), assumingly by forming a salt bridge with the nitrogen on the catecholamine side chain. Catecholamines that are substrates to SULT1A3 include dopamine, serotonin, adrenaline, and noradrenaline [70]. In adults, SULT1A3 is a major extrahepatic (situated outside the liver) enzyme, with especially high expression in the gastrointestinal tract [71], having a direct effect on the oral bioavailability of some drugs [61, 63].

SULT1B1 and SULT1A1 show almost identical substrate specificity profiles with SULT1B1 having considerably lower catalytic affinity in most cases [57]. Nevertheless, thyroid hormones are primarily sulfonated by SULT1B1 [72, 73]. It is the most expressed sulfotransferase in the gastrointestinal tract, and it can also directly affect the bioavailability of some drugs [61, 74].

SULT1E1, also known as estrogen sulfotransferase, has a unique role in hormone homeostasis and biosynthesis as it sulfonates both estrogens and iodothyronines (iodinated derivatives of thyronine). It has a very high selectivity and affinity for estrogens, like 17β-estradiol. It is the most abundant SULT isoform in the lungs, however it is only expressed at relatively low levels in the liver and small intestine [63]. The sulfoconjugation of estrogens inhibits their interaction with the estrogen receptor and so it modulates the biological function of these hormones [75].

SULT2A1 is widely expressed in human tissues. It is the second most abundant sulfotransferase in the liver after SULT1A1. However, its expression levels are considerably lower in other tissues like the kidney, the lungs, and the small intestine [57]. SULT2A1 catalyzes the sulfoconjugation at hydroxyl groups of bile acids and different steroids. Sulfonation by SULT2A1 is the primary pathway in bile acid elimination in humans which takes place in the liver [76]. Sulfated steroids are circulating precursors for the biosynthesis of receptor-active hormones such as 17β-estradiol, testosterone, and dihydrotestosterone. Together with the biosynthetic enzymes, desulfation by steroid sulfatases and transport of the steroid sulfates (in and out of the cell) by active transporters are responsible for delicately maintaining the tissue levels of active steroid hormones [55, 57].

2.2.2. Glucuronidation

The primary sugar conjugation route in humans is glucuronidation (conjugation with α-D-glucuronic acid) although conjugation with glucose, xylose, and ribose are also possible (e.g. in insects, conjugation with glucose is more prevalent than with glucuronic acid, it is also of importance in plants, and can also be found in mammals, however only to a limited extent) [18]. Apart from sulfonation, glucuronidation is the other major phase II reaction type in humans. Uridine 5'-diphosphate-glucuronosyltransferases (UGTs), that catalyze the conjugation of glucuronic acid to a nucleophilic substrate to form glucuronides (by a second order nucleophilic substitution reaction [77]), are one of the major classes of conjugative enzymes involved in phase II drug metabolic reactions [47]. Most human UGTs are

physiologically highly expressed in the liver, the primary site of xenobiotics metabolism, but are also present in other tissues like the intestine, the kidneys, the stomach, and the lungs [78]. The broad occurrence of glucuronidation (present in all tissues of the mammalian body) is probably due to the cofactor (or cosubstrate) involved in the catalytic reaction. The high energy donor uridine-diphosphate glucuronic acid (UDPGA) is part of intermediary metabolism (reactions concerned with the storage and generation of metabolic energy, required for the biosynthesis of low-molecular weight compounds and energy storage compounds [79]) and is closely related to glycogen synthesis. The formation of glucuronide metabolites is quantitatively the most important form of conjugation for drugs as well as endogenous compounds, and concerns alcohols, phenols, hydroxylamines, carboxylic acids, amines, sulfonamides, and thiols. [18]. Numerous nucleophiles are capable of being glucuronidated, the corresponding reactions are divided into four types (O-glucuronides, N-glucuronides, S-glucuronides, and C-glucuronides).

In humans there are 19 genes encoding UGT enzymes that are classified into two families and three subfamilies, namely UGT1A, UGT2A, and UGT2B. More recent studies have also focused on other UGT families, UGT3 and UGT8, that opposed to the members of the families UGT1 and UGT2, utilize different UDP-sugar cofactors. UGT1 and UGT2 members primarily use UDP-glucuronic acid in glycosidation reactions whereas UGT3 family members use UDP-glucose, UDP-xylose, and UDP-N-acetylglucosamine, and UGT8 uses exclusively UDP-galactose. The term UGT traditionally only covered UDP-glucuronosyltransferases (the families UGT1 and UGT2, catalyzing the conjugation of glucuronic acid [80]), however it is often used to describe all UDP-glycosyltransferases (UGT1, UGT2, UGT3, and UGT8, expressing 22 isoenzymes in humans [81]). Members of the UGT1 and UGT2 play an important role in pharmacology and toxicology, also contributing to interindividual differences in drug response and cancer risk. These UGTs are physiologically highly expressed in organs of detoxification, whereas UGT3 and UGT8 conjugate relatively few pharmacological agents, their contributions to drug metabolism appear to be relatively minor [81].

UGTs are localized within the ER lumen, anchored to the membrane. UGT location results in a latency of their enzymatic activity, and require specific transporters for both the cofactor and the conjugation products [82]. Glucuronides generated by UGTs leave the lumen and appear in the cytosol for being pumped into the bile or into the blood by ABC transporters of the plasma membrane [83].

UGTs metabolize a wide range of compounds and they have overlapping substrate specificity. In terms of drug metabolism, the clinically most important hepatic UGTs are UGT1A1, UGT1A3, UGT1A4, UGT1A6, UGT1A9, and UGT2B7 [77].

The isoenzyme UGT1A1 is of particular importance considering its exclusive role in the glucuronidation, and therefore the detoxification of the endogenous heme breakdown by-product, bilirubin [84, 85]. Both substrates and inhibitors of UGT1A1 exhibit a large variety in their sizes and nature. It is physiologically expressed in numerous tissues including the liver, intestine, colon, biliary tissue, and stomach [86]. Unconjugated hyperbilirubinemias such as

Gilbert's syndrome and Crigler-Najjar syndromes have been associated with UGT1A1 polymorphisms [87]. Additionally, impaired UGT1A1 enzymatic activity was shown to be related to the toxicity of several drugs [88] (e.g. irinotecan (used to treat colon cancer and small cell lung cancer) [89], lamotrigine (used to treat epilepsy and stabilize mood in bipolar disorder) [90], raloxifene (used to prevent and treat osteoporosis in postmenopausal women and those on glucocorticoids), or protease inhibitors [91]).

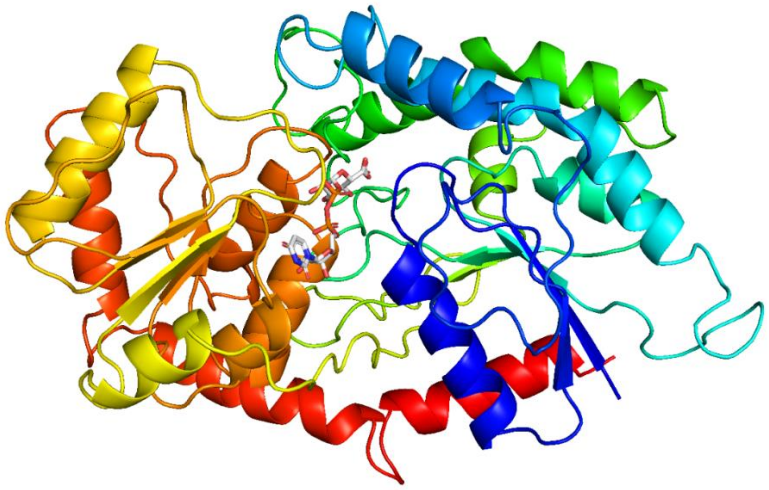

Figure A.4: A homology model (Locuson and Tracy, 2007) of the human UGT1A1 in complex with UDPGA (white sticks).

Several UGT1 isoforms have activity toward steroids. UGT1A1 glucuronidates catechol estrogens, UGT1A3 glucuronidates catechol estrogens and the carboxyl group of lithocholic acid (a bile acid that acts as a detergent to solubilize fats for absorption). UGT1A4 has activity toward the hydroxyl group of several steroids and is the only UGT1A isoform to glucuronidate androsterone [92].

The UGT1A6 isoform catalyzes the glucuronidation of several small planar phenols and primary aromatic amines. It is expressed in the liver and in several extrahepatic tissues. UGT1A9 is abundantly expressed in the liver and kidney, and metabolizes endogenous estrogen and thyroid hormones as well as a variety of therapeutic drugs including entacapone (used in combination with other medications for the treatment of Parkinson's disease), edaravone (used to treat stroke and amyotrophic lateral sclerosis), and dapagliflozin (used to treat type 2 diabetes, certain kinds of heart failure, and chronic kidney disease) [93-96].

UGT2B7 is expressed predominantly in the kidney, liver, and gastrointestinal tract [78]. Among its substrates there are commonly prescribed drugs like the antiretroviral agent zidovudine (AZT), the opioid analgesic morphine [97], and the anticancer agent tamoxifen [98]. UGT2B7 plays an important role in the deactivation of steroids, arachidonic acid, prostaglandins, and leukotrienes [99].

2.2.3. Other phase II reactions

Methylation (conjugation with a methyl group, -CH_3) mostly concerns endogenous compounds, nevertheless some drugs may also be methylated by methyltransferases in the lungs, adrenals, liver, skin, pineal gland, or the kidney. Among substrates for the transfer of a methyl group, there are noradrenaline (catalyzed by Phenylethanolamine N-methyltransferase), histamine (by Imidazole N-methyltransferase), catechols (by Catechol O-methyltransferase), or thiols (by S-Methyltransferase). S-adenosylmethionine (SAM), produced from L-methionine and ATP by the enzyme L-methionine adenosyltransferase, is the necessary cofactor for methyl conjugation. Methylation is an exception from other conjugation reactions as it produces less polar products hindering the excretion of drugs [18].

Acetylation (conjugation with an acetyl group, -$COCH_3$) primarily takes place in the liver, more specifically in its Kupffer cells, catalyzed by the enzyme N-acetyltransferase. The reaction introduces an acetyl group on a target compound via the replacement of active hydrogen to form an acetoxy derivative. Acetylation is commonly found for aromatic amines, sulfonamides, and sulfanilamides. In the case of xenobiotics, acetylation of both nitrogen centers within the amino groups of arylamines and arylhydrazines as well as oxygen centers of arylhydroxylamines have been reported [100]. The reaction requires the cofactor acetyl-CoA. The cofactor can either originate from the glycolysis pathway or from the direct interaction between acetate and coenzyme A (a coenzyme, with an important role in the synthesis and oxidation of fatty acids, and the oxidation of pyruvate in the citric acid cycle). Acetyl metabolic products of sulfonamides (acetylsulfonamides) are more hydrophobic than the corresponding parent drugs and are of particular interest due to their renal toxicity [18].

Amino acid conjugation is a special form of N-acylation (conjugation with an acyl group, R-C=O). The most common amino acids involved in amino acid conjugation are glycine, glutamine, ornithine, arginine, and taurine. The enzymes of amino acid conjugation reside in the mitochondria. Amino acid conjugation is the predominant route of metabolism of salicylic acid, with salicyluric acid (its glycine conjugate) accounting for 75% of aspirin's excretion in urine. Also, xenobiotics containing a carboxylic acid group (-COOH) are commonly used as drugs, herbicides, insecticides, and food preservatives, and their metabolism occurs principally via conjugation with either glucuronic acid or an amino acid, most commonly glycine [101].

Glutathione conjugation is an efficient reaction type for the removal of potentially toxic electrophilic compounds (chemical species that form bonds with nucleophiles by accepting an electron pair). The reaction occurs as a nucleophilic attack by reduced glutathione on nonpolar compounds that contain an electrophilic carbon, nitrogen, or sulphur atom . Phase I reactions can generate strong electrophiles which can then react with glutathione (an antioxidant) to generally produce non-toxic conjugates. Their substrates include halogenonitrobenzenes, arene oxides, quinones, and α,β-unsaturated carbonyls [102]. Such reactions are catalyzed by the enzymes glutathione-S-transferases, mostly within the cytosol of liver, kidney, and gut tissues. The glutathione metabolites can either be directly excreted to the bile or the urine, or more often they can be further metabolized [18].

Fatty acid conjugation (mostly including stearic and palmitic acid) has also been observed for some drugs. Condensation reactions (where two molecules are combined to form a single molecule) have been identified for amines and aldehydes and they do not require enzymatic catalysis. However, little is known about these reactions and their involvement in drug metabolism.

3. Drug transporters

Cells must closely monitor and control their intracellular contents to survive and maintain proper function. Transporters present embedded in the plasma membrane of cells mediate the uptake of nutrients form the extracellular space and the elimination of toxic waste from the cytosol. Transporters have long been identified for endogenous compounds such as glucose, amino acids, nucleosides, water-soluble hormones, or neurotransmitters. Xenobiotics are foreign to the given living organisms and hence are not essential for physiological functions. Nevertheless, they can modulate or even damage such activities and living organisms developed processes to eliminate them [103, 104]. Drug elimination orchestrated by DMEs and drug transporters plays an important role in pharmacokinetics, which comprises drug liberation, absorption, distribution, metabolism, and excretion. Drug transporters have been identified to influence the drug disposition and be involved in drug-drug interactions (DDI) of a large number of drugs and drug candidates [105, 106]. Studies in the last two decades have identified various clinically important drug transporters (e.g. P-glycoprotein (P-gp/ABCB1), breast cancer resistance protein (BCRP/ABCG2), members of the multidrug resistance-associated protein subfamily (MRP/ABCC), organic anion transporter (OAT): OAT1, OAT3, organic cation transporter (OCT): OCT2, or the organic anion transporting polypeptide (OATP): OATP1B1 and OATP1B3). The International Transporter Consortium (comprised of scientists from academia, industry, and regulatory agencies around the world) has repeatedly published whitepapers to emphasize the significance of the *in vitro* and *in vivo* evaluation of the clinically relevant effects on drug disposition and DDI of the most important drug transporters [107-112]. Accordingly, drug agencies worldwide (e.g. the European Medicines Agency and the United States Food and Drug Administration) recommend testing for possible substrate or inhibitor status of drug transporters over the course of drug development [113, 114]. Mechanistically, most drug transporters can be classified as either solute carrier (SLC) transporter or ATP-binding cassette (ABC) transporter [115].

3.1. Solute carrier transporters

SLC transporters utilize an electrochemical potential difference, or an ion-gradient generated by primary active transporters for transporting their substrates across biological membranes [116]. The SLC superfamily includes more than 450 transport proteins that are classified in 65 families based on sequence identity. They are physiologically expressed in various key tissues such as the kidney, liver, intestine, and brain. They play crucial roles in maintaining body homeostasis by carrying a large variety of substrates across cellular membranes. Most SLC transporters are influx transporters. Among their substrates there are sugars, amino acids, vitamins, nucleotides, metals, inorganic ions, organic anions, oligopeptides, and drugs.

Members of this superfamily can be found both in the membrane of almost every organelle throughout the cell, and the plasma membrane (approximately 60 % of SLC proteins with known localizations are at the cell surface on the plasma membrane) [117]. Instead of directly using the energy of ATP hydrolysis (like active transporters), members of the SLC superfamily are either passive facilitative transporters (acting as 'gatekeepers', the compounds move down their gradients) or secondary active transporters (coupling the passage of two or more substances, one of which goes down its electrochemical gradient providing the free energy required for the translocation of the other substance(s)) [118, 119]. Secondary active transporters can be symporters (their substrates cross the membrane in the same direction) or antiporters (the substrates cross in opposite directions). SLC proteins also form the system by which many drugs are thought to cross the plasma membrane and gain access to the different media and so effecting drug pharmacokinetics. A large variety of SLC transporters have been proven to be involved in drug transport, including the SLCO subfamily (OATPs), the SLC22A subfamily (OATs, OCTs), the SLC15A subfamily (peptide transporters, PEPTs), and the SLC47A subfamily (multidrug and toxin extrusion, MATEs), also supporting the hypothesis that most drug uptake occurs through transporters rather than by simple diffusion across the lipid bilayer [117]. OCTs generally transport organic cations, OATPs large and fairly hydrophobic organic anions, OATs smaller and more hydrophilic organic anions. PEPTs are responsible for the uptake of dipeptides, tripeptides, and peptide-like drugs, MATEs for the efflux of organic cations [116]. Among these, the SLC22 and SLCO (former SLC21) families are the best understood in terms of pharmacokinetics [120]. Some members of the SLC superfamily are direct target of FDA approved drugs, most of which belong to the SLC5, SLC6, SLC12, and SLC22 families [121].

3.1.1. Organic anion-transporting polypeptides

OATPs are encoded by the SLCO (former SLC21) genes expressing 11 known transporters. OATP substrates mostly include amphipathic organic compounds with higher molecular weights (>350Da). OATPs transport certain endogenous compounds such as bile acids, thyroid hormones, prostaglandins, or steroids, steroid conjugates as well as a number of xenobiotics like statins, angiotensin-converting enzyme inhibitors, angiotensin receptor blockers, antibiotics, antihistaminics, and anticancer drugs [116]. Interestingly, often the uptake by OATPs is coupled to cooperated efflux by ABC transporters like MRPs, ABCG2/BCRP, P-gp/ABCB1. Additionally, many of the OATP substrates are metabolized by phase I and phase II DMEs in the liver. This cooperative regulation is called the 'drug transporter-metabolism interplay (or alliance)' [122, 123]. Such an interplay also exists in other tissues like the kidney and the intestine. To date the energy-coupling transport mechanisms of human OATPs remain poorly understood, it seems to be pH dependent and electroneutral.

3.1.2. Organic cation transporters

OCTs are encoded by the SLC22 genes. They exhibit extensively overlapping substrate selectivity, their substrates include endogenous organic cations (with widely differing molecular structures), cationic drugs, and toxins [124]. OCTs are involved in the intestinal absorption, hepatic uptake, and renal excretion of hydrophilic drugs. An important example of

drug uptake by hepatocytes mediated by OCTs is of the drug used to treat type 2 diabetes, metformin [125]. They also influence the distribution of endogenous compounds such as thiamine, L-carnitine, and some neurotransmitters (like serotonin, dopamine, and norepinephrine) [126].

3.1.3. Organic anion transporters

OATs are a family of multispecific transporters also encoded by the SLC22 gene family (approximately half of the SLC22 genes encode OATs) that expresses 10 OAT proteins [127]. They are of particular interest because of their role in the transport of common drugs (antibiotics, antivirals, diuretics, nonsteroidal anti-inflammatory drugs), toxins (mercury, aristolochic acid), and nutrients (vitamins, flavonoids) [128]. They can be found at most barrier epithelia of the body (e.g. endothelium) demonstrating their roles in the regulation of the transcellular movement of many small organic anionic compounds across the epithelial barriers and between different body fluid compartments [116]. Even though generally they are capable of bidirectional substrate transport, most of them are considered as influx transporters.

3.1.4. Peptide transporters

PEPTs, members of the proton-coupled oligopeptide transporter family, are encoded by the SLC15 genes. Dipeptides, tripeptides, and peptide-like drugs are actively transported by a process which is coupled to the movement of protons down an electrochemical proton gradient as was earlier discovered by brush border membranes vesicles studies [129, 130]. As the transport of PEPT substrates is coupled to proton movement (PEPTs are proton-driven symporters, the H^+ gradient is maintained via the Na^+/H^+ exchanger and/or ATP-driven H^+-pump), they belong to the class of secondary active transporters. Dietary proteins introduced to the human body undergo a series of degradative steps, at the end of which they are broken down into free amino acids or small peptides. To reach the circulation, they are taken up by intestinal epithelia cells. The uptake of protein digestion products to the small intestine is primarily in the form of small peptides, mediated by PEPTs [131]. Additionally, PEPTs expressed in the kidney are responsible for the conservation of protein digestion products, and in the brain, for the homeostasis of neuropeptides. They can transport almost all dipeptides, tripeptides, and peptide-like drugs of very different physicochemical characteristics, molecular weight, charge, and polarity [116]. PEPTs have also been proven to be responsible for the absorption and disposition of a number of pharmacologically important agents including some aminocephalosporins, angiotensin-converting enzyme inhibitors, some β-lactam antibiotics, or antiviral prodrugs [132].

3.1.5. Multidrug and toxin extrusion

MATE transporters are encoded by the SLC47A gene family. They are an exception within the SLC family as they function as efflux transporters [133], MATEs are responsible for the efflux of organic cations from cells. The free energy necessary for their transport is ensured by the oppositely directed proton gradient ($[H^+]_{in}<[H^+]_{out}$, MATEs are electroneutral proton-driven antiporters), also MATEs belong to the class of secondary active transporters [116]. They are predominantly expressed in the proximal tubule epithelial cells in the kidney and the liver

hepatocyte cells, contributing to the excretion of cationic endogenous substances and xenobiotics. Clinically used drugs have also been identified as substrates to MATEs, e.g. metformin (the main first-line medication for the treatment of type 2 diabetes [134]) and cimetidine (a histamine H_2 receptor antagonist that inhibits stomach acid production) [135].

3.2. ATP-binding cassette transporters

The ATP-binding cassette (ABC) transporter superfamily genes represent the largest family of transmembrane proteins. In addition to SLC transporters, they are the other major drug transporter type. ABC transporters bind ATP and harvest the energy of ATP hydrolysis in order to selectively translocate a variety of substrates (including sugars, amino acids, metal ions, peptides, proteins, hydrophobic compounds, and their metabolites) across membranes. ABC transporters are integral membrane proteins and in humans they can be predominantly found in the plasma membrane, but they are also present in the intracellular membranes of the ER, peroxisome, and mitochondria.

Members of the ABC transporter superfamily are classified based on the sequence and organization of their ATP-binding domains, also referred to as nucleotide-binding domains (NBDs) [136]. Over 40 ABC transporters have been identified in humans which are divided into 7 subfamilies (ABCA to ABCG). At least 11 members (including ABCB1/P-gp, ABCCs/MRPs, ABCG2/BCRP) have been proven to be involved in the development of multidrug resistance (MDR). Such transporters are physiologically expressed in various tissues such as the liver, intestine, kidney, and brain where they influence the absorption, distribution, and excretion of drugs. ABC transporters are also involved in various cellular processes such as the maintenance of osmotic homeostasis, antigen processing, cell division, immunity, cholesterol, and lipid trafficking [116]. Functional ABC transporters (functioning either as monomers or dimers) generally contain two NBDs and two transmembrane domains (TMDs). The NBDs contain characteristic motifs necessary for ATP-binding and hydrolysis, such as the Walker A, Walker B, and the signature motif ('LSSGQ') and other conserved regions like the A-loop, H-loop, D-loop, and the Q-loop. The TMDs account for the substrate specificity of the different transporters. The NBD motions resulting from the binding and hydrolysis of the ATPs are coupled to TMD motions via the coupling helices, the structure of which are conserved among the different transporters.

ABC transporters are mostly unidirectional under physiological conditions and in eukaryotes they primarily function as exporters (translocating substrates from the cytoplasm to the extracellular space, or to the ER, mitochondria, or peroxisome) whereas in bacteria they mostly import essential compounds (like sugars, vitamins, metal ions) into the cell [136]. The TMDs alternate between inward- and outward-facing conformations, that way ensuring the unidirectional active transport of their substrates. The exact mechanisms of ABC efflux transporter-mediated substrate translocation are not fully understood, several different models have been proposed. Binding of a substrate to the TMDs; binding of two ATPs and coordinating Mg^{2+} ions to the NBDs; dimerization of the NBDs; conformational transition between the inward-facing and outward-facing TMD configurations; ATP-hydrolysis;

phosphate, ADP, and transport substrate release; and NBD dissociation are all elementary steps of the transport cycle, however, the exact details and order of these steps can be transporter dependent and remain partially unclear [137]. Among the ABC protein superfamily, ABCB1/P-gp, ABCG2/BCRP and members of the ABCC subfamily are known to be important drug and drug metabolite transporters [138].

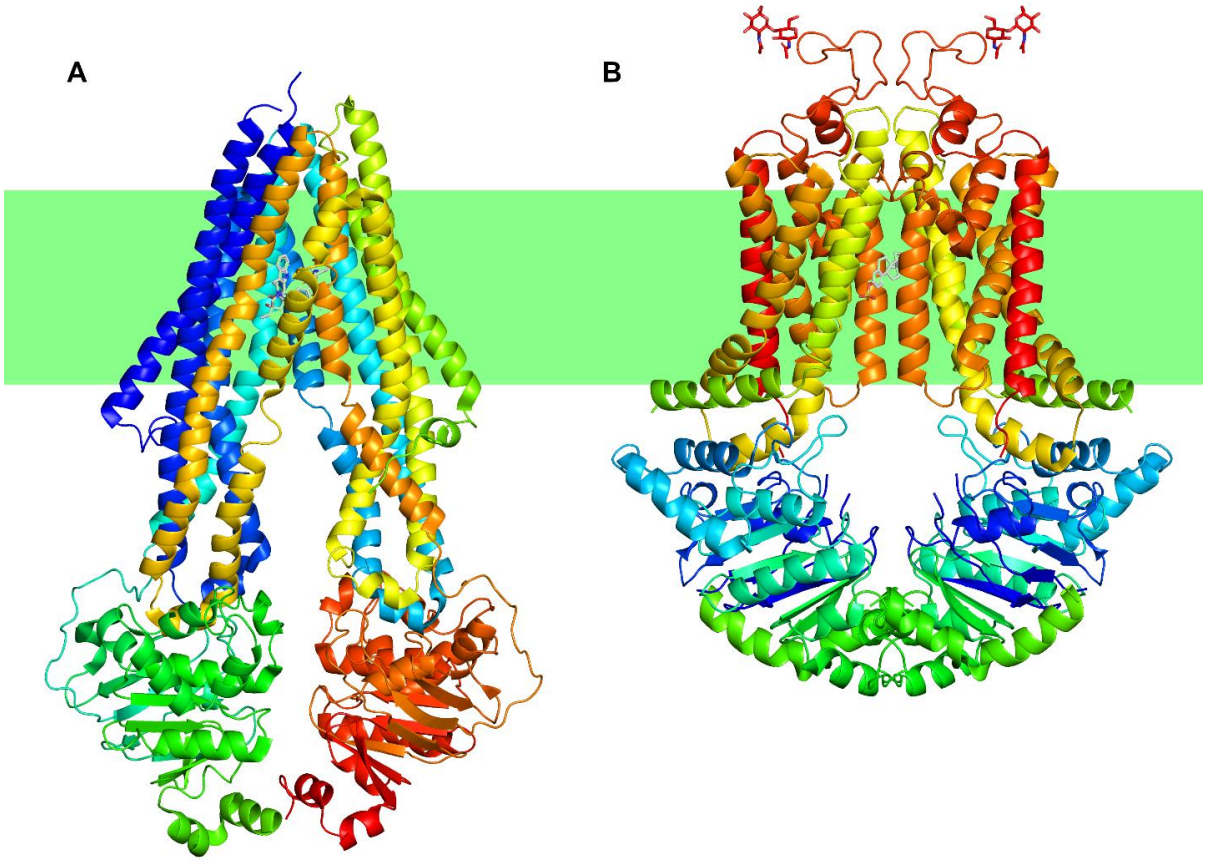

Figure A.5: Cryo-EM structures of ABC transporters in the substrate-bound inward-facing state. A) ABCB1/P-gp (PDB 7A69) and B) ABCG2/BCRP (PDB 6HCO).

3.2.1. ABCA transporters

There are 12 known functional transporters belonging to the ABCA subfamily in humans, all ABCA proteins are full transporters (they function as monomers containing 2 NBDs and 2 TMDs). Many of them have been identified to be involved in the transport of a variety of physiologic lipid or lipid-related compounds [139], they play an important role in cholesterol homeostasis. ABCA members show a broad tissue specificity. Multiple members of the ABCA subfamily have been linked to monogenic recessive disorders recently, ABCA1 (Tangier disease, significantly reduced levels of high-density lipoprotein (HDL) in the blood), ABCA3 (neonatal surfactant deficiency, a lung disorder), ABCA4 (Stargardt disease, an eye disease that happens when fatty material builds up on the macula), ABCA12 (harlequin ichthyosis, a severe disorder that affects the skin), and ABCA5 (congenital generalized hypertrichosis terminalis) [140]. Some members are also involved in more complex diseases like atherosclerosis (ABCA1), pediatric interstitial lung diseases (ABCA3), age-related macular degeneration (ABCA4), and Alzheimer disease (ABCA1, ABCA2, ABCA7, and ABCA5) [116].

3.2.2. ABCB transporters

The ABCB subfamily contains both half transporters (containing 1 TMD and 1 NBD, functioning as homo- or hetero-dimers) and full transporters. There are 11 known functional transporters belonging to the ABCB subfamily. ABCB1, also known as P-Glycoprotein (P-gp) or multidrug resistance protein (MDR1), is the first human ABC transporter cloned and characterized through its ability to promote a multidrug phenotype to tumor cells [136]. ABCB1 has a broad substrate spectrum and is physiologically widely expressed, it mediates drug transport in various tissues like the intestine, the liver, and the brain (especially at the blood-brain barrier). ABCB1 exports various hydrophobic compounds including therapeutic drugs, alkaloids, flavonoids, and other hydrophobic natural toxic compounds. Among the drugs that are substrates to ABCB1 there are chemotherapeutic drugs, HIV protease inhibitors, immunosuppressive agents, antiarrhythmics, calcium channel blockers, analgesics, antihistamines, antibiotics, natural products, fluorescent dyes, and pesticides [116].

3.2.3. ABCC transporters

The ABCC subfamily contains 12 known transporters, all of which are full transporters. They exhibit a divers functional spectrum including ion transport, cell surface receptor, and toxin secretion activities [136]. With one exception (the pseudogene ABCC13) ABCC genes code transporters that are termed multidrug resistance-associated proteins (MRPs), they play an important role in MDR development [141]. The ABCC transporters are divided into short or long structure subgroups. The long ABCC family members, ABCC1, 2, 3, 6, and 10 contain an additional N-terminal TMD0 domain followed by an L0 linker segment connected to the transmembrane region [142]. Several ABCC transporters are linked to genetic diseases such as ABCC2 to Dubin-Johnson syndrome and ABCC7 to cystic fibrosis [116]. Among the ABCC members, ABCC1 (MRP1) is considered to be of the highest clinical importance with respect to drug resistance in cancer, a major obstacle to successful chemotherapy. ABCC1 typically exports structurally diverse amphipathic organic anions, most of which are conjugated with, glutathione, glucuronide, or sulfate (generally metabolites of phase II reactions). ABCC2 is expressed at major physiological barriers, such as the canalicular membrane of liver cells, and always localizes in the apical membranes. It transports a variety of amphiphilic anions, and it also displays a preference for phase II conjugates [143]. ABCC3 is a broad specificity organic anion transporter, it is involved in the efflux of organic anions including monovalent bile acids. Its drug substrate spectrum considerably overlaps with other transporters including ABCC1 and ABCC2. ABCC4 is physiologically widely expressed in most tissues including the lung, kidney, bladder, gallbladder, small intestine, and tonsil, and most abundantly in the prostate. Its substrates include antiviral, antibiotic, cardiovascular, and endogenous molecules [116]. ABCC4 and ABCC5 are both organic anion transporters, they have the outstanding ability to transport nucleotides and nucleotide analogs [144]. ABCC6 is primarily localized in the basolateral plasma membrane of hepatocytes in the liver and in proximal tubules of the kidneys [145], mutation in the ABCC6 gene is responsible for pseudoxanthoma elasticum (a disease that causes mineralization of elastic fibers) [146]. ABCC7 functions as an ATP-gated chloride channel and is not believed to mediate active transport, however it shares the conserved domain architecture

characteristic of the ABC superfamily [147]. ABCC11 is a cyclic nucleotide efflux pump that is able to confer resistance to nucleoside-based agents, it is expressed in axons of neurons, in the human central and peripheral nervous systems and mediates the efflux of neuromodulatory steroids. The functional characteristics of ABCC12 remain unclear [116, 148].

3.2.4. ABCD transporters

The ABCD subfamily contains four genes that encode four half transporters. All members are located in peroxisomes. Members of the ABCD subfamily have distinct but overlapping substrate specificities for different acyl-CoA esters. They play a role in the regulation of very long chain fatty acid transport [116, 136].

3.2.5. ABCE and ABCF transporters

The ABCE and ABCF subfamilies contain genes that have ATP-binding domains that are clearly derived from ABC transporters but have no TMDs and are not known to be involved in any membrane transport functions [136].

3.2.6. ABCG transporters

There are 6 half transporters belonging to the ABCG subfamily. They are all 'reversed' compared to other ABC transporters, they have their NBDs at their N-terminus while the TMDs at the C-terminus [136]. They either form homodimers (ABCG1, ABCG2, and ABCG4) or an obligate heterodimer (ABCG5 and ABCG8). ABCG1, ABCG4, ABCG5, and ABCG8 are involved in the ATP-dependent translocation of steroids and lipids, playing a significant role in the efflux transport of cholesterol. In addition to ABCB1/P-gp and ABCC1/MRP1, ABCG2/BCRP has been implicated to be a major efflux transporter responsible for multidrug resistance in cancer cells. ABCG2, also known as breast cancer resistance protein (BCRP) and mitoxantrone resistance-associated protein (MXR), was originally discovered in a multidrug-resistant breast cancer cell line where it was found to confer resistance to chemotherapeutic agents such as mitoxantrone and topotecan. Like ABCB1 and ABCC1, ABCG2 possesses a very broad substrate and inhibitor specificity that is different from, but substantially overlaps with that of ABCB1 or ABCC1 [149]. ABCG2 is a key player in preventing the absorption of toxic compounds from the gut; increasing their hepatobiliary clearance, and it also plays an essential protective role at different tissue barriers like the maternal-fetus barrier, the blood-brain barrier (BBB), and the blood-testis barrier [150]. High ABCG2 expression has been found especially in the placenta and intestine, but also in the brain endothelium, prostate, testes, ovaries, liver, adrenal gland, uterus, and central nervous system (CNS) [151, 152]. Among its substrates there are physiological compounds like estrone-3-sulfate, 17β-estradiol, 17-β-d-glucuronide, or uric acid. Drugs identified as ABCG2 substrates include chemotherapeutics (e.g. mitoxantrone, camptothecin derivates, flavopiridol, and methotrexate), some fluorescent compounds (e.g. the fluorescent dye Hoechst 33342), conjugated or unconjugated organic anions. Further drug substrates include prazosin (an antihypertensive), glyburide (an antidiabetic medication), cimetidine (a histamine H_2 receptor antagonist that inhibits stomach acid production), sulfasalazine (used to treat inflammatory bowel disease), and rosuvastatin (used to prevent cardiovascular disease) [149]. ABCG2 also transports many phase II metabolites such as sulfate, glutathione, or

glucuronide conjugates. In general, sulfated conjugates seem to be better ABCG2 substrates than glutathione and glucuronide conjugates [116]. A variety of chemical toxicants have also been shown to be transported by ABCG2 [153].

4. Drug-drug interactions

Pharmacokinetic drug-drug interactions (DDIs) occur when a drug alters the disposition (absorption, distribution, elimination) of a co-administered agent. Such interactions may provoke the increase or the decrease of plasma drug concentrations. The mechanism of DDIs involve contributors to drug metabolism, namely DMEs and drug transporters, and also orphan nuclear receptors that regulate the expression of enzymes and transporters at the transcriptional level. Increase in the intracellular drug concentration can be the result of the inhibition of DMEs and drug transporters (a quick effect (24-28h) obstructing the metabolism and excretion of the drug), whereas the decrease of drug concentrations can be the result of the activation of orphan nuclear receptors by inducers that lead to the increase in the expression of DMEs and drug transporters (a more slowly effect (7-10 days) leading to faster deactivation and elimination of the drug) [154]. Such interactions should be avoided in case the drug combination decreases the clinical activity or increases the probability of adverse drug effects.

Induction or inhibition of CYP enzymes is a major mechanism that underlies DDI. CYP enzymes can be transcriptionally activated through receptor-dependent mechanisms and CYP inhibition is a principal mechanism for metabolism-based DDI [155]. Induction of CYP1 genes occurs by the aryl hydrocarbon receptor (AhR) mechanism, additionally, three distinct orphan receptors that belong to the nuclear receptor/steroid receptor superfamily have also been identified to induce CYP1 transcription. AhR transactivates human CYP1A1, CYP1A2, and CYP1B1, as well as some phase II metabolizing enzymes. After its ligand activation, AhR translocates from the cytosol to the nucleus, where it forms a complex that docks onto genes containing the xenobiotic response element (XRE) and transactivates them (e.g. CYP1A1) [155, 156]. AhR ligands include polycyclic aromatic hydrocarbons (PAHs), halogenated aromatic hydrocarbons (HAHs), and some clinical drugs, caffeine, and eicosanoids [157]. CYP enzyme inhibition usually occurs as competition with another drug for the same catalytic binding site. For example, desipramine (a tricyclic antidepressant) metabolism by CYP2D6 is strongly inhibited by binding of fluoxetine (an antidepressant of the selective serotonin reuptake inhibitor class) to the same isoenzyme [158]. Inhibition, however, may also be non-competitive when the compound binds to a site other than the catalytic active site (allosteric inhibition). Enzyme inhibition can be either reversible or irreversible. Inhibition of CYPs can lead to the toxicity (by impaired clearance of drugs) or lack of efficiency (in case of prodrugs) of a given drug [159]. As an example, tyrosine kinase inhibitors (TKIs), a class of anticancer agents, whose chemical structures as well as their metabolism and general pharmacokinetic characteristics are rather variable, are often involved in DDIs as many of them are substrates and/or act as inhibitors of different CYP isoenzymes [160]. Another well-known example is ketoconazole (an antifungal medicine) inhibition of terfenadine metabolism (an antihistamine, withdrawn from

markets worldwide due to its clinical consequences like cardiac arrhythmias) catalyzed by CYP3A, leading to a 35-fold increase in drug exposure measured by area under the plasma concentration versus time curve (AUC) [161].

There are more examples of inhibitors than enhancers of sulfonation. Modulators of SULT enzymes include natural products ingested as part of the human diet as well as environmental chemicals and drugs. Most inhibitors of SULTs are hydroxylated compounds with the potential to be substrates, however the mechanism of inhibition is not always competitive. Example of potent SULT inhibitors include flavonoids and also several therapeutic drugs, mainly non-steroidal anti-inflammatory drugs (NSAIDs) that have been shown to inhibit sulfotransferases and therefore they contribute to DDI [162].

Pharmacokinetic DDI studies, that have been conducted for drugs cleared by glucuronidation, have shown that changes in drug exposure are typically less than two-fold (AUC of plasma concentration over time) compared to what is observed in the absence of a given UGT inhibitor [22, 163]. An example is valproate (used to treat epilepsy and bipolar disorder and prevent migraine headaches) coadministration which increases lorazepam (used to treat anxiety disorders, trouble sleeping, severe agitation, active seizures, alcohol withdrawal, and chemotherapy-induced nausea and vomiting) plasma concentration by 20 % [164]. Evidence of toxicity as a result of inhibition of UGTs is rare, an exception is lamotrigine (used to treat epilepsy and stabilize mood in bipolar disorder) which if coadministrated with valproic acid, increases the risk of rash [22].

The other important mechanism underlying DDIs is the induction or inhibition of drug transporters that mediate the cellular uptake and efflux of xenobiotics, especially in the small intestine, the liver, and the kidney, where transporter-mediated DDI can significantly alter the pharmacokinetics and clinical effects of drugs [165]. Such DDI occur when translocation of a drug by one or more of the drug transporters is influenced by a second drug via inhibition or induction. The intestine has high physiological expression of ABCB1 and ABCG2, which limits the bioavailability of orally administrated substrates. Coadministration of drugs with ABCB1 or ABCG2 inhibitors results in higher bioavailability of substrate drugs, resulting in an enhanced activity or toxicity. Opposed to this, coadministration of ABCB1 or ABCG2 inducers reduces the bioavailability of substrate drugs leading to therapeutic failure. Hepatocytes of the liver highly express uptake transporters (e.g. OATP1B1, OATP1B3, OATP2B1, and OCT1) that mediate the uptake of substrate drugs into hepatocytes for consequent metabolism by DMEs as part of drug elimination. At the same time, as a following step of drug elimination, efflux transporters (e.g. ABCB1, ABCG, ABCC2, and SLC47A1) at the canalicular membrane of hepatocytes excrete drugs and their metabolites into the bile. In the proximal tubular cells of the kidney, uptake of cationic drugs from blood is mediated by SLC22A2 located in the basolateral membrane while their subsequent efflux into urine is mediated by SLC47A1 and SLC47A2 located in the luminal membrane. The inhibition of such drug transporters results in reduced renal clearance of substrate drugs [116].

Digoxin (used to treat various heart conditions) is one of the most extensively investigated compounds involved in ABCB1-mediated DDIs. Its metabolism in humans is negligible, hence alteration in ABCB1 activity (due to expressional or functional modulation of the transporter) has a direct impact on digoxin pharmacokinetics, ABCB1 mediates its renal and biliary secretion [166]. For example, orally administered pretreatment with the antibiotic rifampin before digoxin results in considerably reduced digoxin plasma concentrations. When the same pretreatment was applied intravenously, the activity of digoxin AUC was reduced much less. These results demonstrated that induced ABCB1 activity in the small intestine (in the duodenum, 3.5-fold increase in the ABCB1 expression) increases the presystemic digoxin extraction [167]. Interestingly, depending on its concentration, rifampin may also act as an inhibitor of ABCB1. Shortly after drug administration, high local rifampin concentrations in the gut could also inhibit ABCB1-mediated digoxin uptake, increasing its oral bioavailability [165]. An example for ABCG2-mediated DDIs is the coadministration of the ABCG2 inhibitor GF120918 molecule with topotecan (a chemotherapeutic agent which is substrate to ABCG2), resulting in a considerably increased oral bioavailability of topotecan [168].

An example of uptake-transporter-involved DDIs are statins (which are lipid-lowering 3-hydroxy-3-methylglutaryl-coenzyme A reductase inhibitors, i.e. cholesterol-lowering drugs). Hepatocellular uptake eliminates statins from portal veinous blood and systemic circulation, thereby decreasing their plasma concentrations and minimizing the risk of myotoxic damage (the risk of myopathy occurrence increases with statin plasma concentrations). Also, hepatocellular statin uptake is a prerequisite for their subsequent pharmaceutical activity which is the inhibition of 3-hydroxy-3-methylglutaryl-coenzyme A reductase in the hepatocytes. Statin uptake inhibition increases the risk of myopathy and decreases their therapeutic efficacy. For example, gemfibrozil (also used to treat abnormal blood lipid levels) and its major metabolite (gemfibrozil-1-O-b-glucuronide) are both inhibitors of multiple members of the SLCO1B subfamily. Concomitant oral administration of gemfibrozil e.g. with pravastatin or rosuvastatin resulted in an increase of their AUC [169, 170].

5. Multidrug resistance

Cancer is a leading cause of death worldwide, accounting for nearly 10 million deaths in 2020, i.e. nearly one in six deaths [171]. Even though there are several different methods of cancer treatments, including radiation therapy, surgery, immunotherapy, endocrine therapy, and gene therapy, chemotherapy still remains the most common method of cancer healing. Resistance of tumor cells to chemotherapeutic agents is a great challenge, multidrug resistance (MDR) is responsible for over 90 % of deaths in cancer patients receiving traditional chemotherapeutics or novel targeted drugs. The mechanisms of MDR include elevated drug metabolism by DMEs, enhanced drug efflux by drug transporters, increased DNA repair capacity, and genetic events such as gene mutations, amplifications, and epigenetic alterations [172].

Several studies have proven that exposure to anticancer agents can induce the expression of gene products protecting cells, including DMEs that catalyze phase I and phase II drug metabolism. CYP isoenzymes CYP1A6, CYP1A2, CYP1B1, CYP2C9, CYP2B6, CYP2C19,

CYP3A4/5, and CYP2D6 are essential for phase I drug metabolism and detoxification [172]. The overexpression of CYP1B1, which contributes to the metabolism of mitoxantrone, flutamide, docetaxel, and paclitaxel [173], of CYP2A6, which is involved in the metabolism of several anticancer agents, including ifosfamide, cyclophosphamide, aflatoxin, and fluorouracil, as well as of other CYP isoforms has been reported in different tumor tissues [174]. Enhanced expression of phase II DMEs including UGTs and SULTs in cancer cells may also contribute to their MDR phenotype by inactivating substrate drugs [173]. The search for selective phase II DME inhibitors has become part of the strategy in overcoming cancer MDR as demonstrated e.g. for the isoenzyme UGT1A4 [175].

The activity of drug transporters, especially the efflux ABC transporters, can significantly decrease drug bioavailability, intracellular drug concentrations, and drug transition through the blood-brain barrier (BBB). Chemotherapeutic drug penetration can be drastically reduced to the specific sites, e.g. in brain tumor treatment where in the case of ABCB1 or ABCG2 overexpression (the physiological function of which includes brain protection from potentially damaging compounds), anticancer agents are generally incapable of passing through the BBB. In most cases, higher drug concentrations are necessary to overcome the restriction posed by the efflux transporters, however, higher drug concentrations can easily lead to systemic toxicity. ABCB1, ABCG2, and members of the ABCC subfamily can eliminate a structurally and functionally wide variety of anticancer agents from cells to the extracellular space. Such compounds include e.g. epipodophyllotoxins, anthracyclines, vinca alkaloids, bisantrene, colchicine, taxanes, imatinib, saquinavir, camptothecins, thiopurines, actinomycin D, methotrexate, and mitoxantrone [172]. Notably, overexpression of ABCB1 has been observed in approximately half of all human cancers, existing both as a response to or independently from treatment with anticancer agents.

6. Pharmacogenetics and pharmacogenomics

The field of pharmacogenetics and pharmacogenomics researches the relationship between genotype (such as polymorphism and genetic mutations), gene expression profiles (the level of physiological expression of the different genes), and phenotype in terms of inter-individual variability in drug response or toxicity. Pharmacogenetics is identified as the study of variability in drug response due to heredity, in relation to genes determining drug metabolism, whereas pharmacogenomics is a broader term encompassing all genes in the genome that may determine drug response involving complex multigene patterns within the genome. The distinction however, is arbitrary and the two terms can be used interchangeably [176]. Pharmacogenetics and pharmacogenomics are key factors to the success of personalized medicine, i.e. to find the best treatment based on the individual genetic and biological profile of a patient. Inter-individual variability in drug disposition is a major cause of drug treatment inefficiency and adverse effects. Genetic polymorphisms effecting actors involved in the mediation of drug pharmacokinetics such as DMEs and drug transporters have a great influence on drug disposition. Genetic polymorphisms are alterations in the individual genomes that remain constant throughout a person’s lifetime. Genetic mutations on the other hand are

acquired changes in the genome that occur only in certain cells. Such genetic variations effect drug metabolism and transport, and collectively can contribute to the variability in drug pharmacokinetic and pharmacodynamic processes [177].

Many hepatically cleared drugs are first metabolized by the CYP enzymes, mainly by the members of the CYP1, CYP2, and CYP3 families. Genes encoding CYP enzymes are highly variable with significantly different allele distributions between populations. Evidence of the clinical impact especially of the isoenzymes CYP2C9, CYP2C19, and CYP2D6 on substrate drug metabolism has been reported. Functional changes of the different variant alleles can range from no function to increased functions, poor metabolizers to ultrarapid metabolizers, the major variability in CYP activity is due to single nucleotide polymorphisms (SNPs) in the CYP gene locus [178]. CYPC2C9 is one of the most abundant CYP in human liver that metabolizes more than 100 clinical drugs (antibiotics, anticancer agents, anticoagulants), 85 allelic CYP2C9 variants have been identified to date (2022 August, https://www.pharmvar.org/gene/CYP2C9) [29]. For example, an R144C SNP (CYP2C9*2, 430C>T) is found in Caucasians at a frequency of 10-15 % resulting in a decreased enzymatic activity [179], whereas a complete loss of function with no enzymatic activity can be observed for the allele CYP2C9*6, a 273frameshift, found only in African Americans [180]. To date, 39 allelic variants have been identified for the isoenzyme CYP2C19. For example, CYP2C19*2 (681G>A, splice defect) is its most common loss-of-function allele, with an allele frequency of 29-35% in Asians. CYP2D6 is the most polymorphic CYP gene, to date, 163 allelic variants have been identified. An example for increased enzymatic activity can be found as a result of gene duplication (two or more copies of the same gene sequence, e.g. CYP2D6*1xN, an ultrarapid metabolizer), whereas a splice defect (CYP2D6*4, 1847G>A) is the most frequent null allele in Caucasians at a frequency of 20-25 % in the population [181].

Genetic evolution has created a broad family of cytosolic SULTs that modify a large variety of xenobiotic substrates by sulfoconjugation, members of the SULT superfamily have extremely broad substrate specificities and as a consequence, corresponding genetic alterations may affect drug metabolic enzyme functions influencing individual responses to drug treatments, in terms of therapeutic efficiency or adverse effects. An example of SNP which occurs at frequencies of 31 to 36 % (per chromosome) in Caucasian people is G638A (SULT1A1*2), which translates to an R213H conversation (spatially located distant from the cofactor- and substrate-binding site) in SULT1A1, drastically lowering its enzymatic activity and thermal stability [182]. Another frequent SULT1A1 SNP results in the amino acid substitution M223V (SULT1A1*3) and is well represented in African Americans also reducing its enzymatic activity, however, to a smaller extent than in the case of SULT1A1*2 [60].

A considerable number of prevalent, functional polymorphisms have been identified in several UGT genes as determinants of cancer risk or response to chemotherapy [183]. In terms of drug metabolism, the clinically most important hepatic UGTs are UGT1A1, UGT1A3, UGT1A4, UGT1A6, UGT1A9, and UGT2B7 [77]. Rare mutations in the UGT1A1 gene (point mutations, deletions, and insertions) have been described associated with two severe forms of the

unconjugated hyperbilirubinemia syndromes (Crigler–Najjar type I and II disorders), however, only a few of these were found in the general population at high frequencies (>1 %) to be classified as polymorphisms. Wild type UGT1A1 has 6 TA repeats in its regulatory promoter TATA box and UGT1A1 promoter activity progressively decreases with the number of TA dinucleotide repeats (decrease in the rate of transcription initiation of the UGT1A1 gene). The most common variant allele (0-3%, 2-3%, and 16-19% in the Asian, Caucasian, and African populations, respectively) that has 7 TA repeats (UGT1A1*28) is a low activity allele and is associated with the mild form of the inherited unconjugated hyperbilirubinemia syndrome (Gilbert's syndrome, found in approximately 6–12% of the population) [184, 185]. Studies showed that mutations in the UGT1A1 gene, especially homozygosity for the UGT1A1*28 allele, are responsible for Gilbert's syndrome. Patients affected by Gilbert's syndrome also display lower glucuronidation rates for a number of therapeutic drugs [186].

Genetic alterations affecting drug transporters also have a great influence on drug disposition and therefore also on the inter-individual variability in drug response or toxicity. Re-sequencing of various human ABC transporter and SLC transporter genes has revealed a number of naturally-occurring allelic variants, many of which appear to affect the functional activity of the encoded protein *in vivo*. The most common allelic variant of ABCB1 is the wild-type transporter (ABCB1*1). Another common allele is ABCB1*2 which contains three SNPs simultaneously, C1236T in exon 12, G2677T in exon 21, and C3435T in exon 26. C1236T and C3435T are synonymous/silent polymorphisms (i.e., not causing a change in the amino acid), yet the polymorphisms were associated with changes in pharmacokinetics of a number of drugs, most probably due to the effects on folding by the usage of a rare codon. Opposed to this, the G2677T polymorphism causes the amino acid conversion A893S. In vitro experiments revealed higher vincristine (a chemotherapy drug belonging to vinca alkaloids) transport rates for the A893T variant than in the case of the wild-type transporter [187, 188]. The SNP C3435T together with the formerly mentioned G2677T and C1236T is also frequent, however its exact impact on drug pharmacokinetics remains unclear.

Today, 36 genetic variants in the human ABCG2 are associated with a drug phenotype in literature and are therefore annotated by PharmGKB [189], its most extensively studied variants are C421A in exon 5 leading to the amino acid exchange Q141K (located in the Walker A motif) as well as G34A in exon 2, leading to the amino acid conversion V12M. Several *in vitro* and *in vivo* studies demonstrated that Q141K results in a reduced global ACBG2 protein expression while other studies also reported reduced ATPase and corresponding reduced transport activity [187, 190, 191]. The V12M amino acid exchange is located in a flexible tail close to the N-terminus. Currently available experimental structures lack the first 30 residues, therefore the structural impact of such a substitution on protein folding is not completely understood. However, cell experiments have indicated that it could potentially disrupt the localization and insertion of ABCG2 into the plasma membrane [192]. Two additional naturally occurring variants, T434M and S476P, which are located in the TM helices, abolish the transport activity of ABCG2 [193, 194].

7. *In silico* modeling of drug metabolism and transport

The outcome of administrated drugs is highly influenced by DMEs and drug transporters, they are of primary interest over the course of drug discovery and development. A wide variety of experimental technologies have been developed to provide insights into the fate of drugs and drug candidates, however, due to the high costs of such experimental studies, an increasing number of computational approaches have been involved in the prediction of the metabolic outcome of drug candidates. Screening campaigns of large numbers of chemical compounds are widely used in order to identify a small number of promising therapeutic candidates with good ADME-Tox profiles. *In silico* studies have both addressed the mechanisms of DMEs and drug transporters and their interactions with small molecules. Drug interactions with DMEs and drug transporters are of great clinical interest, modulation of their activities may lead to inefficacy or toxicity. At the same time, DMEs and drug transporters also determine the ADME-Tox properties and bioavailability of drug candidates. The prediction of such interactions can help to reduce drug candidate failure at an early drug development stage. Various approaches have been applied to achieve such predictions *in silico*, the use of computational tools can lower the need for animal testing, and reduces the costs associated with drug development. Structure-based approaches as molecular dynamics and molecular docking simulations have become powerful tools to address conformational flexibility of proteins and their interactions with small compounds, meanwhile ligand-based approaches such as structure-activity relationship (SAR) models have also been created to identify possible substrates and inhibitors of DMEs and drug transporters. Statistical models predicting interactions of ligands with proteins can integrate information on the ligands and/or the protein and their interactions.

7.1. Structure-based methods

7.1.1. Functional dynamics of DMEs and drug transporters, interactions with ligands

The increasing number of experimental 3D structures of the different CYP isoenzymes (mostly determined through X-ray crystallography) have largely contributed to the understanding of the structural basis of ligand binding. The resolved CYP structures have revealed a considerable flexibility, in particular around their active site. Ligand-induced conformational changes have also been described [195, 196]. To further examine the structural flexibility of the CYP enzymes, molecular dynamics (MD) simulations have been employed both in the presence and in the absence of bound ligands. MD is a computational approach which is based on the numerical integration of Newton's equations of motion (computing the interaction forces acting on each atom and then propagating the velocities and positions of the atoms at each simulation step) and can be used to generate successive configurations of a given system as a function of time, starting from an initial conformation which is generally an experimental structure or a (homology) model [197]. MD simulations can be employed with various purposes. Conformational exploration can be useful before molecular docking simulations to consider flexibility of the protein, or the effects of either point mutations or the presence of a bound ligand on the dynamical behavior can be elucidated. Additionally, the stability of different docking positions can be further evaluated with the help of MD simulations. Answers related

to the binding free energy, interactions between the enzyme and its substrates or inhibitors, active-site residues, and some key distances such as e.g. in the case of CYP enzymes, the distance between the heme iron and the substrate reacting group, can be addressed [198]. MD simulations have primarily focused on the isoenzymes CYP2C9, CYP2C19, CYP2D6, CYP3A4, but have also included e.g. CYP1A1, CYP1B1, CYP1A2, CYP2A6, CYP2B1, or CYP2B6. The effects of SNPs have been broadly tested in the different isoforms.

Banu et al. have demonstrated the effect of two alleles of CYP2C9 (R144C and I359L) on its catalytic activity using MD simulations, namely the reduction in size of the substrate entry access channel which leads to reduced metabolic activity of the enzyme [199]. Another study by Sano et al. has also investigated the mechanism of the decreased catalytic activity of CYP2C9 polymorphic variants, they identified alterations in the fluctuations (flexibility) of structural regions e.g. important for substrate binding which may destabilize the enzyme-substrate complex and reduces the enzymatic activity [200].

Substitutions in some residues distant from the catalytic site can also influence enzymatic activity as was demonstrated in CYP2C9 polymorphic variants by Lertkiatmongkol et al. They found that mutations that are located outside the binding pocket can induce pocket conformational changes following allosteric regulation [201] and such allosteric effects were also detected in other isoforms like CYP1A2 [202] and CYP2B4 [203]. Zhang et al. studied the effects of a peripherical mutation in CYP1A2 (F186L) with the help of MD simulations and found that in response to the mutation, the overall structural fold was maintained, however, the flexibility of the enzyme increased and the substrate access channel closed [202]. The reduced enzymatic activity of another allele of CYP2C9 (L90P) was investigated using MD simulations by Zhou et al., a rearrangement in the backbone configuration of residues at the substrate entry was identified as the dominant reason for the catalytic activity reduction of this specific CYP2C9 allele [204]. Louet et al. also addressed the reduced catalytic activity of an allelic variant of CYP2C9 (CYP2C9*30, A477T). They performed MD simulations on both the wild type and the allelic variant in the presence of heme and with or without bound substrate. They observed increased rigidity of the key substrate recognition sites as well as decreased channel access of the substrates [32].

Isvorana et al. investigated the role of amino acid substitutions in the CYP2C subfamily, and used different *in silico* tools to reveal the molecular mechanisms related to CYP2C polymorphisms. The functional effects of missense mutations (with dramatically altered drug metabolism) were analyzed using different online tools predicting the potential consequences of amino acid substitutions on the protein structure and/or function. They identified steric clashes, local rigidity alterations, and hydrogen-bonding/salt-bridge network perturbations among other effects as consequences of the substitutions that can lead to altered drug metabolism [27].

For the isoenzyme CYP2C19, Cui et al. have investigated the mechanism of mutants of experimental interest (A161P and P227L) and the corresponding reduced enzymatic activities, and found that the overall protein topologies were maintained, yet the effect of conformational

changes got manifested at more distant regions through the propagation of favorable interactions [205]. MD simulations were also used to study the effects of different genetic polymorphisms of the isoenzyme CYP2D6 by Fukuyoshi et al., they have also identified amino acid mutations having an effect on distant structural regions which are part of the active site access channel, accounting for the reduced enzymatic activity [206]. For the isoform CYP2A6, Yadav et al. performed MD simulations on four allelic variants and have also observed that non-synonymous SNPs in CYP2A6 did not induce global changes in the physiochemical properties of the enzyme, however, they caused local-trivial changes that are very crucial for the metabolic activity [207]. Similarly, the effect of point mutations on the enzyme dynamics of other isoforms has also been addressed, MD simulations to determine effects of the amino acid mutations have been performed for CYP1A2, CYP17A1, CYP19A1, CYP2A6, CYP2B4, and CYP2B6 [202, 208-212].

In addition to the effect of point mutations, MDs have also been employed to identify differences between the mechanisms of various isoenzymes. Skopalik et al. have investigated the differences in flexibility of the isoenzymes CYP3A4, CYP2C9, and CYP2A6, and have revealed that increased flexibility correlates with higher substrate promiscuity. Among the three isoforms, CYP2A6 has the narrowest substrate range and is the most rigid whereas CYP3A4 is the most promiscuous known CYP and is the most flexible [213]. Mustafa et al. have demonstrated that the small sequence and structural differences between the two highly similar isoforms, CYP2C9 and CYP2C19, that both have key roles in drug metabolism, alters the interactions and orientations of the enzymes in the ER membrane bilayer and so affects the substrate access tunnels, accounting for their differing substrate specificities [214]. Cui et al. have also focused on the differences within the CYP2C family, dynamics of its members bound to a shared substrate (diclofenac) revealed that sequence divergence at the active site residues causes heterogeneous variations in its secondary structures and affects the shape and chemical properties of the substrate-binding site [215].

A broader comparative study including CYP1A1, CYP1A2, CYP2A6, CYP2A13, CYP2B6, and CYP3A4 was performed investigating the regioselectivity of CYPs toward an abundantly present tobacco carcinogen [216]. Additionally, inhibitor selectivity can also be elucidated with the help of MD simulations as was shown by Wright et al. for clobetasol against CYP3A5 versus the structurally very similar CYP3A4 isoenzyme [217]. Fischer et al. used MD simulations to study the conservation and functionality of a superficial allosteric site, that was previously identified in the prokaryotic CYP101A1 and is involved in the regulation of ligand access to its buried binding pocket in the nine most relevant mammalian CYPs. They revealed that several mammalian enzymes of the CYP2 family could possess such an allosteric site [218]. A comprehensive review on the molecular modeling of CYP polymorphism using structure-based *in silico* tools was published by Martiny and Miteva [29].

Similarly to CYP enzymes, MD simulations have been also performed for phase II DMEs to better understand their mechanisms. Cook et al. have proposed a restricted substrate access mechanism explaining SULT selectivity using MD simulations. They suggested that the

substrate-binding pockets of SULT1A1 and SULT2A1 open and close in response to the binding of nucleotide with the active-site cap (a structural region controlling access to the active site) either excluding or admitting large substrates. They showed that access of a large substrate (fulvestrant) to the acceptor-binding pocket is restricted by cofactor (PAPS) binding, whereas access to small substrates is not affected. They hypothesized that at saturating PAPS levels, the concentration of the open enzyme form decreases to a minimum and as a consequence its affinity towards large substrates is weakened while its affinity towards small acceptors are not influenced [65]. In a different study, Cook et al. have also proposed a mechanism for the positive-synergy of a large group of SULT1A1 substrates that induce enzymatic activity. There are SULT1A1 substrates the affinities of which dramatically increase with saturating nucleotide levels. According to their model, such substrates induce a 'sandwich-like' residue-organization around the substrate phenolic moiety which stabilize the substrate nucleophilic hydroxyl in a reactive position [219].

The inhibitory mechanism of a potent, highly specific SULT1A1 inhibitor (mefenamic acid which is an NSAID), was investigated using MD simulations by Wang et al. and ligand-binding residues were identified [220]. Furthermore, Isvoran et al. performed MD simulations on the wild type and the allelic variants of SULT1A1, SULT1A1*2 (R213H) and SULT1A1*3 (M223V), to investigate the effects of amino acid substitutions on the enzyme dynamics in its apo and holo states. In particular, they identified increased flexibility at the loop regions surrounding the substrate-binding site [221].

Besides SULT1A1, SULT1E1 has also been the subject of studies including MD simulations. Rakers et al. have used MD simulations to investigate enzyme flexibility and sample protein conformations of the isoenzyme SULT1E1 before applying ensemble docking. They have observed large flexibility, especially at the lip region (the first loop of the three forming the gate to the active site) of the enzyme, which can significantly modulate the shape of the active site and therefore influence binding events. Furthermore, they identified a lysine residue (K85) on the lip region to be an essential element for regulating substrate access and selectivity. The inward flip of this residue causes blockage of the active site entry, which constrains ligand binding to smaller molecules [222].

In contrast to SULTs, less simulations have been performed on the UGT superfamily, primarily due to the lack of experimental structures available for human UGTs that would contain the substrate-binding domain. However, multiple studies have constructed homology models to promote the understanding of the enzymatic activity and its inhibition [223-225]. Nair et al. built a homology model of the human UGT2B7 enzyme based on a plant homolog crystal structure (UGT85H2 of *Medicago truncatula*) and performed MD simulations for the wild-type and two mutant forms (R259A and R259L) in the presence of either UDP-glucuronic acid or UDP-glucose, to identify key interactions with the different cofactors and investigate the differences in cofactor binding accounting for its selectivity towards UDP-glucuronic acid. They proposed that residues of the substrate-binding domain contribute to UDP-sugar binding, and that such residues confer UDP-sugar selectivity, in particular the residue R259 [226]. For

the isoenzymes UGT1A8 and UGT1A9, Fujiwara et al. built homology models based on a bacterial homolog (TDP-epi-vancosaminyltransferase of *Amycolatopsis orientalis*) and performed MD simulations at different temperatures (310 K and 360 K) to analyze the dynamical changes of UGT1A9 and UGT1A8. Human UGT1A9 is uniquely stable against heat treatment and they identified critical residues responsible for its thermal stability [227]. Subedi et al. have constructed a homology model for UGT1A1 based on a bacterial homolog crystal structure (Yijc of *Bacillus subtilis*) and used MD simulations to test the stability and trustworthiness of the homology model together with the docking results of cortisone and prednisone in the enzyme [228].

Running all-atom classical MD simulations on drug transporters may be unfeasible for biologically relevant time scales in the presence of explicit solvent and phospholipid membranes. Nevertheless, MD has been applied to assess stability of homology models as was in the case of the human SLC47A1 efflux transporter. Zhang et al. built a homology model of the human SLC47A1 (MATE1) transporter based on a bacterial transporter homolog crystal structure (NorM of *Vibrio cholerae*), and confirmed the stability of their model using MD simulation [229]. Tsigelny et al. modeled the structure of SLC22A6 (OAT1) based on the template of the bacterial glycerol-3-phosphate transporter structure (of *Escherichia coli*) and investigated its dynamics in a lipid bilayer [230, 231]. Adla et al. have built a homology model of the human SLCO1A2 (OATP1A2) transporter and tested predicted binding mode stabilities of various synthetic compounds, and identified key interacting transporter residues [232]. Similarly, Gebauer et al. built homology models of SLC22A1 (OCT1) and SLC22A2 (OCT2) based on a plant homolog crystal structure (sugar transport protein 10 of *Arabidopsis thaliana*), and used MD simulations to investigate the stability of docking modes of fenoterol enantiomers [233].

Given their outstanding importance in the mediation of drug disposition, many studies applying MD simulations have focused on ABCB1 and ABCG2. Before the appearance of high-resolution experimental structures of human ABC transporters embedded in lipid bilayer, thanks to the breakthrough advances in cryogenic electron microscopy, homology models were built for P-gp in various studies and such models were validated by using MD simulations [234-238]. A similar study was also published for ABCG2 [239].

Ever since, more reliable initial structures have been available for MD simulations and different studies have generated conformational diversity for molecular docking simulations and have investigated the drug transporter efflux mechanisms. Lagares et al. demonstrated with the help of MD simulations, that the presence of active and inactive compounds bound to ABCB1 influence the conformational distribution and dynamics of the nucleotide-binding domains, active compound-binding induced higher flexibility [240]. Xing et al. used atomistic MD simulations in explicit membrane and solvent environment to explore the effects of substrate and inhibitor binding on the conformational dynamics of ABCB1, and found that in the presence of substrates, the nucleotide-binding domains were closer and better aligned, suggesting that substrate binding may promote ATP hydrolysis, whereas inhibitors stabilized

them in a much more separated configuration, possibly impairing ATP-hydrolysis [241]. Zhang et al., working on the mouse P-gp structure, demonstrated that three drug molecules is the maximum that can simultaneously bind in the ABCB1 cavity [242].

To overcome the time limitation of classical MD simulations, enhanced MD simulations such as targeted MD, accelerated MD, and coarse-grained MD simulations (modeling where molecules are not represented by individual atoms, but by 'pseudo-atoms' approximating groups of atoms) have also been performed on ABCB1. Wang et al. have investigated the transport of an ABCB1 substrate (doxorubicin) [243] and later compared it to the transport of an inhibitor (verapamil) by the human ABCB1 using targeted MD simulations and identified the driving forces responsible for the translocation and concluded that they are identical for substrates and inhibitors, however, the residues involved are different [244].

Zhang et al. also used targeted MD simulations, to simulate the conformational rearrangements starting from the inward- to outward-facing states, using a homology-model for the inward-facing state, and identified both translational and rotational movements between the nucleotide-binding domains during the conformational transition [245]. Random accelerated MD combined with classical MD was performed on ABCB1 by Zhang et al. to simulate the efflux process of drugs and was complemented with metadynamics simulations (applied to estimate the free energy of a system, where ergodicity is hindered by larger energy barriers in most cases) to determine corresponding interaction free energies, the study used the mouse P-gp structure for the simulations [246].

Using coarse-grained MD simulations, Domicevica et al. examined protein-lipid interactions in a membrane mimicking the composition of brain epithelial cells [247], whereas Barreto-Ojeda et al. identified lipid pathways for ABCB1 in the inward-facing conformation in bilayers with different PC/PE lipid ratios [248]. Behmard et al. used steered dynamics simulations, introducing an imaginary external force to the small molecule to drive it through the transporter, combined with umbrella sampling, and analyzed interactions between the transporter and its substrate drugs. They hypothesized that van der Waals interactions are the main driving force in hindering the efflux of drugs in ABCB1 [249].

Several MD simulations have also been performed for ABCG2. Vesga et al. used MD to validate the docking poses of sixteen tetrahydroquinoline/4,5-dihydroisoxazole derivatives to develop new selective and potent inhibitors of ABCG2 through evaluating their stability [250]. Zhang et al. analyzed the stability of and provided information on the molecular interactions between regorafenib and the ABCG2 substrate binding cavity (cavity 1) using the pose from induced-fit docking simulation [251].

Long-timescale MD was performed complementing the experimental identification of a selective, porphyrin derivative inhibitor of ABCG2 capable of overcoming multidrug resistance *in vitro*, to describe interactions between the inhibitor and the transporter [252]. Ibrahim et al. identified eight promising high affinity ABCG2 inhibitors by screening molecules in the eMolecules, ChEMBL, and ChEBI databases, combining molecular docking and MD simulations

[253]. In a different study, they investigated the binding affinities of 181 drug candidates in clinical-trial or investigational stages to identify potential ABCG2 inhibitors, similarly combining molecular docking and MD simulations followed by molecular mechanics-generalized Born surface area (MM-GBSA) binding energy calculations, and identified three promising inhibitor candidates [254]. Again in a different study, they screened compounds in the Naturally Occurring Plant-based Anticancer Compound-Activity-Target (NPACT) database containing 1574 compounds and applied the same methodology as before [255]. The interactions and stability of different 2,4-disubstituted pyridopyrimidine derivatives were investigated using molecular docking and MD simulations in the study of Tadayon et al., they recommended further investigation of two promising compounds as ABCG2 inhibitors [256]. Wang et al. reported that a RAF kinase inhibitor effectively antagonizes ABCG2-mediated MDR *in vitro*, and their docking and MD simulations revealed that it binds to the substrate-binding cavity (cavity 1) [257].

In addition to testing the stability of the binding of different compounds for which classical MD can be a powerful tool, enhanced MD simulations are useful to simulate transporter events on larger timescales. Nagy et al. combined targeted MD and metadynamics simulations to simulate the translocation of a small substrate molecule of ABCG2, uric acid. They proposed the existence of drug binding cavities other than the central binding site, and observed an accelerated transport mechanism in the presence of membrane cholesterol [258].

Many of the previously discussed studies included MD simulations in combination with molecular docking. Molecular docking aims to predict possible ligand orientations with respect to an acceptor molecule. This is achieved by generating multiple protein-ligand complex configurations, using different search algorithms, and corresponding scoring by different scoring functions to energetically rank such complexes. MD can be performed both before molecular docking in order to generate a conformational ensemble and account for the flexibility of the protein, and after the docking to evaluate the stability of a given binding mode, identify statistically relevant interactions, and evaluate the dynamical behavior of the protein-ligand complex. Examples of such combinational studies include the identification of potential substrates or inhibitors of DMEs [222, 228, 259-262] and drug transporters [233, 250, 253-256].

7.1.2. QM/MM modeling of drug metabolism

The prediction of the intrinsic reactivity of the different functional groups of substrates requires more detailed descriptions. QM/MM (quantum mechanics/molecular mechanics) simulations combine *ab initio* quantum calculations in the proximity of a chemical process (at the active site region) and classical molecular mechanics at more distant regions that estimates dynamics on an atomic level [263]. QM/MM can be used to deduce reactivity coupled to the metabolism of drugs, however only at an extremely high computational cost [264]. As a consequence, QM/MM cannot be used for the filtering of large datasets, but it can be a powerful tool for the evaluation of the metabolism in terms of the reaction mechanism and corresponding activation energy of a given substrate molecule.

Several comprehensive reviews have been published on the application of QM/MM in the modeling of CYP enzyme drug metabolism [265-271]. Furthermore, in case of the CYP enzymes, two classes of inhibitors are known, type I binders that behave like 'normal protein ligands', and type II binders that bind directly to the iron-ion in the heme group through a semi-covalent bond. Such interactions are better described at the quantum level, and have been modeled in multiple studies [272-274].

For SULT1A1, Ma et al. investigated the metabolic mechanism of hydroxylated bromodiphenyl ethers and identified proton abstraction and sulfation steps during the reaction [275]. Besides the high computational costs, the major limitation of applying QM/MM is the requirement of an accurate 3D starting structure of the enzyme-ligand complex. No QM/MM studies have been performed to date on the human UGT superfamily, most probably due to the lack of available experimental structures of human UGTs that would contain the substrate-binding domain.

7.2. Ligand-based methods

7.2.1. Quantitative structure-activity relationship and machine learning models

In silico models can be used to predict small molecule interactions with DMEs and drug transporters. Predictions based on the properties of the ligands, widely used for DMEs and ABC transporters, make use of molecular descriptors (or variables) which are mathematical representations of different properties of the molecule. They quantify their topological, geometrical, physical, and chemical information such as the molecular weight, the number of different functional groups, or logP (which is a quantitative representation of the lipophilicity, and its value is obtained by measuring the partitioning of the molecule between an aqueous phase and a lipophilic phase which consists usually of water/n-octanol).

The descriptors can be classified as one-dimensional (1D, calculated from the molecular formula of the molecule, e.g. the number of different atoms, the molecular weight), two-dimensional (2D, calculated from the 2D chemical structure of the molecule, e.g. number of benzene-rings, number of H-bond donors), or three-dimensional (3D, representing structural information derived from the 3D conformation of the molecule, e.g. the solvent accessible surface, polar and nonpolar surface area) [276]. Such descriptors can be used to perform similarity searches in great molecular libraries, to identify candidate molecules which have similar physical/chemical properties to known active compounds based on the values of the different descriptors [277]. Molecular structures can also be described with the help of molecular fingerprints converting them to bit strings (vectors). The use of fingerprints facilitates the fast identification of structural similarities which is useful as similar structures may show similar biological activities [278].

By virtual screening in general, simple models are used enabling fast calculations at stages where high number of candidate molecules needs to be analyzed, whereas more complex models are employed at more advanced stages during drug development, requiring more time-consuming calculations and at the same time having higher prediction accuracy. Statistical models in general require a dataset of a large number of compounds with known

activity in order to create high accuracy, robust models. The structural diversity within the dataset determines the applicability domain of a given model [279].

Quantitative structure-activity relationship (QSAR) models have been developed to predict biological activity both towards DMEs and drug transporters. Models can be classified into two categories, regression or classification. Regression is the estimation of a continuous quantity (such as IC50 values or Ki) while classification is the prediction of discrete class labels (such as different classes of inhibitor potency) based on independent variables [280]. There are several statistical methods that are used in the construction of such models. Multiple linear regression (MLR) presumes a linear relationship between a scalar response and the independent variables, and the models make use of a linear predictor function the unknown parameters of which need to be estimated. MLR models have been built for different CYP isoforms, e.g. CYP1A2 [281, 282], CYP2D6 [283], CYP2C9 [284], or CYP3A4 [285]. Similar simplistic models based on MLR, predicting inhibitor/binding affinity or substrate uptake rate also exist for drug transporters, e.g. ABCB1 [286-288] or ABCG2 [289]. A more robust method is partial least squares (PLS) regression, which at first reduces the number of predictors by extracting a set of components that describe maximum correlation between the estimated quantity and the independent variables, then similarly to MLR, performs least-squares regression by minimizing the sum of the squares of the residual errors between prediction and observation. PLS models have been built for CYP isoenzymes [281, 282, 285, 290-292], for SULT and UGT isoforms [293], and drug transporters [294-297].

More complex machine learning approaches such as random forest (RF) or the more robust support vector machine (SVM), and even neural networks have been employed in the construction of prediction models. RF has been an appealing choice for its simplicity. Additionally, in some studies, the authors have also made use of the clear interpretation of its generated models thank to the easy determination of individual feature importance. RF is built on a large number of decision trees where each tree splits the data at several branches based on random subsets of features and makes its own prediction. The final prediction of the forest is obtained after aggregating the individual results by taking their average (in case of regression) or the majority vote (in case of classification) of the predictions.

Plonka et al. made use of RF classification to train models on large datasets to identify inhibitors of different CYP isoforms (CYP1A2, CYP2C9, CYP2C19, CYP2D6 and CYP3A4) [298]. RF models have also been applied to predict site of metabolism (SOM) of different enzymes. Information on SOM is useful in drug development, if an active metabolite has improved pharmacological, pharmacokinetic and toxicological profiles compared to the parent drug, it can be conveniently used as a lead or even advanced to the clinic whereas if a metabolite is unwanted, the SOM information can guide the structure modification to a direction that will deactivate or eliminate the unstable sites in order to avoid undesirable biotransformations [299]. An example of RF used for the prediction of SOM of CYP enzymes is the work of Sicho et al. who developed models for the prediction of global and isoform-specific regioselectivity of CYP3A4, CYP2D6, and CYP2C9 [300]. For the prediction of UGT-mediated metabolism,

Mazzolari et al. trained RF models to further classify UGT substrates whether the glucuronidation reaction occurs on an oxygen or on a nitrogen atom [301].

The prediction of interactions between ligands and drug transporters has also been addressed with RF models. Schwaha et al. built RF models among others to distinguish between ABCB1 substrate and non-substrate molecules and concluded that on the dataset they used, RF was the most suitable classification method (outperforming binary QSAR [302] and SVM) [303]. Poongavanam et al. used the combination of different machine learning techniques, among which RF, to build models for the prediction of ABCB1 substrates and inhibitors using a set of fingerprints representing the presence or absence of various functional groups [304]. Ohashi et al. developed a simplified *in vitro* screening method to evaluate ABCB1 substrates, and complemented it by building regression models to predict the ABCB1 mediated efflux, and three-class classification models to predict ABCB1 substrate potential (low/medium/high), both using RF machine learning. They also found that on their dataset, RF outperformed other machine learning techniques, SVM, neural networks, k-nearest neighbors, and AdaBoost in the classification and regression [305].

For ABCG2, Ghosh et al. developed a multi-QSAR approach using different methods based on structural fingerprints, in order to identify ABCG2 inhibitors and gain insight into the different important structural fingerprints modulating ABCG2 inhibition [306]. RF produced the best model out of the four machine learning approaches they used, gradient boosting, RF, SVM, and k-nearest neighbors. They concluded that among others, the presence of the nitro group at the para position of a substituted benzene ring is beneficial for ABCG2 inhibition, and also that the presence of an aromatic nitrogen attached to another nitrogen atom in a ring followed by branching is advantageous to design ABCG2 inhibitors with enhanced potency.

Support vector machine (SVM) is another supervised learning algorithm (it uses a labeled training data) that can be used for both classification and regression. In classification, SVM maximizes the width of gap separating the different classes in a given feature space and the use of different kernel functions enables both linear and non-linear classifications. Many of the previously discussed studies constructed models with both RF and SVM, as well as various other machine learning approaches. SVM have been long and broadly employed to build prediction models to identify substrates and inhibitors of different CYP isoforms [307-313]. For phase II DMEs, ML prediction models are not so widespread. For the classification of substrates and non-substrates of 12 human UGT isoforms, Sorich et al. developed and compared different ML approaches, partial least squares discriminant analysis, Bayesian regularized artificial neural network, and SVM based on 2D molecular descriptors and concluded that in their study, SVM outperformed the other two methodologies [314].

SVM has also been applied to drug transporters. The previously discussed study of Poongavanam et al. combined different ML approaches, a wrapper subset evaluator, RF, k-nearest neighbors, and SVM for the prediction of ABCB1 substrates based on fingerprints representing the presence or absence of various functional groups [304]. Leong et al. focused on the inhibitors of ABCB1, they predicted EC50 values *in silico* by combining pharmacophore

models and subsequent SVM regression [315]. Eric et al. used SVM and artificial neural networks to build prediction models of ABCB1 and ABCG2 substrates and inhibitors, and also highlighted similarities and distinctions of the molecular basis for transport and inhibition [316]. Montanari et al. addressed a similar question focusing on inhibitors of ABCB1 and ABCG2, they built models to predict inhibitors of both drug transporters and identify selective inhibitors based on fingerprints, MACCS keys, and 2D molecular descriptors [317]. They compared model performance using different ML tools including k-nearest neighbors, RF, and SVM. They found that on the carefully cleared dataset they used for the classification as either selective ABCB1, ABCG2, or common inhibitor, or non-inhibitor resulted in mediocre results, most probably due to the small number of molecules in the dataset. They also built models to interpret similarities and differences of known ABCB1 and ABCG2 inhibitors, with only two descriptors (number of hydrophobic atoms and number of aromatic atoms, combined in a single decision tree) they managed to separate ABCB1 inhibitors from ABCG2 inhibitors with a relatively good accuracy (>80 %).

Further studies using SVM for the prediction of ABCG2 substrates and inhibitors include the work of Zhong et al who combined a genetic algorithm for the feature selection, a conjugate gradient method for parameter optimization, and SVM for the training of ABCG2 substrate prediction models [318]. Similarly, for the prediction of ABCG2 substrates Hazai et al. built an SVM model using molecular descriptors [319]. Jiang et al. compared the performance of seven ML approaches for the prediction of ABCG2 inhibitors, and concluded that on their dataset SVM, deep neural networks, and extreme gradient boosting outperformed other approaches, and SVM yielded the best predictions [320]. Ding et al. estimated inhibitor IC50 values using a combined pharmacophore ensemble and SVM regression [321]. ABCG2 inhibitor classification models using several ML approaches including SVM, k-nearest neighbor, and neural networks were built by Belekar et al., they observed that SVM performed the best [322].

Other more complex ML approaches have also been employed as detailed for some of the studies discussed earlier, e.g. artificial neural networks [305, 314, 316, 320, 322]. The limitation of all *in silico* prediction models remains, however, the lack of a large amount of high quality, comparable experimental data. In most cases, the experimentally measured activities originate from different studies under different conditions, and it remains a challenging task to predict inhibitor or substrate activity with high accuracy using ligand-based *in silico* approaches.

7.3. Integrated structure-based and machine learning modeling

7.3.1. Prediction of DME ligands

In case the structure of the target protein is known, ligand-based techniques can be complemented with structural information to improve the quality of ligand-binding predictions. In 2013, the host laboratory trained the first machine learning classification models that integrated ligand- and structure-based information for the prediction of DME inhibition in the case of different SULT isoforms. The first models were trained by Martiny et al. on the SULT isoforms 1A1, 1A3, and 1E1 [323], directly followed by the models of Cook et al. for the isoforms 1A1 and 2A1 [324]. Martiny et al. explored protein flexibility using MD simulations and

subsequent docking to predict ligand binding affinities. They combined the predicted binding affinities with ligand-based information (extended connectivity fingerprints) in the training of different machine learning models for the prediction of SULT inhibition. Cook et al. also performed MD simulations complemented with docking, and based on experimental results correlated to the docking energies, a binding cutoff value of the docking score was determined to differentiate between ligands and inactive compounds.

For the sulfotransferase SULT1E1, an SVM model was trained by Rakers et al. for the prediction of substrates and inhibitors, they also combined ligand- and structure-based information [222]. MD simulations were performed to account for the important flexibility of the enzyme, which was followed by ensemble docking simulations, and finally by pharmacophore extraction which represented characteristic states of inhibition and sulfonation. Their prediction was based on hit identification via pharmacophore screening and further evaluation by SVM modeling to classify hits as substrates, inhibitors, or substrates with inhibitory potency at increasing concentrations.

In the host laboratory Martiny et al. created integrated models for a phase I DME, the isoenzyme CYP2D6 [36]. Protein conformational variability was accounted for by performing MD simulations, and docking simulations were performed. They used the predicted binding affinities together with extended connectivity fingerprints to train SVM, RF, and Naive Bayes classifier machine learning models. Recently, Goldwaser et al. created prediction models for the inhibition of the isoenzyme CYP2C9 [262]. Similarly, MD simulations were performed followed by ensemble docking, and the binding energies together with ligand-based physicochemical descriptors were combined in machine learning modeling. Their study was complemented with experimental validation of predicted inhibitors.

Huang et al. created models for the prediction of CYP-mediated site of metabolism by integrating flexible docking and reactivity calculations with machine learning algorithms [325]. They focused on the isoenzymes CYP1A2 and CYP2A6 and trained their models on known substrates. Further examples on the combination of ligand- and structure-based computational modeling for different CYP isoforms for the improvements in the prediction accuracy of DDIs can be found in the recent review by Kato [280].

7.3.2. Prediction of ABC transporter ligands

The integration of the two types of modeling, structure- and ligand-based, has also gained more importance recently in the prediction of ABC transporter ligands. Examples include the work of Esposito et al. who combined machine learning and MD to predict ABCB1 substrates [326]. They calculated MD fingerprints containing information from short MD simulations of the molecules in different environments (in solvent, in membrane, and in the ABCB1 substrate-binding pocket), and used them as descriptors for the training of machine learning models. Another study by Mahmud et al. included virtual screening based on molecular docking, and used QSAR (with PLS regression) to model the predicted binding affinities. The authors identified two compounds with outstanding docking scores, the stability of which were further validated by using MD simulations [327].

High resolution transporter structures embedded in lipid bilayers are crucial for accurate structure-based modeling. As discussed in the previous chapter dedicated to structure-based modeling, several computational studies have been performed on the analysis of ABC transporters and their interactions with ligands using recently published cryo-EM protein structures. More examples on either ligand-based or structure-based modeling for ABC transporter-ligand interactions are given in the reviews of Demel et al. [328] and Montanari et al. [329]. In the future it is expected that, similarly to DMEs, a more complex understanding and more accurate predictions of transporter-ligand interactions will be achieved with the help of novel integrated structure- and ligand-based studies.

B. Computational Modeling Tools

Computational modeling of proteins and their interactions with ligands is of increasing importance in drug discovery and development. Fast ways are necessary for early-stage drug discovery to search large ligand databases with drug-like properties that could have an activity against a target protein. Both ligand-based methods that are based on already known ligands, as well as structure-based methods in case the structure of the target protein (or a similar one) is known, can be used to identify substrates or inhibitors. Structure-based approaches include docking simulations that can be employed to find binding sites and poses quickly, and atomistic simulations such as molecular dynamics that can be performed to refine complexes and extract more accurate binding affinities.

1. Structure-based modeling of proteins

1.1. Interatomic energies, atomistic Force Fields

In silico modeling of the structures and dynamics of proteins requires a model of interatomic interactions. Force fields are used to define the potential energy as a function of atomic coordinates. The force fields used in protein studies are generally semiempirical, and interatomic interactions are treated with a classical mechanics approximation. The most commonly used software packages (NAMD, CHARMM, GROMACS, and AMBER) use the following terms in the force field definition:

$$U(\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_N) = \quad \text{The semi-empirical potential energy function}$$

$$\sum_i \frac{k_l}{2}(l_i - l_i^0)^2 + \quad \text{Bond stretching}$$

$$\sum_i \frac{k_\theta}{2}(\theta_i - \theta_i^0)^2 + \quad \text{Bond bending}$$

$$\sum_i \left\{\sum_{n=1}^{3} k_\varphi [1 + cos(n\varphi_i - \varphi_i^0)]\right\} + \quad \text{Bond torsion}$$

$$\sum_{i,j} \left(\frac{a_{i,j}}{r_{i,j}^{12}} - \frac{b_{i,j}}{r_{i,j}^{6}}\right) + \quad \text{Van der Waals interactions}$$

$$\sum_{i,j} \frac{332 q_i q_j}{\varepsilon r_{i,j}} \quad \text{Electrostatic interactions} \qquad (B.1)$$

where $\mathbf{r}$ refers to atomic coordinates, $\mathbf{l}$ to bond lengths, θ to bond angles, φ to dihedral angles, r_{ij} to interatomic distances, $\mathbf{q}$ to atomic charges. The superscript $\mathbf{0}$ indicates constants known from experiments, the different $\mathbf{k}$'s represent force constants, a_{ij} is the Lennard-Jones (LJ) repulsion coefficient, b_{ij} the LJ attraction coefficient, and ε the effective dielectric constant. The factor 332 arises from the conversion to obtain energy in kcal/mol given that distances are expressed in Ångströms, and charges as multiples of the charge on a proton.

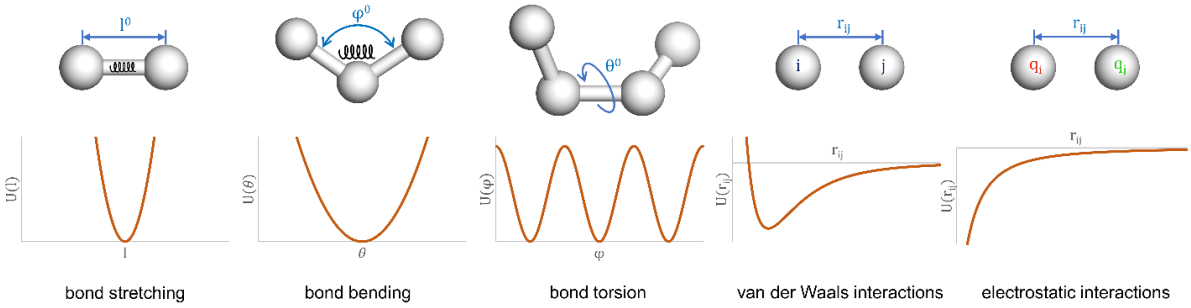

Figure B.1: Energy terms in the semi-empirical potential energy function.

Bond stretching energies arise from the stretching or compression of covalent bonds using a quadratic approximation, similar to ideal springs, relating the instantaneous bond length to its equilibrium, which is the most probable bond length, in terms of potential energy.

Bond-angle energies are similar, this term presumes that each bond angle has a most probable equilibrium state and deviations from it result in an increase in the potential energy, similar to an angular spring. Importantly, bond stretching is energetically more expensive than bond-angle bending, which is expressed in the corresponding force constants: $k_l >> k_\theta$.

Torsional-angle energies introduce small energy barriers between states differing in torsional degrees of freedom. Different dihedral configurations (of the backbone) account for the largest conformational variability between different stable states of a given protein. Sidechains can also adapt different dihedral configurations accounting for the fine adaptation to their environment. In general, multiple prominent torsional states are possible with corresponding minima, and the corresponding energy term is expressed with a periodic (cosine) function. The so far discussed bonded terms were the first to be used to describe intramolecular potential energies.

More complicated conformational properties can be described by introducing non-bonded energy terms. Such terms describe the interactions between atoms that are not covalently bonded and are separated by at least three (or in some cases four) covalent bonds in between them.

Two neutral atoms exhibit weak attraction from a distance (a so-called dispersion force) as a consequence of the dipoles that arise in their electron clouds due to the distortion caused by the presence of another atom. This attraction is short-ranged and decreases as $1/r_{ij}^6$ with increasing interatomic distances. Two neutral atoms also exhibit a repulsion, which is even shorter-ranged than the attraction, $1/r_{ij}^{12}$, arising from the Pauli exclusion principle to avoid the overlap of electron clouds and steric clashes. The balance of attraction and repulsion results in a stable separation of the two atoms, a stable non-covalent interaction. This is referred to as the van der Waals, Lennard-Jones (LJ), or 6-12 potential. The parameters a_{ij} and b_{ij} are specific to the atom types engaging in the interaction. More accurate functional forms also exist, but due to its simplicity, the 6-12 LJ potential is widely used in most semi-empirical biomolecule force fields.

Finally, between atoms that are polar or (partially) charged, an additional Coulomb interaction arise. Coulomb's principle describes that two like charges repel each other whereas opposite charges attract each other, and that the absolute value of the electric potential is proportional to the product of the magnitudes of the two charges and inversely proportional to the distance separating them. The dielectric constant captures the weakening effect of polarizable media where the given medium shields the two charges from each other. Polar atom, unlike charged atoms, have partial charges which are estimated from quantum mechanical modeling of small molecules. Hydrogen bonds are generally not explicitly present in the potential energy formula, but are simplistically captured by electrostatic attraction between the partial charges on the atoms forming the given hydrogen bond [330].

Solvent interactions can be handled either explicitly, which treats waters and solvent ions as individual molecules, or implicitly, treating solvent as a homogenous continuum. Explicit solvent modeling is more accurate, however, it requires considerably more computing power.

1.2. Conformational sampling and molecular simulations

Molecular simulations can be used in order to predict energetically favorable, (most populated) stable states, and understand dynamic processes of biomolecules. Such states correspond to low free energies, which requires on the one hand low enthalpy and on the other hand high entropy, following the equitation:

$$G = H - TS \qquad (B.2)$$

where G is the free energy, H the enthalpy (related to the potential energy), T the temperature, and S the entropy.

1.2.1. Energy minimization

The goal of energy minimization is to identify conformations of low potential energy by exploring the closest local minimum. Using a potential energy function (similar to what is defined in Equation (B.1), local minima are described by having their first derivative, the gradient vector as null, and their second derivative, the Hessian matrix as positive semidefinite. Various methods have been developed to follow gradients downhill on mathematical surfaces, such as steepest descent, conjugate-gradient, or Newton-Raphson, to eventually arrive at a local minimum. The main limitations of energy minimization are that it reveals no dynamics, it cannot cross energy barriers as it only moves downhill and as a consequence the identified minimum is likely to be only a local minimum, and it only minimizes potential energy, not free energy, which may correspond to an improbable, unpopulated state.

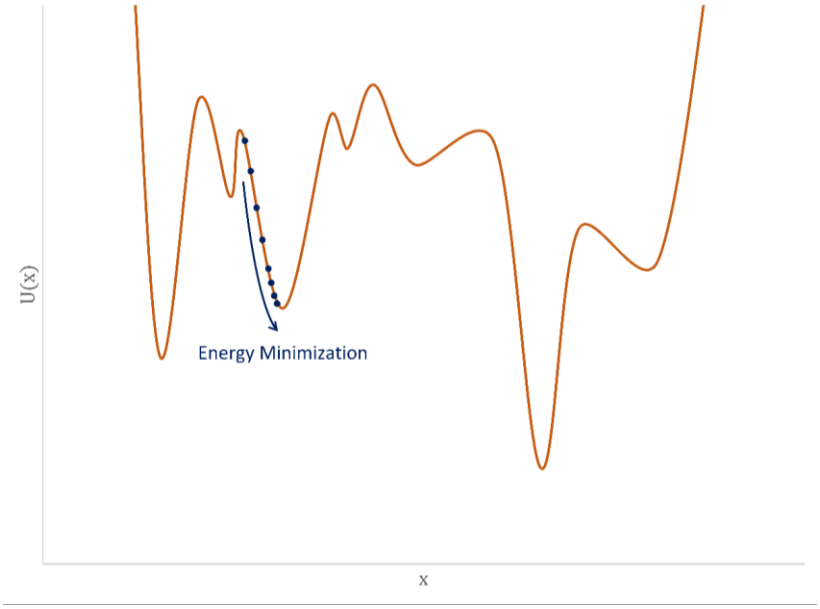

Figure B.2: Schematic view of energy minimization.

1.2.2. Molecular dynamics

To provide a proper description of the conformational population, free energies need to be estimated so that a combination of low energy and high entropy can be ensured. Molecular dynamics (MD) simulations solve Newton's equations of motion iteratively thereby creating a time-dependent conformational trajectory. MD is capable of crossing energy barriers and reaching states with lower free energy while it also provides dynamical (time evolution) information. For each atom, Newton's equation of motion can be applied, which states that the acting force on a given body (in case its mass is constant) is equal to the mass of the body multiplied by the acceleration of its center of mass (i.e. the second derivative of its position function):

$$\boldsymbol{f}_i = m_i \frac{d^2 \boldsymbol{r}_i}{dt^2} = -\nabla_{\boldsymbol{r}_i} U \qquad (B.3)$$

where $\mathbf{f_i}$ is the force acting on atom $\mathbf{i}$, $\mathbf{m_i}$ is its mass, $\mathbf{r_i}$ its position function, and $\mathbf{U}$ the potential energy function. The acting force on the given atom is given by the negative gradient of the potential energy function with respect to the three components (xyz) of the atom's position vector, the negative sign demonstrates that the acting force points towards lower potentials and the atom accelerate in that direction. In computers, the differential equations in Equation B.3 are solved numerically, the equations are expressed in terms of finite differences, and integrating over previous time steps determines the new state of the system. MD simulations are deterministic, however, they exhibit chaotic behaviors, i.e. small perturbations in the initial conditions are amplified exponentially and can lead to vastly different, unpredictable behavior [331]. To obtain unique solutions of the differential equations, initial and boundary conditions need to be defined. Then, to calculate new positions, velocities, and accelerations knowing previous states, different numerical integrations, such as the widely applied algorithms in MD simulations: the Verlet or the leapfrog algorithms, can be used.

The Verlet algorithm relies on the Taylor series expansion of the position vector $\mathbf{r}(t)$ to determine the state after a timestep, at $t+\Delta t$. As initial conditions, the Verlet algorithm takes two successive position vectors. The expression of the position vector at time steps t-Δt and t+Δt using the Taylor series expansion around $\mathbf{r}(t)$ are the following:

$$\boldsymbol{r}(t+\Delta t) = \boldsymbol{r}(t) + \boldsymbol{v}(t)\Delta t + \frac{1}{2}\boldsymbol{a}(t)(\Delta t)^2 + \cdots \; higher\; order\; terms \qquad (B.4)$$

$$\boldsymbol{r}(t-\Delta t) = \boldsymbol{r}(t) - \boldsymbol{v}(t)\Delta t + \frac{1}{2}\boldsymbol{a}(t)(\Delta t)^2 - \cdots \; higher\; order\; terms \qquad (B.5)$$

where $\mathbf{r}(t)$ is the instantaneous position vector, $\mathbf{v}(t)$ the velocity, $\mathbf{a}(t)$ the acceleration, and Δt the integration time step. The sum of the Equations B.4 and B.5, truncated at their third order terms (i.e. the precision of the equation is till the third order), results in the Verlet equation for the updated position vector:

$$\boldsymbol{r}(t+\Delta t) = 2\boldsymbol{r}(t) - \boldsymbol{r}(t-\Delta t) + (\Delta t)^2\boldsymbol{a}(t) \qquad (B.6)$$

The velocities can be evaluated at each time step by taking the difference of Equations B.4 and B.5, which results in the velocity formula:

$$\boldsymbol{v}(t) = \frac{\boldsymbol{r}(t+\Delta t) - \boldsymbol{r}(t-\Delta t)}{2\Delta t} \qquad (B.7)$$

The leapfrog algorithm calculates velocities at half-time steps: $\mathbf{v}(t\pm\Delta t)$. The leapfrog algorithm takes a position vector and a set of velocities for each atom as initial conditions. The initial velocities can be randomly drawn according to the Maxwell-Boltzmann velocity distribution:

$$p(v_{i,x}) = \sqrt{\frac{m_i}{2\pi RT}}\, exp\left(\frac{m_i v_{i,x}^2}{2RT}\right) \qquad (B.8)$$

for each component x,y, and z, where $p(v_{i,x})$ is the probability of the velocity x-component of the velocity vector $\mathbf{v}_i$ for atom i, m_i is its mass, T the temperature of the simulation, and R the universal gas constant. The velocity at $t+\Delta t$ is defined presuming a uniformly accelerated motion (constant force) during $\Delta t/2$:

$$\boldsymbol{v}(t+\Delta t/2) = \boldsymbol{v}(t) + \boldsymbol{a}(t)\,\Delta t/2 \qquad (B.9)$$

The position at the next step is defined presuming a uniform motion (resultant force equals zero) during Δt:

$$\boldsymbol{r}(t+\Delta t) = \boldsymbol{r}(t) + \boldsymbol{v}(t+\Delta t/2)\Delta t \qquad (B.10)$$

Boundary conditions are usually handled through the use of identical replicas of the simulation box around the central box, forming a grid, for the calculation of interactions at the edges (periodic boundary conditions). To maintain concentrations and conserve the total number of molecules, molecules leaving the simulation box are assumed to reenter at the opposite side. All numerical integration methods require the use of a sufficiently small integration time step (Δt) in order to ensure that errors due to their assumptions/approximations remain negligible. Such a time step needs to be shorter than the system's fastest motions, for proteins these are the bond vibrations, a typical time-step used in classical MD is 1-2 fs.

1.2.3. Normal mode analysis

Normal mode analysis (NMA) can be efficiently used for identifying and describing the slowest intrinsic motions of macromolecules, which in nature, generally correspond well to collective functional movements. NMA has become one of the standards to study the dynamics of macromolecules. NMA relies on the harmonic approximation of the potential energy function around a given local minimum. The calculation of all-atomic normal modes in a given forcefield includes the following three steps:

- Potential energy minimization to reach an equilibrium state (a local minimum)
- Calculation of the Hessian matrix (matrix of second derivatives of the potential energy with respect to the mass-weighted atomic coordinates)
- Diagonalization of the Hessian matrix

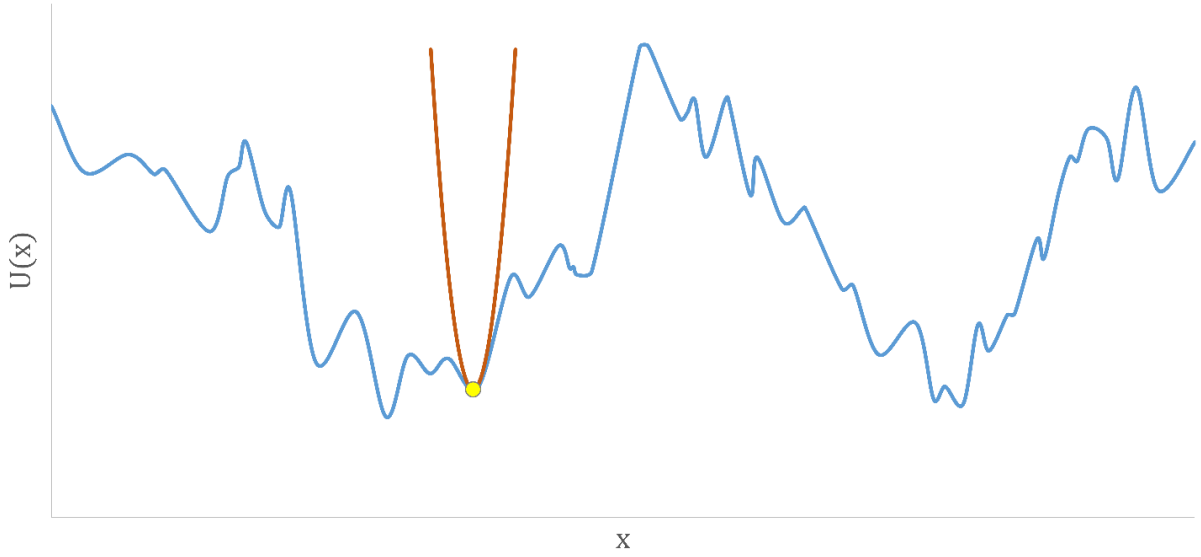

Figure B.3: Schematic representation of the harmonic approximation of the potential energy surface (blue) by NMA (red).

NMA is usually performed in vacuum, the potential energy thus is a function of 3N coordinates, N being the number of atoms in the system. The potential energy can be expanded in a Taylor series around a local minimum. The multivariable Taylor expansion is expressed as:

$$U(\boldsymbol{r}) \cong U(\boldsymbol{r}^0) + \boldsymbol{\nabla} U(\boldsymbol{r}^0)^T(\boldsymbol{r}-\boldsymbol{r}^0) + \frac{1}{2}(\boldsymbol{r}-\boldsymbol{r}^0)^T \boldsymbol{H}_0(\boldsymbol{r}-\boldsymbol{r}^0) + \cdots higher\ ord.terms \quad (B.11)$$

where $\mathbf{U(r)}$ is the potential energy function, $\mathbf{r}$ is 3N dimensional position vector, $\mathbf{H}$ the Hessian matrix, and the superscripts $\mathbf{0}$ correspond to the equilibrium state. The gradient in an energy minimum is equal to the null vector. If the reference $U^0 = 0$ is taken as reference state and terms higher than the quadratic are neglected, the approximation of the potential energy function becomes quadratic:

$$U(\boldsymbol{r}) \cong \frac{1}{2}(\boldsymbol{r}-\boldsymbol{r}^0)^T \boldsymbol{H}_0(\boldsymbol{r}-\boldsymbol{r}^0) = \frac{1}{2}\sum_{\boldsymbol{r}_i}\sum_{\boldsymbol{r}_j} \left.\frac{\partial^2 U(\boldsymbol{r})}{\partial \boldsymbol{r}_i \partial \boldsymbol{r}_j}\right|_{\boldsymbol{r}_0} (\boldsymbol{r}_i - \boldsymbol{r}_i^0)(\boldsymbol{r}_j - \boldsymbol{r}_j^0) \quad (B.12)$$

The first derivative of the potential energy function with respect to $\mathbf{r}$ is expressed as:

$$\nabla U(\boldsymbol{r}) = \boldsymbol{H}_0(\boldsymbol{r} - \boldsymbol{r}^0) \qquad (B.13)$$

Using Newton's equation of motion, the following formula is obtained:

$$-\nabla U(\boldsymbol{r}) = -\boldsymbol{H}_0(\boldsymbol{r} - \boldsymbol{r}^0) = \boldsymbol{M}\frac{d^2\boldsymbol{r}(t)}{dt^2} \qquad (B.14)$$

where M is the mass matrix, containing atomic masses in its diagonal. Introducing mass-weighted coordinates, $\boldsymbol{q}_i = \sqrt{m_i}\boldsymbol{r}_i$, where $\mathbf{r}_i$ is the position vector and m_i the mass of atom i, $\boldsymbol{q} = \sqrt{\boldsymbol{M}}\boldsymbol{r}$, and the mass-weighted Hessian, $\boldsymbol{K}_0 = \sqrt{\boldsymbol{M}}^{-1}\boldsymbol{H}_0\sqrt{\boldsymbol{M}}^{-1}$, Equation B.14 can be simplified (using the notation $\ddot{\boldsymbol{r}}$ for the second derivative with respect to time):

$$\boldsymbol{M}\ddot{\boldsymbol{r}} = -\boldsymbol{H}_0\boldsymbol{r} \qquad (B.15)$$
$$\ddot{\boldsymbol{r}} = -\boldsymbol{M}^{-1}\boldsymbol{H}_0\boldsymbol{r} \qquad (B.16)$$
$$\sqrt{\boldsymbol{M}}\ddot{\boldsymbol{r}} = -\sqrt{\boldsymbol{M}}\boldsymbol{M}^{-1}\boldsymbol{H}_0\boldsymbol{r} = -\sqrt{\boldsymbol{M}}\sqrt{\boldsymbol{M}}^{-1}\sqrt{\boldsymbol{M}}^{-1}\boldsymbol{H}_0\boldsymbol{r} = -\sqrt{\boldsymbol{M}}^{-1}\boldsymbol{H}_0\boldsymbol{r} \qquad (B.17)$$
$$\sqrt{\boldsymbol{M}}\ddot{\boldsymbol{r}} = -\sqrt{\boldsymbol{M}}^{-1}\boldsymbol{H}_0\sqrt{\boldsymbol{M}}^{-1}\sqrt{\boldsymbol{M}}\boldsymbol{r} \qquad (B.18)$$
$$\ddot{\boldsymbol{q}} = -\boldsymbol{K}_0\boldsymbol{q} \qquad (B.19)$$

The set of 3N coupled differential equations can be easier solved after coordinate transformation to a set of 3N decoupled differential equations. The Hessian matrix and the mass-weighted Hessian matrix are both symmetric which is a sufficient condition to be diagonalizable. The decoupling of the differential equations is ensured through the coordinate transformation by an orthonormal matrix T for which $\boldsymbol{T}^T\boldsymbol{T} = \boldsymbol{T}\boldsymbol{T}^T = \mathbf{1}$, furthermore it is required that T diagonalizes the mass-weighted Hessian (K_0) so that each differential equation becomes decoupled from all others, the matrix D_0 should be diagonal and is defined as $\boldsymbol{D}_0 = \boldsymbol{T}^{-1}\boldsymbol{K}_0\boldsymbol{T}$. Then, the column vectors of T form a basis consisting of eigenvectors of K_0, whereas the diagonal matrix D_0 has the corresponding eigenvalues as its diagonal elements. Equation B.15 can be rewritten after transformation to the space spanned by the eigenvectors of K_0, using the substitution $\boldsymbol{q} = \boldsymbol{T}\boldsymbol{y}$ and $\ddot{\boldsymbol{q}} = \boldsymbol{T}\ddot{\boldsymbol{y}}$ (as T is time-independent) and thus $\ddot{\boldsymbol{y}} = \boldsymbol{T}^{-1}\ddot{\boldsymbol{q}}$:

$$\boldsymbol{T}^{-1}\ddot{\boldsymbol{q}} = -\boldsymbol{T}^{-1}\boldsymbol{K}_0\boldsymbol{q} = -\boldsymbol{T}^{-1}\boldsymbol{K}_0\boldsymbol{T}\boldsymbol{T}^{-1}\boldsymbol{q} \qquad (B.20)$$
$$\ddot{\boldsymbol{y}} = -\boldsymbol{D}_0\boldsymbol{y} \qquad (B.21)$$

and thus in the space of the eigenvectors of $\mathbf{K}_0$, the 3N differential equations become decoupled. Each differential equation has the form of an ideal harmonic oscillator $\ddot{y}_i = -D_{0,ii}y_i = -\lambda_i y_i$, where λ_i is the i^{th} eigenvalue of $\mathbf{K}_0$. The solution to such differential equation has the form: $y_i(t) = y_i^0 + C_i\cos(\omega_i t + \varphi_i)$, with $\omega_i \coloneqq \sqrt{\lambda_i}$, which is the vibrational frequency (in an energy minimum the Hessian is positive semidefinite, all of its eigenvalues are nonnegative), C_i is the magnitude of the oscillation, y_i^0 represents the equilibrium state and φ_i an arbitrary phase factor. The solutions can be transformed back into the Cartesian-space by using $\boldsymbol{q} = \boldsymbol{T}\boldsymbol{y}$ and accounting for the initial condition:

$$\boldsymbol{q}(t) = \boldsymbol{q}^0 + \sum_{i=1}^{3N}\boldsymbol{a}_i C_i(\omega_i t + \varphi_i) \qquad (B.22)$$

where $\mathbf{a}_i$ is the i^{th} eigenvector of $\mathbf{K}_0$ (i.e. the i^{th} column vector of the matrix $\mathbf{T}$) and ω_i the square-root of the i^{th} eigenvalue of $\mathbf{K}_0$. They can be calculated by solving the eigenvalue equation: $\boldsymbol{K}_0\boldsymbol{a}_i = \lambda_i\boldsymbol{a}_i$. The physical interpretation of $\mathbf{a}_i$ gives the direction of the vibration with respect to the equilibrium state, whereas ω_i the corresponding vibrational frequency. All 3N eigenvectors $\{\mathbf{a}_i\}$ are orthonormal, hence the name normal modes, thus they are linearly independent, and they form a new complete basis set for the molecule, enabling the expression of molecule conformations in form of normal mode coordinates. The quadratic approximation of the molecular dynamics can be described as 3N independent vibrations around the equilibrium state, each along a given normal mode with a corresponding frequency. There are six zero eigenvalues of K_0, associated with rigid-body translations and rotations, which do not influence the potential energy. Given the normal modes are of unit length, the actual vibrational magnitude depends on the temperature, and is proportional to $1/\omega_i$ the square fluctuation to $1/\omega_i^2$, the lowest-frequency modes make the largest contribution to the overall motion. The harmonic approximation to the square fluctuation of atomic vibrations is given by the formula:

$$\langle|\Delta\boldsymbol{q}_i|^2\rangle = k_B T \sum_{j=1}^{3N-6} \frac{|\boldsymbol{a}_{i,j}|^2}{\omega_i^2} \qquad (B.23)$$

where k_B is the Boltzmann constant, $\mathbf{T}$ the absolute temperature, $\mathbf{N}$ the number of atoms, and $\mathbf{a}_{i,j}$ consists of the xyz components of the j^{th} normal mode of nonzero eigenvalue corresponding to atom $\mathbf{i}$.

1.3. Analyzing conformational ensembles

Given a set of conformations (e.g. trajectories generated by molecular dynamics simulation), different measures can be calculated in order to extract statistically relevant information. Measures like the root mean square deviation and the radius of gyration provides information per conformation, and their variation can be evaluated within the ensemble, whereas root mean square fluctuation provides atomistic statistics that already incorporate all the conformations in the ensemble.

1.3.1. Root Mean Square Deviation (RMSD)

Given a conformation in the 3D space, a reference structure is required to compute RMSD. It measures the square root of the average Euclidean squared-distances between all matching atom pairs for the given conformation and the reference structure. Its value is dependent on rigid-body transformations (translation and rotation), generally the RMSD denotes its minimum. An optimal overlap is achieved between two conformations if the corresponding RMSD is minimized with respect to relative translation and rotation movements. RMSD is given by the formula:

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left\|\boldsymbol{r}_i - \boldsymbol{r}_i^0\right\|^2} \qquad (B.24)$$

where $\mathbf{N}$ is the number of matching atom pairs considered for the RMSD calculation, $\boldsymbol{r}_i$ denotes the xyz coordinates of atom $\mathbf{i}$ in the given conformation, and $\boldsymbol{r}_i^0$ the xyz coordinates of the

same atom in the reference structure. To obtain statistical information within a conformational ensemble, the distribution of the RMSD calculated for the conformations with respect to the same reference can be evaluated. Furthermore, if the ensemble is time dependent (a conformational trajectory),its time evolution $RMSD(\boldsymbol{r}(t))$ can also be monitored.

1.3.2. Radius of gyration (R_{gyr})

The calculation of the R_{gyr} does not require a reference structure it is entirely defined by a given conformation. It provides information about the overall compactness of a given conformation, about the distribution of atoms with respect to the center of mass. A low R_{gyr} value indicates that the atoms in the given conformation are close to their center of mass, whereas a high R_{gyr} value indicates the opposite, that the atoms are far from their center of mass. Mathematically, the R_{gyr} measures the square root of the mass-weighted average Euclidean squared-distances between the individual atoms and the center of mass. The R_{gyr} is given by the formula:

$$R_{gyr} = \sqrt{\frac{1}{\sum_{i=1}^{N} m_i} \sum_{i=1}^{N} m_i \|\boldsymbol{r}_i - \boldsymbol{r}_{CM}\|^2} \qquad (B.25)$$

where N is the number of atoms in the conformation, $\boldsymbol{r}_i$ denotes the xyz coordinates and $\mathrm{m_i}$ the mass of atom i; and $\boldsymbol{r}_{CM}$ the xyz coordinates of the center of mass which is defined by the formula:

$$\boldsymbol{r}_{CM} = \frac{1}{\sum_{i=1}^{N} m_i} \sum_{i=1}^{N} m_i \boldsymbol{r}_i \qquad (B.26)$$

Similarly to RMSD, in case of a conformational ensemble, the distribution of the R_{gyr} can be analyzed, and in case the ensemble is time-dependent, its time evolution $R_{gyr}(\boldsymbol{r}(t))$ can also be monitored to e.g. identify larger conformational transitions or follow convergence.

1.3.3. Root Mean Square Fluctuation (RMSF)

The calculation of atomic fluctuations already incorporates statistical evaluation. Generally, internal fluctuations are of interest. Global fluctuations originating from translational and rotational movements of the molecule can be eliminated during a preceding superposition step by minimizing the RMSD of all conformations with respect to a given reference structure. The RMSF provides information about the flexibility of a given atom among a set of conformations. Mathematically the RMSF of a given atom measures the square root of the average squared Euclidean distances between its instantaneous positions and its average position over the conformational ensemble. The RMSF of a given atom is defined by the formula:

$$RMSF_i = \sqrt{\frac{1}{E} \sum_{s=1}^{E} \left\| \boldsymbol{r}_{i,s} - \boldsymbol{r}_{i,avg} \right\|^2} \qquad (B.27)$$

where E is the number of conformations in the ensemble, $\boldsymbol{r}_{i,s}$ denotes the xyz coordinates of atom i in the sample conformation s, and $\boldsymbol{r}_{i,avg}$ the xyz coordinates of the average position of atom i in the conformational ensemble. The average position of atom i is defined by the formula:

$$\boldsymbol{r}_{i,avg} = \frac{1}{E} \sum_{s=1}^{E} \boldsymbol{r}_{i,s} \qquad (B.28)$$

Typically, the RMSF of alpha carbon atoms are calculated to identify rigid protein regions (small RMSF values) and regions of high flexibility (high RMSF values).

1.3.4. Principal Component Analysis (PCA)

Principal Component Analysis is a broadly used algorithm in many different fields, it is used to reduce dimensionality of the data in order to increase interpretability while still preserving as much information possible. For conformational ensembles, the aim can be formulated as to decrease the dimensionality from the full detail description of all degrees of freedom, down to a small number of reduced coordinates that still capture the essential features encoded in the ensemble. This is achieved by an orthogonal linear coordinate transformation from the original Cartesian space to a space spanned by a set of uncorrelated collective variables (the principal components, PCs) that are ranked according to the corresponding variation described by the given PC. Similarly to RMSF calculations, PCA is usually preceded by a superposition step to filter out translational and rotational movements.

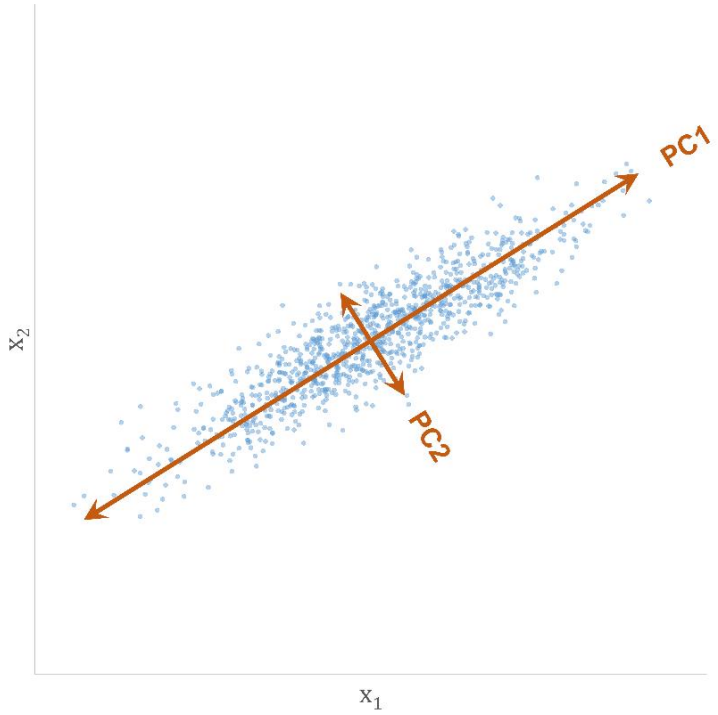

Figure B.4: Principal Components of a two-dimensional distribution. The characteristics of the distribution can be well described by using only PC1.

The underlying mathematics of PCA can be formulated as follows. The observation matrix is given, containing all the sample conformation coordinates as column vectors

$$\boldsymbol{R} = [\boldsymbol{r}_1 | \dots | \boldsymbol{r}_E] \qquad (B.29)$$

where E is the number of conformations in the ensemble, and $\boldsymbol{r}_s,\ s = 1 \dots E$ are the Cartesian coordinates of a given sample conformation s. A linear function defined by the unit vector $\mathbf{p}_1$ generate projections of the sample conformations:

$$\boldsymbol{p}_1^T \boldsymbol{R} = [\boldsymbol{p}_1^T \boldsymbol{r}_1 \quad \dots \quad \boldsymbol{p}_1^T \boldsymbol{r}_E] \qquad (B.30)$$

The corresponding estimated variance is:

$$Var(\boldsymbol{p}_1^T\boldsymbol{R}) = \frac{1}{E-1}\sum_{s=1}^{E}\left(\boldsymbol{p}_1^T\boldsymbol{r}_s - \boldsymbol{p}_1^T\boldsymbol{r}_{avg}\right)^2$$

$$Var(\boldsymbol{p}_1^T\boldsymbol{R}) = \frac{1}{E-1}\sum_{s=1}^{E}\left[\boldsymbol{p}_1^T\left(\boldsymbol{r}_s - \boldsymbol{r}_{avg}\right)\right]^2 = \frac{1}{E-1}\sum_{s=1}^{E}\boldsymbol{p}_1^T\left(\boldsymbol{r}_s - \boldsymbol{r}_{avg}\right)\cdot\boldsymbol{p}_1^T\left(\boldsymbol{r}_s - \boldsymbol{r}_{avg}\right)$$

$$Var(\boldsymbol{p}_1^T\boldsymbol{R}) = \frac{1}{E-1}\sum_{s=1}^{E}\boldsymbol{p}_1^T\left(\boldsymbol{r}_s - \boldsymbol{r}_{avg}\right)\times\left(\boldsymbol{r}_s - \boldsymbol{r}_{avg}\right)^T\boldsymbol{p}_1$$

$$Var(\boldsymbol{p}_1^T\boldsymbol{R}) = \boldsymbol{p}_1^T\left[\frac{1}{E-1}\sum_{s=1}^{E}\left(\boldsymbol{r}_s - \boldsymbol{r}_{avg}\right)\times\left(\boldsymbol{r}_s - \boldsymbol{r}_{avg}\right)^T\right]\boldsymbol{p}_1$$

$$Var(\boldsymbol{p}_1^T\boldsymbol{R}) = \boldsymbol{p}_1^T\boldsymbol{S}\boldsymbol{p}_1 \qquad (B.31)$$

where $\boldsymbol{r}_{avg}$ denotes the average conformation over the ensemble, and $\mathbf{S}$ the estimated covariance matrix of the original Cartesian variables and is given by the formula:

$$\boldsymbol{S} = \frac{1}{E-1}\sum_{s=1}^{E}\left(\boldsymbol{r}_s - \boldsymbol{r}_{avg}\right)\times\left(\boldsymbol{r}_s - \boldsymbol{r}_{avg}\right)^T \qquad (B.32)$$

The problem of finding the first PC can be formulated as identifying the unit vector $\boldsymbol{p}_1$ that maximizes the variance of the corresponding projections:

$$\underset{\boldsymbol{p}_1}{argmax}\ \boldsymbol{p}_1^T\boldsymbol{S}\boldsymbol{p}_1\ \ subject\ to\ \boldsymbol{p}_1^T\boldsymbol{p}_1 = 1 \qquad (B.33)$$

The constrained maximization problem can be expressed with the help of a Lagrangian: $L(\boldsymbol{p}_1, \lambda) := f(\boldsymbol{p}_1) - \lambda g(\boldsymbol{p}_1)$, where $f(\boldsymbol{p}_1) = \boldsymbol{p}_1^T\boldsymbol{S}\boldsymbol{p}_1$ is the target function and $g(\boldsymbol{p}_1) = \boldsymbol{p}_1^T\boldsymbol{p}_1 - 1$ the constraint function with the corresponding Lagrange multiplier λ arising from the constraint $\boldsymbol{p}_1^T\boldsymbol{p}_1 = 1$. Finding a stationary point of the Lagrangian function $L(\boldsymbol{p}_1, \lambda)$ is a necessary condition to find the maximum of the target function with the given constraints. The Lagrange function has the form:

$$L(\boldsymbol{p}_1, \lambda) = \boldsymbol{p}_1^T\boldsymbol{S}\boldsymbol{p}_1 - \lambda(\boldsymbol{p}_1^T\boldsymbol{p}_1 - 1) \qquad (B.34)$$

In a stationary point of the Lagrange function its gradient is zero: $\boldsymbol{\nabla}L(\boldsymbol{p}_1^*, \lambda^*) = \mathbf{0}$. Then the partial derivative with respect to the vector $\mathbf{p}_1$ is:

$$\frac{\partial L(\boldsymbol{p}_1^*, \lambda^*)}{\partial \boldsymbol{p}_1} = 2\boldsymbol{S}\boldsymbol{p}_1 - 2\lambda\boldsymbol{p}_1 = \mathbf{0} \qquad (B.35)$$

which becomes the following eigenvector equation:

$$\boldsymbol{S}\boldsymbol{p}_1 = \lambda\boldsymbol{p}_1 \qquad (B.36)$$

where $\boldsymbol{p}_1$ is an eigenvector of the estimated covariance matrix $\mathbf{S}$, and λ is the corresponding eigenvalue. In order to maximize the target function:

$$f(\boldsymbol{p}_1) = \boldsymbol{p}_1^T\boldsymbol{S}\boldsymbol{p}_1 = \boldsymbol{p}_1^T\lambda\boldsymbol{p}_1 = \lambda\boldsymbol{p}_1^T\boldsymbol{p}_1 = \lambda \qquad (B.37)$$

the largest eigenvalue must be chosen (λ_1), its value directly equals to the maximal variance, and the corresponding eigenvector $\mathbf{p}_1$ is the first PC of the ensemble $\mathbf{R}$.

Selecting the second PC maximizes the variance of the projections by $\mathbf{p}_2$:

$$\underset{\boldsymbol{p}_2}{argmax}\, \boldsymbol{p}_2^T \boldsymbol{S} \boldsymbol{p}_2 \; subject\; to\; \boldsymbol{p}_2^T \boldsymbol{p}_2 = 1 \; and \; \boldsymbol{p}_1^T \boldsymbol{p}_2 = 0 \qquad (B.38)$$

which means that in addition to the previous maximization, an additional constraint is added i.e. the projections by $\mathbf{p}_2$ must be uncorrelated from the projections by $\mathbf{p}_1$. The Lagrangian becomes:

$$L(\boldsymbol{p}_2, \lambda_2) = \boldsymbol{p}_2^T \boldsymbol{S} \boldsymbol{p}_2 - \lambda_2 (\boldsymbol{p}_2^T \boldsymbol{p}_2 - 1\,) - \varphi \boldsymbol{p}_1^T \boldsymbol{p}_2 \qquad (B.39)$$

and its partial derivative with respect to $\mathbf{p}_2$:

$$\frac{\partial L(\boldsymbol{p}_2^*, \lambda_2^*)}{\partial \boldsymbol{p}_1} = 2\boldsymbol{S}\boldsymbol{p}_2 - 2\lambda_2 \boldsymbol{p}_2 - \varphi \boldsymbol{p}_1 =$$
$$2\boldsymbol{p}_1^T \boldsymbol{S} \boldsymbol{p}_2 - 2\lambda_2 \boldsymbol{p}_1^T \boldsymbol{p}_1 - \varphi \boldsymbol{p}_1^T \boldsymbol{p}_1 =$$
$$2\boldsymbol{p}_2^T \boldsymbol{S} \boldsymbol{p}_1 - 2\lambda_2 \boldsymbol{p}_1^T \boldsymbol{p}_1 - \varphi \boldsymbol{p}_1^T \boldsymbol{p}_1 =$$
$$2\boldsymbol{p}_2^T \lambda_1 \boldsymbol{p}_1 - 2\lambda_2 \boldsymbol{p}_1^T \boldsymbol{p}_1 - \varphi \boldsymbol{p}_1^T \boldsymbol{p}_1 =$$
$$2\lambda_1 \boldsymbol{p}_2^T \boldsymbol{p}_1 - 2\lambda_2 \boldsymbol{p}_1^T \boldsymbol{p}_1 - \varphi \boldsymbol{p}_1^T \boldsymbol{p}_1 =$$
$$\varphi \boldsymbol{p}_1^T \boldsymbol{p}_1 = \varphi = \mathbf{0} \qquad (B.40)$$

This means that φ must equal to zero, and as a consequence:

$$2\boldsymbol{S}\boldsymbol{p}_2 - 2\lambda_2 \boldsymbol{p}_2 = \mathbf{0}$$
$$\boldsymbol{S}\boldsymbol{p}_2 = \lambda_2 \boldsymbol{p}_2 \qquad (B.41)$$

hence the second PC is another eigenvector of the estimated covariance matrix $\mathbf{S}$ with a corresponding eigenvalue λ_2. To maximize the target function $\boldsymbol{p}_2^T \boldsymbol{S} \boldsymbol{p}_2 = \lambda_2$, the second largest eigenvalue must be chosen. This process can be repeated to obtain 3N principal components, 3N being the dimension of the original coordinate space (the 3D Descartes coordinates of the N atoms). The 3N orthonormal eigenvectors of the estimated covariance matrix $\mathbf{S}$ form the set of PCs, they are ranked by the corresponding eigenvalues that directly equal to the variance along the given PC. In most cases, a low number of PCs can efficiently describe the characteristics encoded in the ensemble $\mathbf{R}$. The calculation of PCs can be achieved by the diagonalization of the estimated covariance matrix to obtain its eigenvectors and corresponding eigenvalues (more details are given about the diagonalization in the section on Normal Mode Analysis). Typically, only a subset of atoms is included in the PC calculations, e.g. backbone atoms or alpha carbons.

1.3.5. Quasiharmonic Mode Analysis

A closely related tool to PCA is quasiharmonic mode analysis. However, while PCA attempts to decompose system fluctuations into independent (pairwise linearly uncorrelated) motional modes, quasiharmonic analysis aims at analyzing a conformational trajectory by assuming an

underlying effective harmonic model. Nevertheless, the mathematics behind the two analysis tools are very similar. The motivation for quasiharmonic mode analysis is to connect simulation trajectories (e.g. from molecular dynamics) to normal modes.

In analogy with diagonalizing the estimated covariance matrix ($\mathbf{S}$) to obtain the PCs, in quasiharmonic mode analysis the mass-weighted estimated covariance is diagonalized: $\sqrt{\boldsymbol{M}}\boldsymbol{S}\sqrt{\boldsymbol{M}}$. Similarly to normal modes, where the potential energy is harmonically approximated, quasiharmonic analysis supposes that atomic fluctuations result from an underlying harmonic potential:

$$\widetilde{U}(\boldsymbol{r}) = \frac{1}{2}(\boldsymbol{r} - \tilde{\boldsymbol{r}}_0)^T \widetilde{\boldsymbol{H}}_0 (\boldsymbol{r} - \tilde{\boldsymbol{r}}_0) \qquad (B.42)$$

where $\widetilde{\boldsymbol{H}}_0$ is an effective Hessian matrix and $\tilde{\boldsymbol{r}}_0$ an effective equilibrium. The covariance matrix given by the normal mode analysis is defined as:

$$\Sigma = \frac{2}{k_B T} \boldsymbol{H}_0^{-1} \qquad (B.43)$$

where $\Sigma = \langle (r - r_0) \times (r - r_0) \rangle$ is the covariance matrix of the Cartesian variables, $\mathrm{k_B}$ the Boltzmann constant, T the absolute temperature, and $\boldsymbol{H}_0^{-1}$ the inverse of the Hessian matrix at a given energy minimum. To match the dynamics described by the trajectory to normal modes around a given minimum, quasiharmonic analysis approximates the equilibrium state as the average conformation over the trajectory, $\tilde{\boldsymbol{r}}_0 = \langle \boldsymbol{r} \rangle$, and estimates an effective Hessian matrix by:

$$\widetilde{\boldsymbol{H}}_0 = k_B T \boldsymbol{S}^{-1} \qquad (B.44)$$

where $\mathbf{S}$ is the estimated covariance matrix from the trajectory. Similarly to normal modes, to obtain the quasiharmonic modes, the mass-weighted effective Hessian matrix, $\widetilde{\boldsymbol{K}}_0 = \sqrt{\boldsymbol{M}}^{-1} \widetilde{\boldsymbol{H}}_0 \sqrt{\boldsymbol{M}}^{-1} = k_B T \sqrt{\boldsymbol{M}}^{-1} \boldsymbol{S}^{-1} \sqrt{\boldsymbol{M}}^{-1}$ needs to be diagonalized. Given an invertible matrix, the eigenvectors of the matrix are identical to the eigenvectors of its inverse matrix, and the corresponding eigenvalues of the inverse matrix are the inverse of the eigenvalues of the original matrix. Therefore, instead of $\widetilde{\boldsymbol{K}}_0$ the mass-weighted estimated covariance matrix, $\sqrt{\boldsymbol{M}}\boldsymbol{S}\sqrt{\boldsymbol{M}}$ can be diagonalized, and its eigenvectors form the set of quasiharmonic modes, and the matching quasiharmonic frequencies are given by the formula:

$$\omega_i = \sqrt{\frac{k_B T}{\lambda_i}}, i = 1 \dots 3N \qquad (B.45)$$

Similarly to PCA, a preceding superposition step is required to eliminate translational and rotational movements, and this will manifest in six quasizero frequencies. Furthermore, if geometrical constraints are present (e.g. on bond lengths), additional quasizero frequencies appear. Importantly, to properly match quasiharmonic modes and frequencies to normal modes, all atoms must be present in the calculations.

1.4. Modeling protein-ligand interactions

1.4.1. Molecular docking simulations

If the target protein structure is known (either resolved experimentally or modeled e.g. by homology), docking methods can be efficiently used to find binding sites and poses for ligands, and predict binding affinities. Docking on the one hand is computationally fast compared to atomistic simulations (e.g. molecular dynamics), however it is also less accurate as it approximates true physical energies with simplified energetics and solvation while performing a limited conformational search, mostly treating the receptor proteins as rigid. Due to its effectiveness, however, molecular docking is one of the most frequently used methods in structure-based drug design.

Molecular docking can be defined as an optimization problem where docking seeks to identify the best orientation of a ligand with respect to a receptor (in terms of free energy of the complex), taking into account the flexibility of the ligand (and possibly some residue sidechains of the receptor). The docking simulation evaluates rigid body transformations such as translations and rotations and internal changes including torsion angle rotations. Docking simulations require a search algorithm and a scoring function.

Search algorithms

An exhaustive search, in theory, would evaluate all possible conformations and respective orientations between the receptor and the ligand, however, it is computationally unfeasible. Docking algorithms can be classified as systematic, or stochastic (or even deterministic e.g. molecular dynamics).

Systematic search algorithms try to explore all the degrees of freedom in a molecule represented by the rotations of the bonds and angles and size of increments which is both computationally expensive and slow [332]. To avoid combinatorial explosion, the search can be done by building the ligand from different fragments: choosing a fragment as anchor, followed by sequential addition of combinations of the remaining fragments (e.g. the DOCK [333] or FlexX [334] software) [335, 336].

Stochastic search is randomized, it is based on making random changes to the ligand which are evaluated with a predefined probability function. The two major methods used in stochastic search are Monte Carlo (MC) and genetic algorithms (GA). MC starts from a random initial ligand configuration within the active site, scoring is done via a predefined scoring function. Small changes are made, and a new configuration is generated, which is always accepted if it outscores the previous configuration, otherwise it is only accepted with a probability according to e.g. the Boltzmann-based Metropolis criterion:

$$p_{accepting\ x_{k+1}} = \frac{p(x_{k+1})}{p(x_k)} = exp\left\{-\frac{U(x_{k+1})-U(x_k)}{RT}\right\}. \qquad (B.46)$$

The algorithm continues until the desired number of configurations is reached. MC is more robust in finding a global minimum than energy minimization as it can overcome energy barriers with a certain probability (MC is used in e.g. the MCDOCK [337], Vina [338], and ICM

[339] software). GA are based on the principle of biological evolution; the method is analogous to gene recombination and mutation in producing next generations. State variables, such as parameters describing translation, rotation, and the conformation of the ligand with respect to the receptor correspond to a gene in the algorithm and are grouped to form a chromosome. A GA population consisting of N sets is evaluated, individual chromosomes are scored based on a scoring function (the receptor-ligand complexes can be built from the state variables encoded in the chromosomes). Then, random pairs of the chromosomes are combined (mated), and new chromosomes are produced through reproduction (identical copy), crossovers (random exchange between the chromosome pair) and random mutations (perturbations introduced to the chromosome). Chromosomes of the new generation are then also evaluated, and a selection is made following some probabilistic selection rule (GA is used in e.g. the AutoDock [340] and GOLD [341] software).

Scoring functions

Docking algorithms generate different receptor-ligand configurations (different poses of the ligand within the binding site). The evaluation and ranking of such configurations are done using scoring functions. They are used for the identification of favorable binding modes, the prediction of binding affinity, and for virtual screening. Generally, such mathematical functions approximate the binding free-energy. Assuming thermodynamic equilibrium conditions for the protein-ligand complex formation, the binding free-energy is related to the binding constant (K_i) according to: $\Delta G^0 = -RTlnK_a$, where $K_a = \frac{k_{on}}{k_{off}}$, the ratio of the on-rate and of-rate constants [342]. Enthalpic and entropic contributions are both important, the following terms can be considered for the approximation of the binding free-energy:

$$\Delta G_{bind} = \Delta G_{solv} + \Delta G_{conf} + \Delta G_{int} + \Delta G_{rot} + \Delta G_{rigid\ bod\ mov} + \Delta G_{vib} \qquad (B.47)$$

where ΔG_{bind} is the binding free-energy, ΔG_{solv} the free-energy difference in interactions with the solvent due to ligand binding, ΔG_{conf} the effect of conformational changes in the receptor and the ligand, ΔG_{int} the contributions of receptor-ligand interactions, ΔG_{rot} the effect of freezing rotatable bonds (entropic contribution), $\Delta G_{rigid\ bod\ mov}$ the loss of degrees of freedom of rigid-body movements upon the association of two bodies to form a complex, and finally ΔG_{vib} the effect of changes in vibrational modes. Scoring functions use assumptions and simplifications to estimate the majority of the various terms [332]. There are three major classes of scoring functions, forcefield-based, empirical, and knowledge-based.

Forcefield-based scoring functions take the sum of the different potentials defined by a given forcefield (see Equation B.1 for general forcefield terms, e.g. the software AutoDock is based on the AMBER forcefield). Solvation is often handled using an implicit solvent medium; however, entropy is mostly not considered. Empirical scoring functions (which still may contain some force-field based terms) incorporate several (simplistic) energy terms, e.g. ionic interaction, hydrogen bonds, lipophilic contacts etc., the weighting of which in the overall score is based on experimental observations and are deduced from known receptor-ligand complexes (can be derived from regression models using experimentally determined binding

energies and resolved complex structures). Empirical scoring functions can be evaluated much faster than forcefield-based scoring functions making them an appealing choice (e.g. the Vina software uses empirical scoring). Knowledge-based scoring relies on statistical observations of ligand-receptor contacts extracted from experimentally resolved complexes (e.g. the DrugScore [342] software uses knowledge-based scoring). Knowledge-based scoring focuses on structural evaluation., prioritizing ligand configurations that are similar to existing experimental observations. Such scoring functions use pairwise atomic potentials. Assuming that experimental complexes represent the optimum placement of the ligand atoms relative to the receptor atoms constrained by covalent bonds, a large number of complexes can be evaluated to derive statistical potentials between protein and ligand atom-types.

1.4.2. Enrichment plots

To evaluate the capability of a docking algorithm using a given target protein structure of accurately modeling *in vitro* experiments, enrichment plots (a.k.a. Receiver Operator Characteristic (ROC) curve) can be created based on a dataset with known activity. A set of active and inactive (or decoy) molecules against a target are collected so that they occupy similar chemical spaces. The assumption is that active ligands have a lower energy score (have higher binding affinity) than inactive or decoy compounds. Enrichment plots demonstrate the trade-off between sensitivity and specificity, i.e. show the true positive rate (**sensitivity**, the ratio between the correctly predicted hits and all hits in the dataset, (P_T/P)) as the function of the false positive rate ($1 -$ **specificity**, the ratio between the incorrectly predicted hits and all the negatives in the dataset, (P_F/N)) evaluating different activity thresholds for the prediction classification. For each threshold, the true positive and false positive rates can be calculated, and a data point can be included in the enrichment curve.

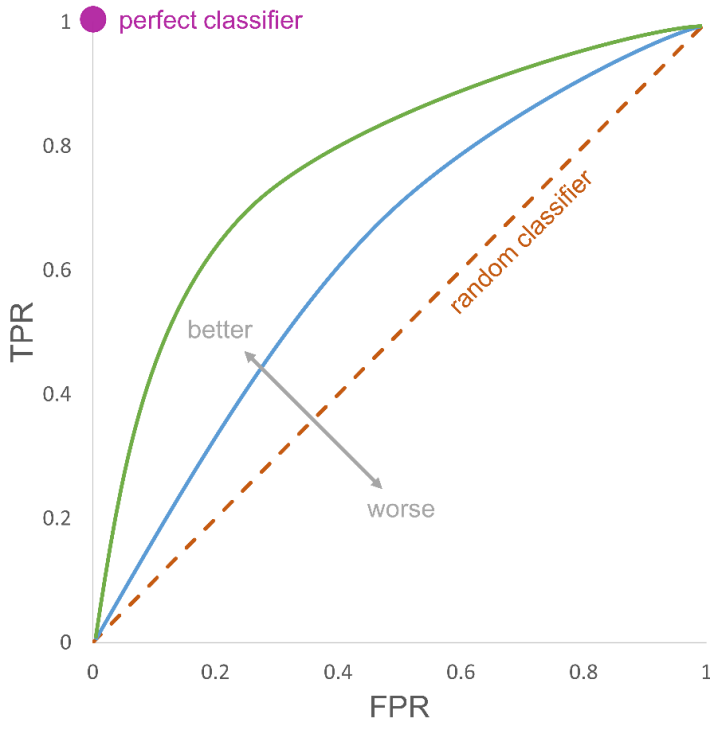

Figure B.5: Enrichment plots of some classifiers. TPR is the true positive rate (sensitivity), FPR is the false positive rate (1-specificity)

Docking results that give curves closer to the top-left corner indicate better performance, a random classifier is expected to give points lying along the diagonal. The overall performance can be described by a single parameter, the area under the curve (AUC) of the enrichment curve, which translates to the probability that a randomly chosen active compound is ranked higher than a randomly chosen inactive compound based on their docking score. However, a classifier with high AUC can sometimes score worse in a specific region of the enrichment curve than another classifier with lower AUC (their corresponding enrichment curves can intersect). The use of decoys instead of known inactive compounds necessitates to rather focus on the initial rise of the enrichment curve (the partial area under the curve (pAUC) allows to concentrate at a specific region of the curve and is usually calculated at early false positive rate values) [343].

A similar type of enrichment curves can be created given the compounds are ranked from best to worst docking scores. Starting from the best ranked compound, one-by-one adding compounds to a sample according to their ranking, the ratio of active compounds within the sample can be plotted as a function of the proportion of the sample size with respect to the total dataset size.

2. Machine learning modeling

Machine learning (ML) is a type of artificial intelligence (AI), ML algorithms can train models capable of predicting variables using some training data. Algorithms can be classified as supervised, unsupervised, or reinforcement learning. In supervised learning the training data is provided with desired outputs, the goal of the algorithms is to learn a general rule mapping the input variables to the output variable(s). Opposed to that, no labels are fed for the training data in unsupervised learning, the algorithm itself needs to find structures within the input data (e.g. the problem of conformational clustering). Reinforcement learning defines a clear goal and a prescribed set of rules for accomplishing that goal, by introducing positive rewards and negative punishments for different actions.

With the help of supervised learning, binary/multi-class classification (dividing data into two or more categories) or regression modeling (predicting continuous values) tasks can be addressed. Within the scope of the PhD thesis, two widely used supervised ML algorithms, random forest and support vector machines are discussed in more details for the binary classification of active and inactive compounds.

2.1. Random forest

Random forest (RF) is one of the most frequently used supervised ML algorithms because of its simplicity end easy interpretability. RF uses ensemble learning which combines many weak classifiers to provide an overall robust prediction of high accuracy and to avoid overfitting to the training data. RF consists of many decision trees, each is trained using only a subset of the original training dataset, and the output of RF models given an input is determined by majority vote of the individual decision trees (for classification, or by taking their average for regression problems).

The decision tree algorithm itself is a supervised ML technique that uses a set of rules for classification (or regression). Decision trees use recursive partitioning of the data into subsets. The graphical interpretation of a decision tree is a flowchart-like structure with if-else conditions. The tree consists of a root node, internal nodes, and leaf nodes. Starting from the root node, after branching at nodes according to different if conditions, the decision is reached in one of the leaf nodes of the tree.

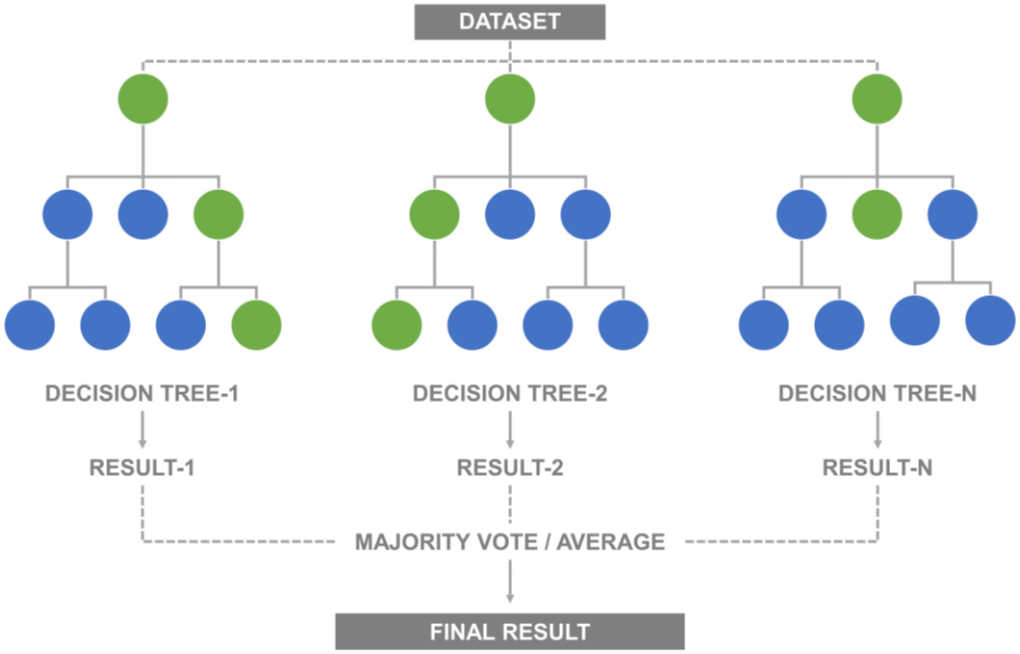

Figure B.6: Schematic flow-chart of the Random Forest algorithm.

Algorithms constructing decision trees rely on the concept of impurity to describe data heterogeneity with respect to the different features at a given node containing training data points possibly from different classes. The calculation of impurity of the different features at a given node helps to identify the feature to be used for the most efficient splitting of the data. Impurity functions have their maximum if the data points have a uniform distribution among the different classes, i.e. in practical terms each class is represented by the same number of training data points at the given node, and have their minimum if only training data points belonging to one single class are present. Entropy and Gini-index are widely-used impurity measures:

$$Entropy = -\sum_{i=1}^{n} p_i \, log(p_i) \qquad (B.48)$$

$$Gini\ index = 1 - \sum_{i=1}^{n} p_i^2 \qquad (B.49)$$

where n is the number of classes in the classification problem, and $\mathrm{p_i}$ denotes the estimated probability (relative frequency) of class i at a given node. Splitting according to a given feature has an impurity measure equal to the weighted sum of its child node impurities (the total Gini index of a split), where the weights are equal to the relative frequencies of each feature value at the node. The procedure of building a complete decision tree starts from the root node and consists of evaluating impurity measures for the different features at each node and selecting

the feature to use for data splitting that has the minimum impurity (maximum information gain). A node is considered as a leaf if its impurity reached its minimum (zero entropy or Gini index, subset is homogenous, all training data points belong to the same class).

RF performs data sampling by random replacement. Each decision tree receives a different dataset which has the same size as the original training dataset, however due to random replacements, some data points are missing whereas for others there will be duplicates (the so-called bootstrapped data). Furthermore, feature sampling can be done by using only a random subset of all features to build each decision tree. Alternatively, a different random subset of all features can be used at each potential splitting during the building of the different trees, increasing the degree of randomness in the algorithm to ensure that the resulting trees are different enough.

Hyperparameters of the RF algorithm include the number of decision trees in the forest and the number of features considered by each tree or at each splitting. Problems related to imbalanced training data (i.e. the number of data points is different for the different classes), can be addressed at the level of data sampling to tune the ratio between the classes in the bootstrapped training data sets provided to the trees.

RF also enables the determination of the individual feature importance in the classification. A frequently used feature importance form is the Gini importance, which is defined as the total decrease in node impurity measured by the Gini index weighted by the probability of reaching that node per tree, and averaged over all trees. The Gini importance is given by the formula:

$$Gini\ importance(X_m) = \frac{1}{N_T}\sum_T \sum_{t\in T: v(s)_t = X_m} p(t)\Delta i(s_t, t) \qquad (B.50)$$

where N_T is the number of trees, t denotes nodes within T where X_m is used for splitting the data, $p(t)$ is the probability of reaching node t within the tree T (estimated by the proportion of training samples reaching node t in tree T), and $v(s_t)$ is the feature used in the split of s_t. In the above equation $p(t)\Delta i(s_t, t)$ is the weighted impurity decrease, the weighted difference between the impurity of the given node and the weighted-sum of impurities of its child nodes; in case the Gini index is used for impurity measures, it results in the Gini importance.

2.2. Support-vector machine

Support-vector machine (SVM) is another very popular set of supervised ML algorithms used for classification and regression problems. In classification, SVM sorts data into two (or more) classes with the help of a hyperplane separating the different classes in some space. A hyperplane in a space $\mathbb{R}^n$ has a dimension of $n\text{-}1$, and has the form: $x_n = \alpha_{n-1}x_{n-1} + \alpha_{n-2}x_{n-2} + \cdots + \alpha_1 x_1 + b$,or alternatively expressed:

$$\boldsymbol{w}\cdot\boldsymbol{x} + b = 0 \qquad (B.51)$$

where $\boldsymbol{w} = (\alpha_1, \alpha_2, \ldots, \alpha_{n-2}, \alpha_{n-1}, -1)$ and $\boldsymbol{x} = (x_1, x_2, \ldots, x_{n-2}, x_{n-1}, x_n)$, and b is some constant. The rule of binary classification for classes separated by a given hyperplane in some space can be written for a data point s_i as:

$$h(\boldsymbol{s}_i) = \begin{cases} +1 \; if \quad \boldsymbol{w} \cdot \boldsymbol{s}_i + b \geq 0 \\ -1 \; if \quad \boldsymbol{w} \cdot \boldsymbol{s}_i + b < 0 \end{cases} \qquad (B.52)$$

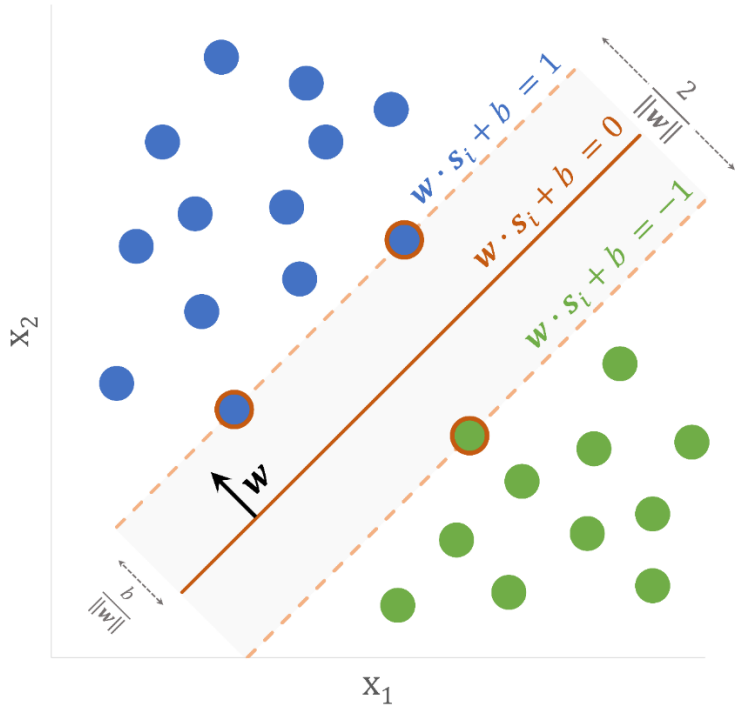

Figure B.7: SVM classification, the two classes are presented in green (-1) and blue (+1), the support vectors are shown with a dark orange contour.

The aim of SVM is to find the optimal hyperplane in some space which best separates the different classes. The metrics used in the hyperplane optimization problem is called the geometric margin (in case of perfect linear separation, it is the size of a buffer zone around the hyperplane where no training data points are present), which is the smallest distance among the training data points from the given hyperplane. The hyperplane optimization problem can be formulated as finding the hyperplane (described by a $\mathbf{w}$ normal vector and a constant $\mathbf{b}$) for which the geometric margin is maximum, i.e. given a training dataset $D = \{(\boldsymbol{s_i}, y_i) | \boldsymbol{s_i} \in \mathbb{R}^n, y_i \in \{-1,1\}\}_{i=1}^{m}$,

$$(\boldsymbol{w}, b)_{opt} = \underset{\boldsymbol{w},b}{argmax} \left(\underset{i=1\ldots m}{min} \frac{y_i}{\|\boldsymbol{w}\|} (\boldsymbol{w} \cdot \boldsymbol{s}_i + b) \right) \qquad (B.53)$$

where the training set consists of data points $1\ldots\mathrm{m}$, and each datapoint $\mathbf{s}_\mathrm{i}$ has a corresponding label $\mathbf{y}_\mathrm{i}$. On the training dataset, correctly classified datapoints will always have $y_i h(\boldsymbol{s_i}) > 0$ (both label and prediction are $+1$, or both are -1).

Scaling a normal vector does not change the defined hyperplane, the problem can be reformulated by scaling $\mathbf{w}$ so that $y_i(\boldsymbol{w} \cdot \boldsymbol{s}_i + b) = 1$ for the closes training data point to the form:

$$(\boldsymbol{w}, b)_{opt} = \underset{\boldsymbol{w},b}{argmax} \frac{1}{\|\boldsymbol{w}\|} \quad s.t. \quad f_i = y_i(\boldsymbol{w} \cdot \boldsymbol{s}_i + b) \geq 1, \forall i \qquad (B.54)$$

which is identical to the convex quadratic minimization problem:

$$(\boldsymbol{w}, b)_{opt} = \underset{\boldsymbol{w},b}{argmin}\frac{1}{2}\|\boldsymbol{w}\|^2 \quad s.t. \;\; y_i(\boldsymbol{w}\cdot\boldsymbol{s}_i + b) \geq 1, i = 1 \dots m \qquad (B.55)$$

The above equations do not tolerate outliers in the training dataset while optimizing the hyperplane separating the two classes (given the condition $f_i = y_i(\boldsymbol{w}\cdot\boldsymbol{s}_i + b) \geq 1, \forall i$, hard margin case) and does not work with non-perfectly linearly separable data. To overcome this hard constraint, the optimization problem can be adjusted by introducing slack variables (ζ_i). The slack variables can be considered as penalizing terms introduced upon the separation by a given plane and margin, data points which are correctly predicted and lie outside the margin have a penalty term of 0, data points which are predicted correctly but lie within the margin will have a penalty term between 0 and 1, and data points which are wrongly predicted will have penalty terms larger than 1,

$$\zeta_i := \begin{cases} 0, & if\ y_i(\boldsymbol{w}\cdot\boldsymbol{s}_i + b) \geq 1 \\ 1 - y_i(\boldsymbol{w}\cdot\boldsymbol{s}_i + b), & if\ y_i(\boldsymbol{w}\cdot\boldsymbol{s}_i + b) < 1 \end{cases} \qquad (B.56)$$

The optimization problem becomes:

$$(\boldsymbol{w}, b)_{opt} = \underset{\boldsymbol{w},b,\boldsymbol{\zeta}}{argmin}\frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m}\zeta_i \quad s.t. \;\; y_i(\boldsymbol{w}\cdot\boldsymbol{s}_i + b) \geq 1 - \zeta_i \; and \; \zeta_i \geq 0, \; i = 1 \dots m \qquad (B.57)$$

where $\boldsymbol{\zeta}$ contains the non-negative slack variables $(\zeta_1, \dots, \zeta_m)$, and $\mathbf{C}$ is the so-called regularization parameter, which determines the degree of tolerance towards misclassification (C close to zero means there is almost no penalty for misclassification, C approaching positive infinite will not tolerate misclassification). In geometrical terms, the smaller the regularization parameter, the wider the margin separating the different classes at the cost of higher misclassification on the training set.

The constrained optimization can be expressed with the help of a Lagrangian function $L(\boldsymbol{w}, b, \boldsymbol{\zeta}, \boldsymbol{\lambda}, \boldsymbol{\mu}) := f(\boldsymbol{w}, b, \boldsymbol{\zeta}) - \boldsymbol{\lambda}^T\boldsymbol{g}(\boldsymbol{w}, b, \boldsymbol{\zeta}) - \boldsymbol{\mu}^T\boldsymbol{h}(\boldsymbol{w}, b, \boldsymbol{\zeta}) = 0$, where $f(\boldsymbol{w}, b, \boldsymbol{\zeta}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m}\zeta_i$ is the function to be optimized, and $\boldsymbol{g_i}(\boldsymbol{w}, b, \boldsymbol{\zeta}) = y_i(\boldsymbol{w}\cdot\boldsymbol{s}_i + b) - 1 + \zeta_i, i = 1 \dots m$ arising from the constraints $y_i(\boldsymbol{w}\cdot\boldsymbol{s}_i + b) \geq 1 - \zeta_i$ and $\boldsymbol{h}_i(\boldsymbol{w}, b, \boldsymbol{\zeta}) = \zeta_i, i = 1 \dots m$ arising from the constraints $\zeta_i \geq 0$, are the Lagrangian constraint functions with corresponding non-negative Lagrangian weights ($\lambda_i and\ \mu_i, i = 1 \dots m$). The Lagrangian becomes:

$$L(\boldsymbol{w}, b, \boldsymbol{\zeta}, \boldsymbol{\lambda}, \boldsymbol{\mu}) := \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m}\zeta_i - \sum_{i=1}^{m}\lambda_i(y_i(\boldsymbol{w}\cdot\boldsymbol{s}_i + b) - 1 + \zeta_i) - \sum_{i=1}^{m}\mu_i\zeta_i \qquad (B.58)$$

and to find the local minimum of the target function $\boldsymbol{f(w, b, \zeta)}$ with respect to the constraints imposed by $\boldsymbol{g(w, b, \zeta)}$ and $\boldsymbol{h(w, b, \zeta)}$, a stationary point of the Lagrangian has to be identified. At a given stationary point $(\boldsymbol{w}^*, b^*, \boldsymbol{\zeta}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$, the gradient equals to zero, i.e. $\nabla L(\boldsymbol{w}^*, b^*, \boldsymbol{\zeta}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) = \mathbf{0}$. The following are true at a stationary point of the Lagrangian:

$\frac{\partial L(\boldsymbol{w}^*,b^*,\boldsymbol{\zeta}^*,\boldsymbol{\lambda}^*,\boldsymbol{\mu}^*)}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{i=1}^{m} \lambda_i\, y_i \boldsymbol{s}_i = \boldsymbol{0}$ *(B.59)*

$\frac{\partial L(\boldsymbol{w}^*,b^*,\boldsymbol{\zeta}^*,\boldsymbol{\lambda}^*,\boldsymbol{\mu}^*)}{\partial b} = -\sum_{i=1}^{m} \lambda_i\, y_i = 0$ *(B.60)*

$\frac{\partial L(\boldsymbol{w}^*,b^*,\boldsymbol{\zeta}^*,\boldsymbol{\lambda}^*,\boldsymbol{\mu}^*)}{\partial \zeta_i} = C - \lambda_i - \mu_i = 0,\ i = 1 \dots m$ *(B.61)*

and the following terms can be substituted in the original Lagrangian:

$\boldsymbol{w} = \sum_{i=1}^{m} \lambda_i\, y_i \boldsymbol{s}_i$ *(B.62)*

$\sum_{i=1}^{m} \lambda_i\, y_i = 0$ *(B.63)*

$\mu_i = C - \lambda_i$ *(B.64)*

and we obtain the so-called dual Lagrangian:

$L_d = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \lambda_i\lambda_j y_i y_j \boldsymbol{s}_i\boldsymbol{s}_j - C\sum_{i=1}^{m}\zeta_i - \sum_{i=1}^{m}\sum_{j=1}^{m} \lambda_i\lambda_j y_i y_j \boldsymbol{s}_i\boldsymbol{s}_j + \sum_{i=1}^{m}\lambda_i - \sum_{i=1}^{m}\lambda_i\zeta_i - C\sum_{i=1}^{m}\zeta_i + \sum_{i=1}^{m}\lambda_i\zeta_i$ *(B.65)*

and after simplification:

$$L_d(\boldsymbol{\lambda}) = \sum_{i=1}^{m}\lambda_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \lambda_i\lambda_j y_i y_j \boldsymbol{s}_i\boldsymbol{s}_j \qquad (B.66)$$

such that $\sum_{i=1}^{m} \lambda_i\, y_i = 0$, and as $\mu_i = C - \lambda_i$ (for the non-negative Lagrange multipliers $\lambda_i and\ \mu_i, i = 1 \dots m$), the following constraints also applies: $0 \le \lambda_i \le C,\ i = 1 \dots m$. Instead of solving the original minimization problem:

$$\min_{\boldsymbol{w},b,\boldsymbol{\zeta}} f(\boldsymbol{w},b,\boldsymbol{\zeta})\ such\ that\ g_i(\boldsymbol{w},b,\boldsymbol{\zeta}) \ge 0\ and\ h_i(\boldsymbol{w},b,\boldsymbol{\zeta}) \ge 0, i = 1 \dots m \qquad (B.67)$$

the Wolfe dual problem of maximization can be solved to obtain the same optimum (according to the Kuhn-Tucker theorem):

$$\max_{\boldsymbol{\lambda}} L_d(\boldsymbol{\lambda})\ such\ that\ \sum_{i=1}^{m} \lambda_i\, y_i = 0\ \ and\ \ 0 \le \lambda_i \le C,\ \ i = 1 \dots m \qquad (B.68)$$

which has the final formula:

$$\max_{\boldsymbol{\lambda}} \sum_{i=1}^{m}\lambda_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \lambda_i\lambda_j y_i y_j \boldsymbol{s}_i^T\boldsymbol{s}_j\ such\ that\ \sum_{i=1}^{m} \lambda_i\, y_i = 0\ \ and\ \ 0 \le \lambda_i \le C,\ \ i = 1 \dots m \qquad (B.69)$$

Support vectors are the training data points for which $y_i(\boldsymbol{w} \cdot \boldsymbol{s}^* + b) - 1 = 0$, data points that lie on the separating margin. Using the retrieved optimal Lagrangian multipliers, the optimal plane is given by $\boldsymbol{w} = \sum_{i=1}^{m} \lambda_i\, y_i \boldsymbol{s}_i$ and knowing that $y_i^2 = 1, \forall i$, the constant can be determined by $b = y_i - \boldsymbol{w} \cdot \boldsymbol{s}^*$, for any $\boldsymbol{s}^*$ support vector.

The introduction of slack variables enables to work on non-perfectly linearly separable data (mostly due to noise). However, SVM can also often be efficiently used for problems where the data cannot be linearly separated in a given space due to the non-linearity present in the characteristics of the data. By applying the so-called kernel trick, SVM can be used as a

nonlinear classifier. By transforming the original data to higher dimensions using a kernel function, the linearly non-separable classes may become linearly separable in the new feature-space. In the linear optimization problem using the dual Lagrangian, the optimization only requires the scalar products between the input samples: $\boldsymbol{s}_i^T \boldsymbol{s}_j$. SVM is powerful for separating data classes in higher dimension spaces due to this property, the individual vector projections are not required, only the scalar product of the different input samples in the higher dimension space needs to be assessed. A kernel function takes two points in the original space and returns their scalar product in the higher dimension space. Some of the most frequently used kernel functions include:

$k_{pol}(\boldsymbol{s}_i, \boldsymbol{s}_j) = (\boldsymbol{s}_i{}^T \boldsymbol{s}_j + r)^d$, the Polynomial kernel (with r=d=0 it is the linear case)

$k_{RBF}(\boldsymbol{s}_i, \boldsymbol{s}_j) = e^{-\frac{1}{2\sigma^2}\|\boldsymbol{s}_i - \boldsymbol{s}_j\|^2}$, the Gaussian radial basis kernel (also called RBF for radial basis function)

A given kernel function is related to the transformation $\Phi(\boldsymbol{x})$, such that $k(\boldsymbol{s}_i, \boldsymbol{s}_j) = \Phi(\boldsymbol{s_i}) \cdot \Phi(\boldsymbol{s_j})$. The hyper plane separating the classes and its corresponding normal-vector is also defined in the transformed space: $\boldsymbol{w} = \sum_{i=1}^{m} \lambda_i \, y_i \Phi(\boldsymbol{s_i})$. By the classification of new data the scalar product is also computed using the kernel function, without the projection of the data point in the higher dimension space: $\boldsymbol{w}^T \boldsymbol{x} = \sum_{i=1}^{m} \lambda_i \, y_i k(\boldsymbol{s}_i, \boldsymbol{x})$. At the optimum (where $\lambda_i(y_i(\boldsymbol{w} \cdot \boldsymbol{s}_i + b) - 1 + \zeta_i) = 0, \ \forall i$), all λ_i equal to zero for datapoints that are not support vectors (they do not lie on the separating margin).In other words, only support vectors can have corresponding non-zero Lagrange multipliers. Therefore, once the model is trained, SVM only requires a minimal number of operations for the binary classification.

The binary classification of SVM can be extended to multiclass problems by breaking down the multiclass problem into multiple binary classifications. Such approaches include one-versus-all or one-versus-one type of classifiers.

2.3. Assessment of classification models

The quality of prediction models can be evaluated on the training dataset. However, an external evaluation gives a better estimate of the classification quality for new data points. For the external validation of a prediction model, a set of datapoints with known labels can be used that are not present in the training of the given model.

2.3.1. Out-of-bag error

Thanks to the bootstrapping used by random forest training, not all datapoints are used for the growing of the different trees. Each tree will have a set of out-of-bag samples from the training data that were not used by the given tree for the training. For each data point in the training set, a classification prediction can be done by majority vote of the trees which did not use the given data point in their training and the overall quality of the random forest model can be evaluated based on the predictions for the training data points.

2.3.2. Cross validation

Cross validation is a more general way that can be used to evaluate prediction models (e.g. both for RF and SVM). Over several iterations, it uses different portions of the training data to train and test a model. A single round of cross validation involves the partitioning of the training data into two complementary subsets, training a model based on the first subset and evaluating the model on the second subset so that the samples used in the evaluation are not part of the training process. Most methods include multiple rounds of cross validation to reduce variability originating from the partitioning, the results of the different rounds are then combined to estimate the overall predictive performance.

One such algorithm is k-fold cross validation where the original training data is randomly partitioned into $\mathbf{k}$ equal sized subsamples. For each iteration, one of the k subsamples is kept for validation and the other $\mathbf{k\text{-}1}$ sets are used to train the model. Altogether $\mathbf{k}$ rounds of iterations are done, one for each subsample used as the validation set. The different results can be averaged then to obtain a single estimation.

2.3.3. Model validation measures

Given a set of data points for which both the label and the predicted class is known, different measures can be calculated to assess the predictive performance of the binary classification model. The data points can be divided into four groups based on their predicted and actual labels:

TP true positives that have a positive label and were correctly predicted as a hit

TN true negatives that have a negative label and were correctly predicted as a miss

FP false positives that have a negative label but were incorrectly predicted as a hit

FN false negatives that have a positive label but were incorrectly predicted as a miss

Sensitivity

The sensitivity describes the proportion of the correctly identified hits with respect to all the actual hits, in other words, how good the model is in finding hits among the data. It is given by the formula:

$$Sensitivity = \frac{TP}{TP+FN} \qquad (B.70)$$

Specificity

Similarly, the specificity describes the proportion of the correctly identified misses with respect to all the actual misses, and in other words, how good the model is in finding misses among the data. It is given by the formula:

$$Specificity = \frac{TN}{TN+FP} \qquad (B.71)$$

Accuracy

To obtain an overall assessment of the model, accuracy can be calculated. The accuracy describes the proportion of correctly identified labels among all the data points, either by correctly finding a hit or a miss. It is given by the formula:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (B.72)$$

Balanced Accuracy

In case the dataset is not well balanced (i.e. the classes are not equally represented), accuracy can be misleading. Instead, the Balanced Accuracy (BA) can be calculated which is defined as the mean of the Sensitivity and Specificity:

$$BA = \frac{Sensitivity+Specificity}{2} \qquad (B.73)$$

Matthews Correlation Coefficient

Another effective solution to overcome class imbalance issues in the evaluation of binary classification is the Matthews Correlation Coefficient (MCC), which ranges between -1 (worst performance) and $+1$ (best performance), and is given by the formula:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}} \qquad (B.74)$$

II. Objectives

„Az iránytű pontosan megmutatja, merre van észak. Azonban mit sem tud a mocsarakról és a szakadékokról. Ha pedig az ember elsüllyed vagy lezuhan, mi értelme volt, hogy tudta, merre van észak?”

“The compass shows you exactly which way is north. However, it knows nothing about swamps and ravines. And if you sink or fall, what was the point of knowing which way is north?”

Náray Tamás

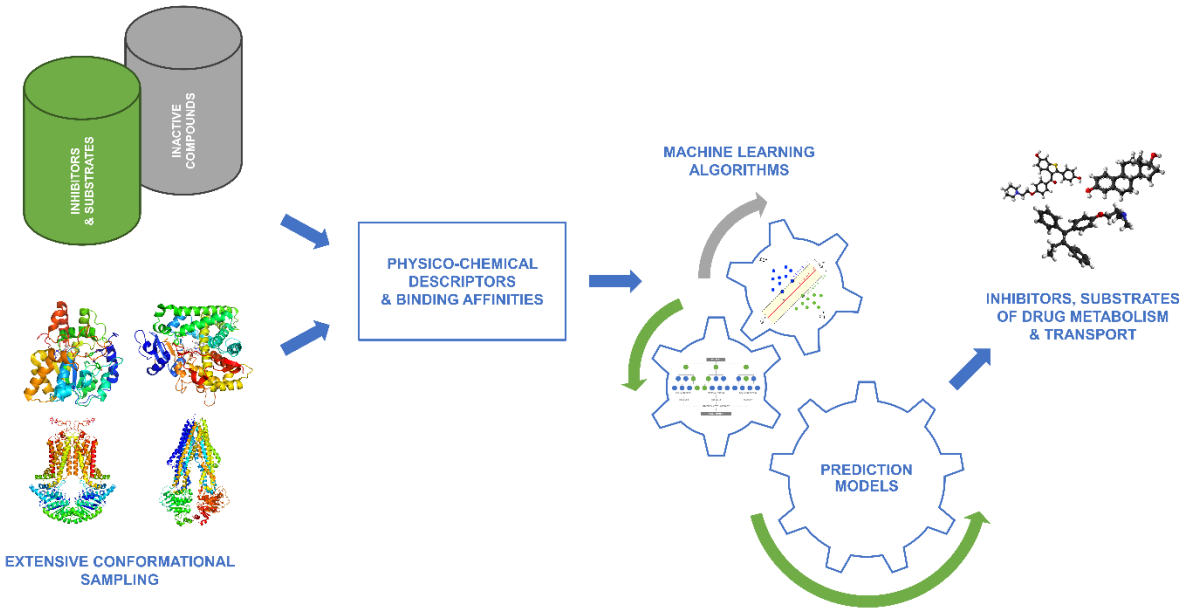

Figure.1: Flowchart of the objective: creation of machine learning models combining structure- and ligand-based information for the identification of substrates and inhibitors of drug metabolizing enzymes and drug transporters.

Drug elimination through metabolism and excretion is a complex process that is governed by several metabolizing enzymes and membrane transporters as discussed in the Introduction. Interactions with DMEs and drug transporters are of great importance for safety treatments. The ADME-Tox profile of substrate drugs is influenced by such enzymes and transporters, whereas the modulation of their activities also influences the pharmacokinetics of other xenobiotic and endogenous compounds, and can evoke severe adverse drug reactions. One of the major reasons for drug candidate failure concerns problems related to their pharmacokinetics and pharmacodynamics discovered during clinical trials. *In silico* prediction of such interactions can help reduce the rate of drug candidate failure at an early drug development stage thereby also reducing associated costs, and can help decreasing the number of animal tests.

Studies on drug metabolism have predominantly prioritized phase I DMEs, in particular cytochrome P450 (CYP) due to their involvement in toxic events. However, phase II DMEs also play an essential role in drug metabolism and are critical for drug administration safety and the prediction of xenobiotics toxicity. Therefore, in my PhD work I have focused on two major phase II DMEs, namely the sulfotransferase SULT1A1 and the UDP glucuronosyl transferase UGT1A1, as well as on an essential drug efflux transporter, the breast cancer resistance protein (ABCG2/BCRP).

The first step towards understanding the complex mechanisms of the different ligand-protein interactions is through revealing the dynamical behavior of the given protein and elucidating its functional movements. The molecular mechanisms guiding the recognition of the diverse substrates and inhibitors of SULT1A1 are related to the conformational flexibility of the enzyme which needs to be elucidated. Little is known about the dynamic behavior of the

coordination between the cofactor- and the substrate-binding domains of UGT1A1 related to the catalyzed enzymatic reaction. Similarly, very little structural information is available for ABCG2 about the large conformational transitions that are the driving forces for the substrate translocation of the transporter. Therefore, the primary goal of the current PhD work has been to decipher the functional movements encoded in the above-listed proteins with the help of different simulation approaches. Furthermore, my objective has been to identify the effects of ligand binding on the dynamics and the functional collective movements of the different proteins.

Using the conformational ensembles generated by the different simulation approaches and the dynamic information extracted from them, the second main objective of my PhD work has been to elucidate the interactions of known active compounds, substrates and inhibitors, with the target proteins, and investigate their binding affinities and binding modes within the substrate binding cavities. Both SULT1A1 and UGT1A1 possess a wide ligand specificity range and for future predictions of inhibition, it is crucial to unravel the molecular mechanisms accounting for their promiscuity. In addition, ABCG2 must guide its substrates along a path to reach the extracellular space, and I aimed at identifying the mechanism underlying the substrate translocation, and reveal substrate interactions in the second substrate-binding cavity.

Ultimately, the final objective of my thesis work has been to build *in silico* predictive classification models capable of distinguishing between active and inactive compounds by integrating protein structure-based and machine learning methodologies. In the framework of the present thesis the construction and application of such models is demonstrated for the enzyme UGT1A1.

III. Results

„Te jól tudod, a költő sose lódit: az igazat mondd, ne csak a valódit,
a fényt, amelytől világlik agyunk, hisz egymás nélkül sötétben vagyunk.”

“You know this well: the poet never lies, The real is not enough; through its disguise
Tell us the truth which fills the mind with light, Because, without each other, all is night.”

József Attila (translated by Vernon Watkins)

A. SULT1A1

Sulfation is one of the major conjugating pathways responsible for the detoxification and subsequent elimination of xenobiotic and endogenous small molecules. Cytosolic sulfation reactions are catalyzed by sulfotransferases (SULTs), the expression of the different SULT enzymes occurs almost in every organ in humans. The isoform SULT1A1 (also known as the thermostable phenol sulfotransferase) has the broadest substrate specificity within the SULT superfamily, and it displays an extensive tissue distribution.

The following chapter on the substrate binding mechanism of SULT1A1 investigates the functional movements of the enzyme in its monomer form bound to the enzymatically active cofactor (PAPS) and its interactions with a set of known SULT1A1 ligands.

Different simulation approaches are used to thoroughly explore the conformational states of the enzyme, with special emphasis on the substrate-binding pocket and the gate that governs the access to the catalytic site formed by three mobile loop regions. Classical MD and an enhanced MD simulation tool, which incorporates collective movements described by low frequency normal modes (Molecular Dynamics with excited Normal Modes), are performed to compare their corresponding capacity to explore large functional changes in the enzyme and to generate conformational ensembles with large diversity for consequent ensemble docking simulations.

The binding modes of the numerous ligands are evaluated in the conformations of the two ensembles generated by classical and enhanced MD simulations to identify binding modes and corresponding protein conformations that form favorable complexes. The affinity of the generated conformations towards different sized ligands is further tested with a comprehensive comparison of the binding of a medium sized and a large substrate, estradiol and fulvestrant. The stability of the predicted binding modes is investigated using classical MD simulations on the complexes. The ultimate goal of selecting a smaller set of diverse enzyme conformations that can accommodate the structurally very different ligands with high affinity is to use information on protein-ligand interactions for activity prediction of drugs, drug candidate molecules, and other xenobiotics.

Insights into the substrate binding mechanism of SULT1A1 through Molecular Dynamics with excited Normal Modes simulations

Balint Dudas[1,2,#], Daniel Toth[3,#], David Perahia[2], Arnaud B. Nicot[4], Erika Balog[3,*], Maria. A. Miteva[1,*]

[1]Inserm U1268 MCTR, CiTCoM UMR 8038 CNRS - University of Paris, Pharmacy Faculty of Paris, France
[2]Laboratoire de Biologie et Pharmacologie Appliquée, Ecole Normale Supérieure Paris-Saclay, UMR 8113, CNRS, Gif-sur-Yvette, France
[3]Department of Biophysics and Radiation Biology, Semmelweis University, Budapest, Hungary
[4]Inserm, Université de Nantes, Centre de Recherche en Transplantation et Immunologie, UMR 1064, ITUN, F-44000 Nantes, France

[#]1st coauthors
[*]corresponding authors: maria.mitev@inserm.fr, balog.erika@med.semmelweis-univ.hu

Published in Scientific Reports on 2021 Jun 23
doi: 10.1038/s41598-021-92480-w

Abstract

Sulfotransferases (SULTs) are phase II drug-metabolizing enzymes catalyzing the sulfoconjugation from the co-factor 3′-Phosphoadenosine 5′-Phosphosulfate (PAPS) to a substrate. It has been previously suggested that a considerable shift of SULT structure caused by PAPS binding could control the capability of SULT to bind large substrates. We employed molecular dynamics (MD) simulations and the recently developed approach of MD with excited Normal Modes (MDeNM) to elucidate molecular mechanisms guiding the recognition of diverse substrates and inhibitors by SULT1A1. MDeNM allowed exploring an extended conformational space of PAPS-bound SULT1A1, which has not been achieved up to now by using classical MD. The generated ensembles combined with docking of 132 SULT1A1 ligands shed new light on substrate and inhibitor binding mechanisms. Unexpectedly, our simulations and analyses on binding of the substrates estradiol and fulvestrant demonstrated that large conformational changes of the PAPS-bound SULT1A1 could occur independently on the co-factor movements that could be sufficient to accommodate large substrates as fulvestrant. Such structural displacements detected by the MDeNM simulations in the presence of the co-factor suggest that a wider range of drugs could be recognized by PAPS-bound SULT1A1 and highlight the utility of including MDeNM in protein-ligand interactions studies where major rearrangements are expected.

1. Introduction

Drug metabolizing enzymes (DMEs) play a key role in the metabolism of endogenous molecules and the detoxification of xenobiotics and drugs (Sun and Scott, 2010, Testa et al., 2012, Shimada, 2006). Phase I metabolism includes hydrolysis, reduction, and oxidation reactions, while Phase II comprises mainly glucuronidation, sulfation, methylation, and glutathione conjugation reactions (Pratt and Taylor, 1990). Sulfotransferases (SULTs) and UDP-glucuronosyltransferases are responsible for most of the Phase II reactions in the body, with the conjugation of approximately 40 % of all drugs (Tibbs et al., 2015). SULTs catalyze the sulfoconjugation from the co-factor 3′-Phosphoadenosine 5′-Phosphosulfate (PAPS) to a substrate hydroxyl or amino group (Dong et al., 2012, Gamage et al., 2006, Bojarova and Williams, 2008, Chapman et al., 2004). DMEs are highly promiscuous, and the relations of their structural plasticity and substrate promiscuity have been widely studied (Tibbs et al., 2015, Martiny et al., 2013, Allali-Hassani et al., 2007, Sun and Scott, 2010, Louet et al., 2018, Martiny and Miteva, 2013, Dong et al., 2012, Gamage et al., 2005, Guengerich et al., 2019, Srejber et al., 2018, Martiny et al., 2015). SULTs show a broad substrate range, metabolizing a wide variety of endogenous compounds like steroids and polysaccharide chains, and participating in the bioactivation of a number of xenobiotics and drugs (Gamage et al., 2006).

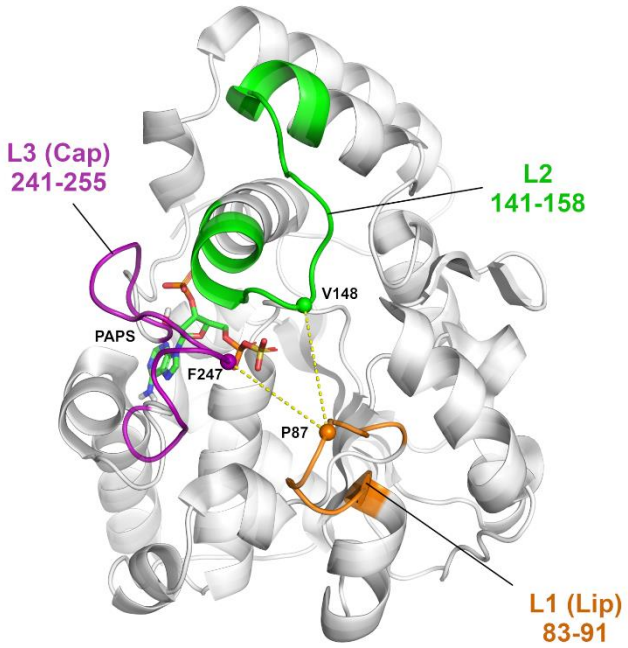

*Figure A.1: Crystal structure of SULT1A1*1, PDB ID: 4GRA. PAP of 4GRA was replaced by PAPS which was retrieved from the structure of SULT1E1 (PDB ID: 1HY3 containing PAPS) and inserted on the same position as that of the nucleotide in 4GRA; it is shown in sticks.*

The molecular bases of substrate specificity, selectivity, and inhibition across different SULT isoforms, have been previously addressed (Dajani et al., 1998, Lee et al., 2003, Wang et al., 2017, Cook et al., 2016, Cook et al., 2015a, Cook et al., 2013b, Cook et al., 2013a, Zhu et al., 2019, Martiny et al., 2013, Rakers et al., 2016, Allali-Hassani et al., 2007). These specificities have proven to be complex as relationships between SULTs pocket characteristics and substrate shape have shown not to be direct, since pocket shape and size have the potential to

fluctuate upon substrate binding (Cook et al., 2015a). Structural displacements can alter the substrate-binding profiles, thus guide enzyme-substrate interactions. It has been demonstrated that the binding of PAPS causes a considerable shift in the PAPS binding domain of SULT, moving a strongly conserved 30-residue active site "Cap", which covers both the nucleotide co-factor and the substrate-binding site, towards "closure" (Figure A.1). This large movement, called "gating", was suggested to participate in an isomerization equilibrium rate controlling the potential of SULT to bind larger substrates (Cook et al., 2013a, Zhu et al., 2019, Cook et al., 2015a, Wang et al., 2016). However, sulfonation data for SULT2A1/raloxifene strikingly revealed that the enzyme was still capable of turnover (Cook et al., 2012) with approximately 5% of SULT2A1 remaining in its open state even at saturating levels of PAPS (Cook et al., 2013a, Tibbs et al., 2015). These data demonstrate that the gating mechanism may not be dependent only on the co-factor binding and that the mechanism of substrate recognition and selectivity should be further elucidated.

Molecular Dynamics (MD) simulations (Mortier et al., 2015) and more recent Normal Mode Analysis approaches (Moroy et al., 2015, Pantaleao et al., 2018) have become major techniques in the arsenal of tools developed to investigate the mode of action of bioactive molecules. A recent approach called MDeNM (Molecular Dynamics with excited Normal Modes) has recently been developed using low-frequency normal mode directions in MD simulations (Costa et al., 2015). This approach considers many different linear combinations of NM vectors, each used in an independent MD simulation in which the corresponding collective motion is kinetically excited. Therefore, a wide variety of large movements can be promoted straightforwardly, which would be costly by standard MD simulations. So far MDeNM has been used successfully to study large functional movements in several biological systems (Fagnen et al., 2020, Gomes et al., 2020, Dudas et al., 2020, Dudas et al., 2021).

In this study, we focused on SULT1A1 (Gamage et al., 2003), which is the most abundant SULT in the human liver. The SULT1A1 enzyme is widely distributed throughout the body, with a high abundance in organs such as the liver, lung, platelets, kidney, and gastrointestinal tissues (Hempel et al., 2007). Human SULT1A1 exhibits a broad substrate range with specificity for small phenolic compounds, including the drugs acetaminophen and minoxidil, and pro-carcinogens such as N-hydroxy-aromatic and heterocyclicaryl amines (Gamage et al., 2006). To elucidate the gating mechanism guiding the recognition of diverse substrates, in this work, we employed the recently developed original approach of MDeNM (Costa et al., 2015) to explore an extended conformational space of the PAPS-bound SULT1A1 (SULT1A1/PAPS), which has not been achieved up to now by using classical MD simulations (Cook et al., 2016, Cook et al., 2015a, Cook et al., 2013b, Cook et al., 2013a, Zhu et al., 2019). The investigation of the generated ensembles combined with the docking of 132 SULT1A1 substrates and inhibitors shed new light on the substrate recognition and inhibitor binding mechanisms. The performed MD and MDeNM simulations of SULT1A1/PAPS as well as MD and docking simulations with the substrates estradiol and fulvestrant, previously suggested to undergo different binding mechanisms (Cook et al., 2013a), demonstrated that large conformational changes of the PAPS-

bound SULT1A1 can occur. Such conformational changes could be sufficient to accommodate large substrates, e.g. fulvestrant, independently of the co-factor movements. Indeed, such structural displacements were successfully detected by the MDeNM simulations and suggest that a wider range of drugs could be recognized by PAPS-bound SULT1A1.

2. Results and Discussion

MDeNM simulations enable an extended sampling of the conformational space by running multiple short MD simulations during which motions described by a subset of low-frequency Normal Modes are kinetically excited (Costa et al., 2015). Thus, MDeNM simulations of SULT1A1/PAPS would allow detecting "open"-like conformations of SULT1A1, previously generated by MD simulations performed in the absence of its bound co-factor PAP(S) (Wang et al., 2017, Cook et al., 2013b, Zhu et al., 2019, Cook et al., 2013a). PAPS was included in the co-factor binding site of SULT1A1 (see Materials and Methods for details) and maintained bound to SULT1A1 in all our simulations, since it was demonstrated that the co-factor is required for the correct folding of the substrate-binding site. Previous crystal structures of co-factor-free SULT have shown significant unfolding of the key loop L3 (FIGURE A.1) covering the co-factor and substrate binding sites (Allali-Hassani et al., 2007). Here, the conformational sampling of SULT1A1/PAPS was performed by running: i) three 200 ns long MD simulations with different initial velocity distributions and ii) the previously developed efficient simulation method, MDeNM (Costa et al., 2015) - with 240 replicas - that combines Normal Mode Analysis (NMA) and Molecular Dynamics. MDeNM performs several simultaneous MD simulations during which motions along different randomized linear combinations of the most relevant low-frequency normal modes are promoted in the form of a velocity increment. The starting crystallographic coordinates for SULT1A1*1 were taken from the Protein Data Bank (Berman et al., 2000), PDB ID 4GRA (Cook et al., 2013a), containing the co-factor PAP. We replaced PAP with PAPS required for the sulfonation catalytic activity of SULT1A1. No substrates/inhibitors were included in the MD and MDeNM simulations to avoid possible ligand-induced biases of the SULT1A1/PAPS structure. The total simulation time was 600 ns for the MD and 48 ns for the MDeNM simulations (see the Methods for details).

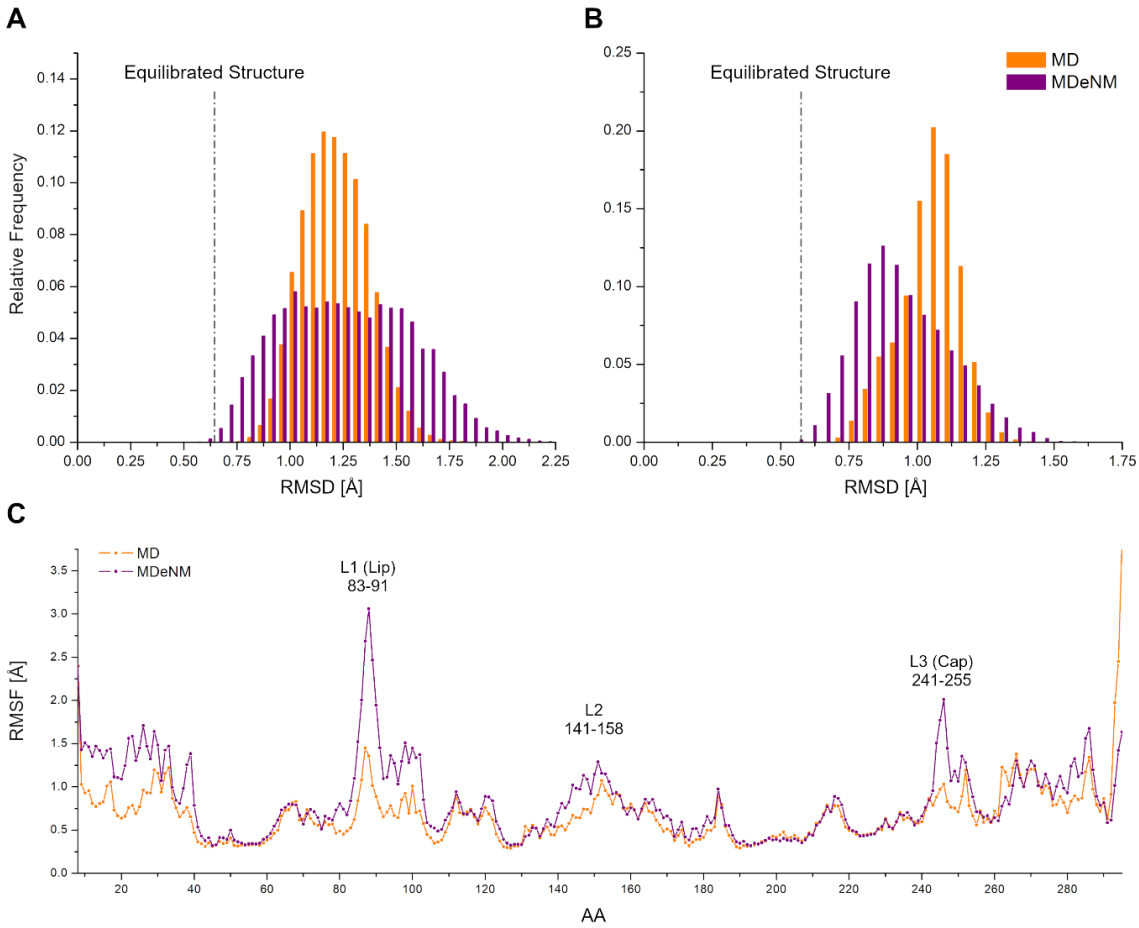

Figure A.2: The Root Mean Square Deviation (RMSD) with respect to the crystal structure PDB ID: 4GRA of the MD (in orange) and MDeNM (in purple) generated structures of SULT1A in the presence of the PAPS. A. calculated on the binding pocket heavy atoms (the residues of the binding pocket are given in the Supporting Information and B. on the backbone of the whole protein; C. Root Mean Square Fluctuation (RMSF) of C_α atoms per amino acid residue (AA) in the MD (orange) and MDeNM (purple) conformational ensembles.

2.1. Structural analysis of the MD and MDeNM generated conformational ensembles

In order to identify similarities and differences in the conformational ensembles generated by the MD and MDeNM simulations, the Root Mean Square Deviation (RMSD) of the binding pocket (its residues are listed in the SI) was calculated with respect to the crystal structure (FIGURE A.2A). The MD conformations distribution covers an RMSD range between 0.75 Å and 1.75 Å with a clear peak around 1.2 Å with respect to the binding pocket of the starting crystal structure. The MDeNM conformations distribution of the binding pocket is more dispersed, even reaching conformations with a binding pocket deviating up to 2.25 Å from the crystal structure. Particularly, the region corresponding to RMSD values above 1.45 Å is more populated by MDeNM. The RMSD distribution of the whole protein backbone, calculated for the MDeNM conformations, showed a peak closer to the starting structure than that of the conformations generated by MD (FIGURE A.2B). However, the MDeNM simulations also generated conformations that deviate more from the crystal structure than those observed by MD, up to 1.5 Å. Larger deviations in the case of our MDeNM simulations originate from significant global movements of the protein. Larger deviations hence imply a more exhaustive

conformational sampling, especially for the binding pocket. Our results suggest that MDeNM performed a more exhaustive conformational sampling of the SULT1A1 binding pocket while maintaining the protein's overall structure closer to the starting structure.

The Root Mean Square Fluctuation (RMSF) of the C α atoms was calculated to identify flexible protein regions of functional importance (FIGURE A.2C). Significant differences are visible at the gate (formed by loops L1, L2, and L3) of the binding pocket of SULT1A1 between conformational ensembles generated by the two methods. MDeNM particularly magnifies motions related to L1 (residues 83-91) and L3 (residues 241-255) and moderately related to L2 (residues 141-158). The fluctuation amplitude of the residues P87 and E246 at the tip of L1 and L3, respectively, is double in the case of MDeNM, indicating that MDeNM explores the gating motions to a greater extent. The Cap L3 has been suggested to play a key role in the gating mechanism of SULT1A1 (Cook et al., 2013a) and SULT2A1 (Cook et al., 2012, Zhu et al., 2019), fluctuating between a closed and an open isomer depending on the nucleotide-binding. L1 (also known as the "Lip" (Cook et al., 2015b)) demonstrates a larger fluctuation than L3 by both MD and MDeNM, implying its involvement in the gating mechanism. Obviously, here the presence of PAPS stabilizes L3, which is known to be completely unfolded in the absence of bound co-factor (Allali-Hassani et al., 2007). Although the RMSF of both MD and MDeNM demonstrates the flexibility of L1, L2 and L3, larger movements of L1 and L3 are observed by the MDeNM simulations than by the MD.

The C α atoms of residues P87, V148, and F247 representing each loop at their tip were selected to follow the relative motions and the gating mechanism of the three loops at the entrance to the binding pocket. Two distances, namely d(L1,L2) and d(L1,L3), were monitored corresponding to the distances d(P87Cα,V148Cα) and d(P87Cα,F247Cα) (see FIGURE A.1). The distribution of all generated conformations along these two distances can be seen in Fig. 3.

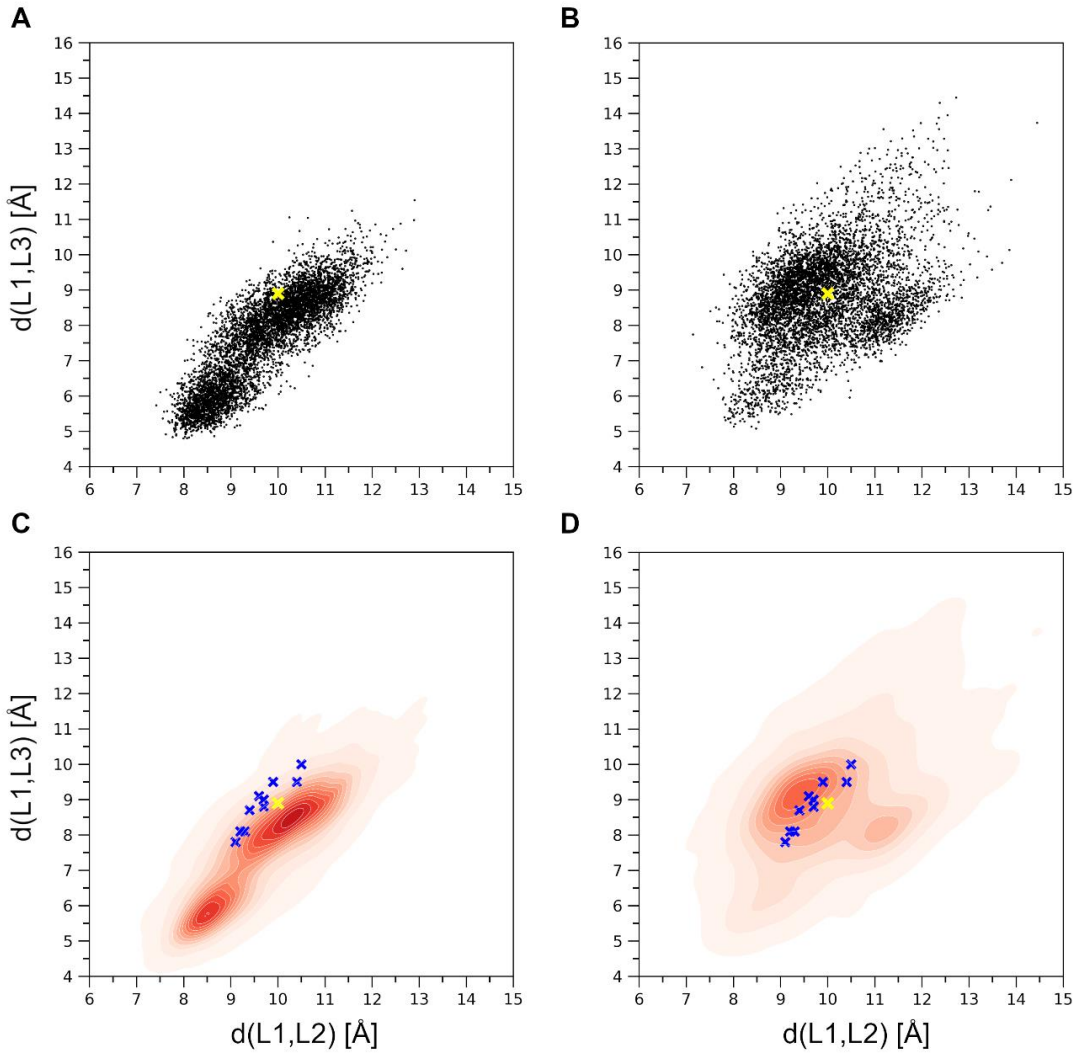

*Figure A.3: **(A)** Distribution of the d(L1,L2) and d(L1,L3) distances corresponding to the MD generated conformations; **(B)** Distribution corresponding to the MDeNM generated conformations; **(C)** Corresponding normalized distribution densities for the MD conformations and **(D)** for the MDeNM conformations. The available crystal structures (denoted by blue 'x'-es) are plotted into the maps of (C) and (D); The location of the crystal structure (4GRA.pdb) is shown in yellow 'x'.*

Conformations reached by MD (FIGURE A.3A) exhibit a strong positive correlation (the correlation being 0.86) between d(L1,L2) and d(L1,L3), restricting thus the opening of the gate to occur along both distances at the same time. Interestingly, there are two dense regions in the MD conformations distribution, one lying close to the initial conformation (4GRA.pdb) denoted by yellow 'x', and another one corresponding to a more closed state. MD did not explore conformations having d(L1,L3) greater than 11.5 Å. The MDeNM distribution (FIGURE A.3B) is more widely spread and less restricted by the d(L1,L2) and d(L1,L3) correlation (the correlation being 0.40). MDeNM reaches conformations with the d(L1,L3) distance 3 Å beyond MD, up to 14.5 Å, corresponding to more widely open conformations, whereas MD maps densely populated tightly closed states. Both MD and MDeNM covered and reached far beyond the gate positions of L1, L2, and L3 - both in the closing and the opening directions - of experimentally available conformations (the apo-forms of SULT1A1*1 and SULT1A1*2 without

bound ligand PDB IDs 4GRA and 3U3J, respectively; the holo-forms of SULT1A1*2 with bound ligand PDB IDs: 1LS6, 2D06, 3U3M, 3U3O, 3U3R, 3U3K; and two ancestral variant b9 PDB IDs: 3QVU, 3QVV) (see FIGURE A.3C and D), which exhibit a very conserved overall structure with slight differences in their gate opening, the RMSD difference calculated on the Cα-s of the whole protein between any two experimental structures being less than 0.51 Å. The observed correlation between d(L1,L2) and d(L1,L3) in addition to the high RMSF values at L1, and visual inspection further confirmed the significant movements of L1 by the opening-closing of the gate, underlining the functional importance of L1 by SULT1A1 as proposed in the work of Rakers et al. for SULT 1E1 (Rakers et al., 2016).

2.2. Ensemble docking of SULT1A1 substrates and inhibitors

The docking of 132 previously known substrates or inhibitors (collected in our previous work (Martiny et al., 2013) and (Paitz and Bowden, 2013, Cook et al., 2012)) was performed into the binding pocket of the conformations collected by MD and MDeNM to gain insight into the mechanism of SULT1A1-ligand interactions.

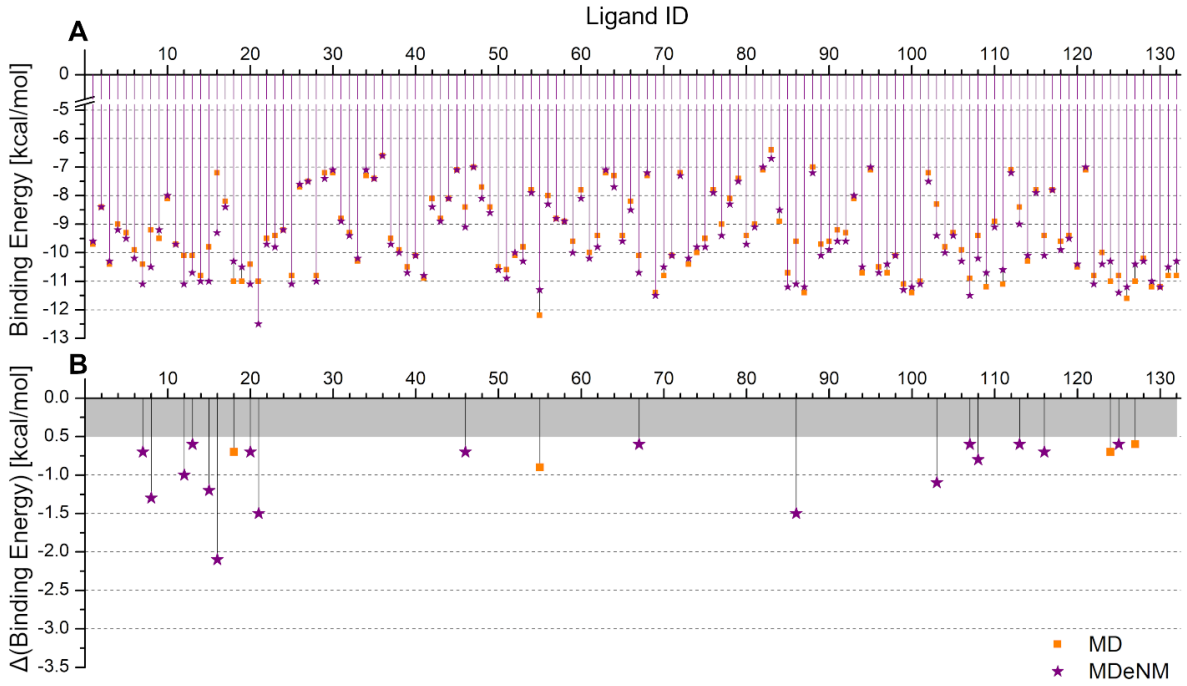

*Figure A.4: **(A)** The lowest binding energy (BE) per ligand resulting from the docking of the set of 132 known ligands to the ensemble of representative structures after clustering of SULT1A1/PAPS obtained from the MD (denoted by orange squares) and MDeNM (denoted by purple stars) simulations. **(B)** Differences between the best BEs retained for the MD and MDeNM conformations; for the better visualization, only differences larger than 0.5 kcal/mol are indicated.*

First, both the MD and MDeNM generated conformations were clustered based on their binding pocket (see the list of residues in SI) to obtain a smaller, representative set of conformations to be used for the docking of all the ligands (see Methods for details). We performed docking on 94 MD and 86 MDeNM centroid SULT1A1/PAPS conformations. For each docking simulation, the best Binding Energy (BE) was retained. As different ligands can be

accommodated in different binding pocket arrangements, for each ligand the best BE over the set of conformations have been taken; the results are plotted in FIGURE A.4A. Many ligands expressed similar docking behavior into the MD and MDeNM set of conformations, the average of the BEs over all the ligands being -9.33 kcal/mol and -9.49 kcal/mol, while the worst BE being -6.4 kcal/mol and -6.6 kcal/mol for MD and MDeNM, respectively. For some ligands, however, considerable differences were observed (see FIGURE A.4B). Most of these compounds (17 out of the 21) showing a difference greater than 0.5 kcal/mol) showed a more favorable BE when docked to the MDeNM set of conformations, demonstrating the benefit of including the MDeNM simulations in addition to MD. We compared the predicted and experimental binding energies reported in the literature for several SULT1A1 ligands (see in SI Table S1 and Figure S1). Predicted binding energies (BE) were calculated by averaging the best scored Autodock Vina energies in the best 10 MD conformations and in the best 10 MDeNM conformations. The comparison between the experimental free binding energies and the scores calculated with Autodock Vina can be only qualitative, yet a correlation with a correlation coefficient R^2 of 0.56 was obtained. Interestingly, the Vina scores distinguished between the low-affinity substrate p-nitrophenol with experimental BE of -5.76 kcal/mol (Lu et al., 2009) and the other higher affinity ligands.

To characterize the binding poses of the substrates, a criterion of having their acceptor hydroxyl or primary amino functional group in the vicinity of the sulfate group of the co-factor PAPS and the catalytic residue H108 was imposed. Docking positions and the corresponding BE of substrates with the d(O,S) or d(N,S) distance greater than 5 Å were rejected, and the best BE satisfying the distance criterion was taken (see in SI Figure S2). For 22 out of the 26 compounds showing a difference greater than 0.5 kcal/mol with the applied distance criterion, docking into the MDeNM conformations outperformed the MD ones. The assessment of ligands for which there was a significant difference between MD and MDeNM (greater than 1 kcal/mol) revealed that most of the compounds for which MDeNM performed better were of big size, occupying a large volume in the binding pocket, and their poses corresponding to the best BE were accommodated within widely open SULT1A1/PAPS conformations. These conformations were either not generated or poorly populated by the MD simulations (see in SI Figure S3).

2.3. Implication of substrate binding and SULT1A1 flexibility for gating mechanism elucidation

To further investigate the gating mechanism and substrate recognition of SULT1A1, we additionally analyzed the docking of two estrogens, the substrates 17β-estradiol (E2) and fulvestrant, previously suggested to be accommodated via different mechanisms depending on the co-factor induced isomerization (Cook et al., 2013a). E2 is a smaller, medium-sized substrate of SULT1A1 that contains a phenolic-hydroxyl group at the C3, and a hydroxyl group at the 17β position. Fulvestrant is an estrogen analogue, a larger substrate of SULT1A1, with an additional 15-atom long functional sidechain at the C7 position. E2 and fulvestrant were both docked into 6000 structures generated by MD and 6000 other structures generated by MDeNM (they were taken every 100 ps during MD and after every second relaxation phase in

MDeNM, respectively). The docking poses of both E2 and fulvestrant were considered acceptable on a given enzyme conformation if the BE was lower than -5 kcal/mol (more favorable binding energy) and the distance between the PAPS sulfate and the ligand's nucleophilic hydroxyl oxygen was less than 5 Å. Although it has been shown that the formation of fulvestrant-3-sulfate/estradiol-3-sulfate is preferable, it is also possible that low levels of fulvestrant-17-sulfate/estradiol-17-sulfate are produced (Edavana et al., 2011). The distribution of conformations capable of accommodating E2 and fulvestrant, along the formerly defined distances d(L1,L2) and d(L1,L3), is shown in FIGURE A.5.

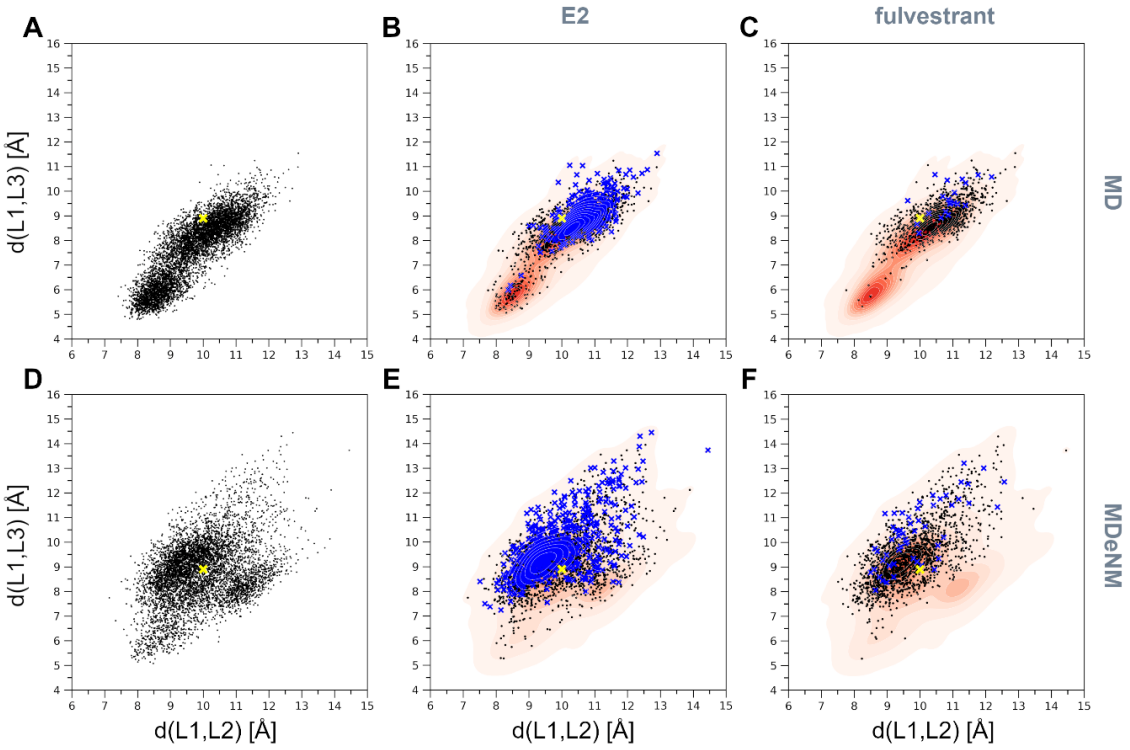

*Figure A.5: Distribution within the space defined by d(L1,L2) and d(L1,L3) distances for **(A)** the MD generated structures, **(B)** MD structures capable of accommodating competent E2 orientations, **(C)** MD structures capable of accommodating competent fulvestrant orientations; **(D)** the MDeNM generated structures, **(E)** MDeNM structures capable of accommodating competent E2 orientations, and **(F)** MDeNM structures capable of accommodating competent fulvestrant orientations. Conformations showing a BE stronger than -5 kcal/mol are indicated in black points and those showing a BE stronger than -10 kcal/mol are indicated in blue 'x'-es on parts B, C, E, and F. The initial crystal structure (4GRA.pdb) is shown in yellow 'x'.*

MD and MDeNM conformations were capable of accommodating E2, regardless of their openness (FIGURE A.5B and E), which agrees with previous kinetic and binding studies showing that E2 can bind to open and closed conformations of SULT1A1 (Cook et al., 2013b). The analysis of the conformations showing the strongest BEs (having a BE to estradiol lower than -10 kcal/mol; denoted by blue 'x') further indicates that the extremely closed state is mostly unfavorable even for estradiol binding. This is in line with the fact that E2 is a medium-size substrate of SULT1A1. Fulvestrant showed, even more, an obvious preference towards open

conformations. Similarly to MD, as mentioned above, the opening along d(L1,L2) and d(L1,L3) is restricted by the high correlation between them; hence opening along both distances is required for fulvestrant to dock (FIGURE A.5C). MDeNM results reveal, however, that the opening along d(L1,L3) rather than d(L1,L2) is essential for fulvestrant (FIGURE A.5F). Analysis of the best docking results of fulvestrant (having a BE lower than -10 kcal/mol; denoted by blue 'x') further confirmed that only conformations with a great d(L1,L3) distance are favorable for fulvestrant docking. MDeNM simulations were capable of generating widely open conformations accessible for fulvestrant, 3 Å along d(L1,L3) beyond MD conformations. Both MD and MDeNM results confirm that, open conformations are still available for big ligands to bind even with the co-factor bound.

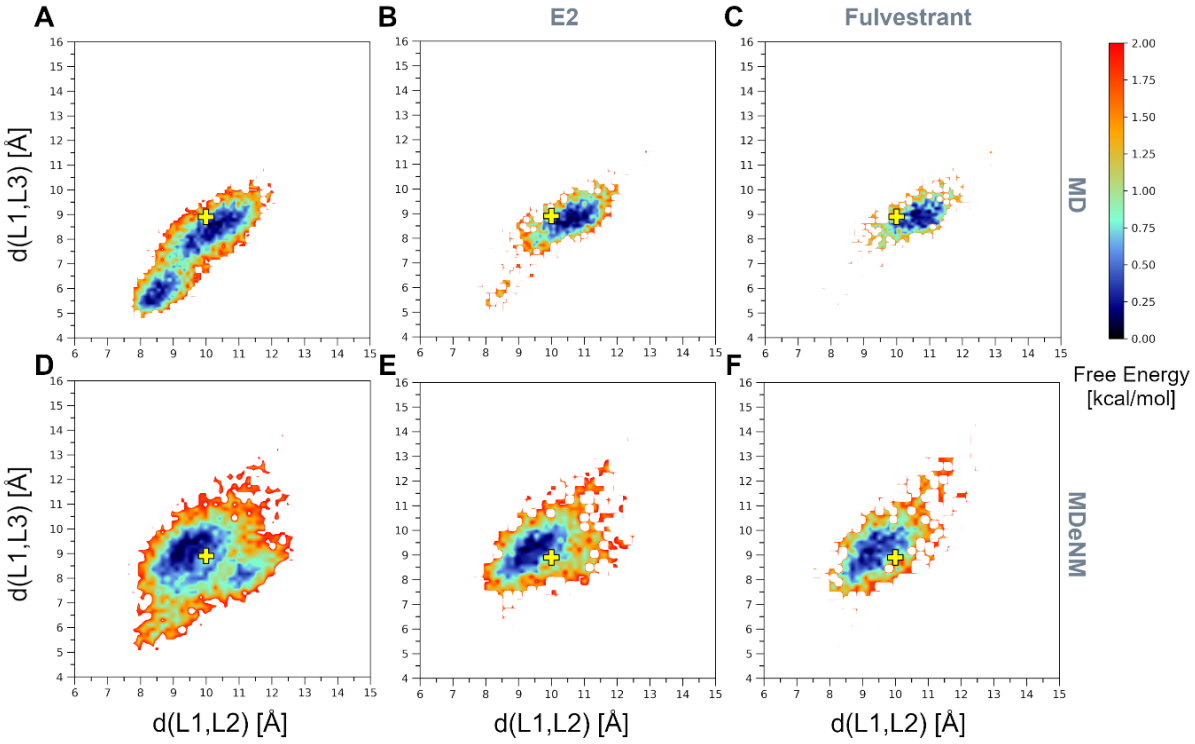

*Figure A.6: Free Energy Landscapes (FELs) in the space defined by the distances d(L1,L2) and d(L1,L3) of **(A)** the MD generated structures, **(B)** MD structures capable of accommodating competent E2 orientations, **(C)** MD structures capable of accommodating competent fulvestrant orientations; **(D)** the MDeNM generated structures, **(E)** MDeNM structures capable of accommodating competent E2 orientations, and **(F)** MDeNM structures capable of accommodating competent fulvestrant orientations. The initial crystal structure (4GRA.pdb) is denoted by yellow '+'.*

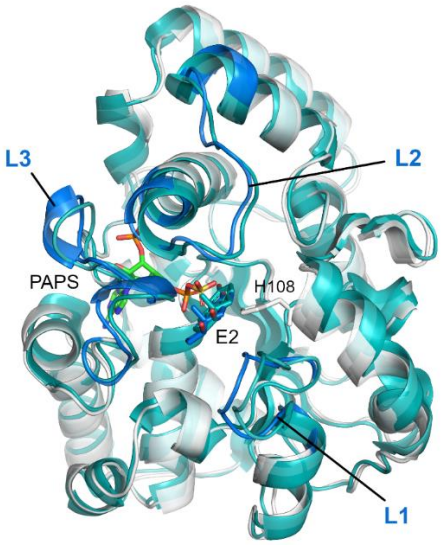

*Figure A.7: A favorable docking position of E2 in an MDeNM generated conformation (in white) superposed to the crystal structure of SULT1A1*2 co-crystalized with E2 (PDB 2D06 in cyan).*

The distribution of conformations shown in FIGURE A.5 were also transformed in Free Energy Landscapes (FEL) according to Equation A.1 (see Materials and Methods) and are shown in FIGURE A.6. Interestingly, most of the conformations capable of accommodating competent E2 and fulvestrant are of low free energies. An example of a favorable position of E2 docked into an MDeNM generated conformation (FIGURE A.7) illustrates the excellent superposition to the bioactive conformation of E2 in the structure of SULT1A1*2 co-crystallized with E2. FIGURE A.8 shows competent docking positions of fulvestrant in three MD and three MDeNM generated conformations. Their comparison with the crystal structure of apo SULT1A1*1 (PDB ID 4GRA) demonstrates the utility of using MDeNM simulations, suggesting a larger opening of the pore than observed by the MD simulations and facilitating thus the accommodation of large substrates as fulvestrant.

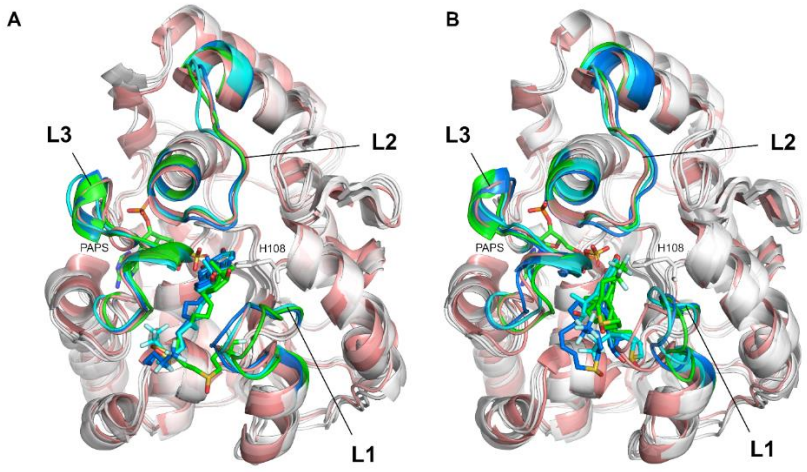

*Figure A.8: Favorable docking positions of fulvestrant in A. three MD and B. three MDeNM generated conformations. The apo crystal structure of SULT1A1*1 (4GRA.pdb) is shown in salmon for reference.*

Further MD simulations were performed for SULT1A1/PAPS bound to a substrate. The best-docked structures for the two substrates E2 and fulvestrant, having the best docking scores and competent positions, were chosen as starting structures for the additional MD simulations. Two docked positions of E2 were chosen, one in an MD - and one in an MDeNM - generated conformations (shown in FIGURE A.7). For the fulvestrant, three and three starting positions were chosen out of the MD- and MDeNM - generated conformations, respectively (shown in FIGURE A.8). In 7 out of the 8 MD simulations, the substrate remained in a stable position keeping a distance between the hydroxyl group of the ligand and the sulfate group of PAPS within 5 Å. The unstable fulvestrant-bound complex, starting from an MDeNM conformation, had a significantly different initial substrate orientation compared to the co-crystallized structure of E2 (see in SI Figure S4F model 2). The binding energies of the two substrates and SULT1A1/PAPS calculated with Autodock Vina scoring function for the complexes' structures before, and after the 100 ns MD simulations are shown in SI Table S2. It is seen that after all MD simulations with a bound substrate, the predicted binding energies for E2 and fulvestrant (SI Table S2) are closer to the experimental ones (SI Table S1) as compared to the energies calculated after docking only (SI Table S2).

To compare the MD simulations with and without bound substrates, the FELs were calculated with respect to the distances d(L1,L2) and d(L1,L3) (see FIGURE A.6 and SI Figure S4). The energetically most stable states of the MD simulations with a bound substrate correspond in all cases to conformations that are more open than the crystal structure 4GRA.pdb, both for E2 and fulvestrant. Interestingly, both MD and, to a greater extent, MDeNM were able to generate open conformations starting from the apo-state (without a bound ligand) (FIGURE A.6), corresponding to these energetically stable MD states in the presence of a bound substrate. Except for the one unstable MD simulation in the presence of fulvestrant as discussed above, both MD simulations with estradiol, and the other five MD simulations with fulvestrant show the induced further opening of the loops in the presence of a bound substrate.

These results are in agreement with previous indications that SULT undergoes a large opening to accommodate very large SULT substrates such as fulvestrant, 4-hydroxytamoxifen, or raloxifene (Cook et al., 2013a, Daniels and Kadlubar, 2013, Falany et al., 2006). However, we should note that the above discussed open SULT1A1/PAPS structures were generated in the presence of PAPS in our case. Thus, our simulations do not entirely support the assumption that recognition of large substrates is dependent on a co-factor isomerization as proposed in (Cook et al., 2013a, Zhu et al., 2019). Furthermore, allosteric binding was previously proposed to occur for some inhibitors in one part of the large cavity, assuring the substrates' access close to the co-factor (Coughtrie and Johnston, 2001). Previous studies suggested that inhibitors like catechins (naturally occurring flavonols) (Coughtrie and Johnston, 2001) or epigallocatechin gallate (EGCG) (Cook et al., 2015a) might inhibit SULT1A1 allosterically close to that cavity. Detailed analysis of our MDeNM results on the flexibility of this large cavity area – constituted by the active site and the pore (also called the catechin-binding site (Cook et al., 2016)), sometimes accommodating a second inhibitor molecule (e.g. p-Nitrophenol, see PDB ID 1LS6

(Gamage et al., 2003)) – showed that some L1 and L3 conformations (e.g. seen in FIGURE A.8B) ensure sufficient opening of the pore to accommodate large inhibitors like EGCG, and thus such binding into the pore (Cook et al., 2016, Cook et al., 2015a) might not be considered as allosteric.

3. Conclusion

In this study, we employed MD simulations and the recently developed MDeNM approach to elucidate the molecular mechanisms guiding the recognition of diverse substrates and inhibitors by SULT1A1. MDeNM allowed exploring an extended conformational space of PAPS-bound SULT1A1, which has not been achieved by using classical MD. Our simulations and analyses on the binding of the substrates estradiol and fulvestrant demonstrated that large conformational changes of the PAPS-bound SULT1A1 could occur independently of the co-factor movements. We argue that the flexibility of SULT1A1 ensured by loops L1, L2, and L3 in the presence of the co-factor is extremely high and may be sufficient for significant structural displacements for large ligands, substrates, or inhibitors. Such mechanisms can ensure the substrate recognition and the SULT specificity for various ligands larger than expected, as exemplified here with fulvestrant. Altogether, our observations shed new light on the complex mechanisms of substrate specificity and inhibition of SULT, which play a key role in the xenobiotics and Phase II drug metabolism (Testa et al., 2012, Bojarova and Williams, 2008). In this direction, the results obtained using the MDeNM simulations were valuable and highlighted the utility of including MDeNM in protein-ligand interactions studies where major rearrangements are expected.

4. Materials and Methods

4.1. Protein structures preparation

Some studies indicate that the SULTs are half-site reactive enzymes, and when the nucleotide is bound at only one subunit of the SULT dimer, the “Cap” of that subunit will spend most of its time in the “closed” conformation (Wang et al., 2016). Although the dimer interface is adjacent both to the PAPS binding domain and the active site “Cap” of the SULTs in some X-ray structures (e.g. PDB ID 2D06 , SULT1A1 co-crystallized with PAP and E2), suggesting that the interaction between the two subunits may play a role in the enzyme activity, SULT monomers retain their activity in vitro (Cook et al., 2015a). Furthermore, in other X-ray structures, a different dimer binding site is observed (e.g. PDB ID 2Z5F, SULT1B1 co-crystallized with PAP). Previously, identical behaviors were observed when simulations were performed with monomers or dimers constructed using the canonical interface (Cook et al., 2013a). Here, all simulations were performed using monomer structures.

Several crystal structures of SULT1A1 are available in the Protein Data Bank (www.rcsb.org). The only available structure of SULT1A1*1 containing R213 and M223 without bound ligand was selected, PDB ID: 4GRA (Cook et al., 2013a). The co-factor PAP present in the 4GRA structure was replaced by PAPS. The PAPS structure was taken of SULT1E1 (PDB ID: 1HY3 (Pedersen et al., 2002)) and superposed to PAP in 4GRA.pdb by overlapping their common

heavy atoms; the differing sulfate group of PAPS did not cause any steric clashes with the protein. The pKa values of the protein titratable groups were calculated with PROPKA (Sondergaard et al., 2011), and the protonation states were assigned at pH 7.0. PAPS parameters were determined by using the CHARMM General Force Field 2.2.0 (CGenFF) (Vanommeslaeghe et al., 2010). The partial charges of PAPS were optimized using quantum molecular geometry optimization simulation (QM Gaussian optimization, ESP charge routine(Frisch et al., 2016)) with the b3lyp DFT exchange correlation functional using the 6-311+g(d,p) basis set.

A rectangular box of TIP3 water molecules with 14 Å in all directions from the protein surface (82 Å x 82 Å x 82 Å) was generated with CHARMM-GUI (Jo et al., 2008, Lee et al., 2016), and the NaCl concentration was set to 0.15 M, randomly placing the ions in the unit cell. The solvated system was energy minimized with progressively decreasing harmonic restraints applied to atomic positions: steepest descent (SD) was first used where the harmonic force constant was decreased every 100 steps adopting the values 50, 10, 1, and 0.1 kcal/mol/$Å^2$. The system was further minimized without harmonic restraints by performing successive cycles of SD and Adopted Basis Newton-Raphson (ABNR) minimizations till a tolerance of RMS energy gradient of 0.01 kcal/mol/Å was reached. The minimization was performed with CHARMM (Brooks et al., 2009) using the additive all-atom CHARMM force field C36m (Huang and MacKerell, 2013). The system was then heated and equilibrated at 300 K for 100 ps in an NVT ensemble followed by a 5 ns NPT run at 1 atm pressure. The equilibration was performed with NAMD (Phillips et al., 2020) using the additive all-atom CHARMM force field C36m(Huang and MacKerell, 2013). For constant temperature control Langevin dynamics was used with a damping coefficient of 1 ps^{-1}. The constant pressure was achieved by using Nose-Hoover method with a piston oscillation period of 50 fs, and a piston oscillation decay time of 25 fs. The integration time step was set to 2 fs. For the energy calculations, the dielectric constant was set to 1. The particle mesh Ewald (PME) method was used to calculate the electrostatic interactions with a grid spacing of 1 Å or less having the order of 6. The real space summation was truncated at 12.0 Å, and the width of Gaussian distribution was set to 0.34 $Å^{-1}$. Van der Waals interactions were reduced to zero by 'switch' truncation operating between 10.0 and 12.0 Å.

4.2. MD simulations

MD simulations were carried out with NAMD (Phillips et al., 2020) using the all-atom CHARMM force field C36m (Huang and MacKerell, 2013). Three parallel 200 ns long MD simulations were performed for SULT1A1/PAPS without bound ligand starting from the equilibrated structure, with random velocities assigned according to the Maxwell-Boltzmann distribution at 300 K. A time step of 2 fs was used, with the coordinates saved every 10 ps. The parameters for the 200 ns runs were identical to those used for the previously described NPT equilibration of 5 ns. Additional 8 MD simulations of 100 ns were performed for SULT1A1/PAPS in the presence of a bound substrate (E2 and fulvestrant) starting from different substrate positions. The

parameters of E2 and fulvestrant were determined by CGenFF. The same MD protocol was then applied as detailed above for SULT1A1/PAPS without bound substrate.

4.3. MDeNM simulations

MDeNM simulations and analyses were performed with CHARMM (Brooks et al., 2009) using the all-atom CHARMM force field C36m (Huang and MacKerell, 2013). Starting from the same equilibrated SULT1A1 structure in solution as for the MD simulations, the MDeNM approach was used to map its conformational surface(Costa et al., 2015) thoroughly. The equilibrated structure was first energy minimized to calculate the normal modes. For energy minimization, we first used the steepest descent (SD) method with harmonic restraining potentials applied to atomic positions whose force constant were decreased from 10, 1, 0.1, and 0 kcal/mol/Å^2 every 500 steps. It was followed by the Adopted Basis Newton-Raphson minimization to reach an RMS energy gradient of 10^{-5} kcal/mol/Å. The normal modes of the energy minimized structure were calculated using the VIBRAN module of CHARMM (Woodcock et al., 2008). For the MDeNM calculations, the three low-frequency normal modes contributing the most to the highest RMSF of atomic displacements were taken.

Then, random linear combinations of these modes were generated such that the RMSDs between 1 Å displaced structures along these combined NM directions were greater than 0.3 Å. This provided the directions for unbiased coverage of the large-scale conformational space of the protein. In total, 240 different directions were created. For each of them, MD simulations were performed within which the motion described by the combined NM vector was kinetically promoted; this was achieved by adding to the current MD velocities an additional velocity in the direction of the NM combined vector corresponding to an overall 2 K increase of the system's temperature. As the excitation energy rapidly dissipates in less than 1 ps, a series of 50 consecutive excitations were achieved after every 4 ps of the MD simulation to allow the system to evolve and relax. Thus, the total MDeNM simulation time was 240 x 50 x 4ps = 48ns. The other MD parameters were the same as the given ones in the previous paragraph on *"MD simulations".*

4.4. Clustering

The Quality Threshold (QT) algorithm (Heyer et al., 1999) as implemented in VMD (Humphrey et al., 1996) was applied to perform conformational clustering of the MD generated conformations. A distance function defined as the RMSD difference calculated for the heavy atoms of the binding pocket (see in SI for its definition) was used with the maximum cluster diameter set to 1.1 Å. The centers of the 94 most populated clusters containing 85 % of all the conformations were then used to dock known substrates and inhibitors of SULT1A1. In the case of the MDeNM generated conformations, the population of clusters is biased due to the common starting structure for each replica and the applied RMSD filtering upon the generation of the excitation directions. A pseudo-uniform selection from all the MDeNM generated conformations was applied with a spacing of 1.1 Å in the RMSD space defined by residues within the binding pocket to create a representative set. A total of 86 structures were retrieved and used for the docking of known substrates and inhibitors of SULT1A1.

4.5. Docking

Docking experiments were performed with AutoDock Vina 1.1.2 (Trott and Olson, 2010) that employs gradient-based conformational docking and an empirical scoring function predicting the protein-ligand binding energy in kcal/mol. A list of 132 known substrates and inhibitors of SULT1A1 were taken, collected in our previous work (Martiny et al., 2013) and (Paitz and Bowden, 2013, Cook et al., 2012). The protein conformations selected for docking were pre-processed with AutoDockTools (Morris et al., 2009), the solvent was removed, non-polar hydrogens were merged, and Gasteiger charges were assigned. The ligands were prepared for the docking using AutoDockTools. A grid box of 24 Å x 24 Å x 24 Å was centered on the binding pocket with a spacing of 1 Å. The grid center was set to x = 27.050 Å; y = 17.520 Å; z = 17.653 Å with respect to the crystal structure 4GRA.pdb. The maximum number of binding modes was set to 20, the exhaustiveness of the global search to 10, the maximum energy difference between the retained best and worst binding modes to 15 kcal/mol. During the docking, the ligands and the binding site residues K106 and F247 observed to change their side-chain conformations easily during the MD and MDeNM simulations were handled flexibly; the rest of the protein and the co-factor were kept rigid.

4.6. Free Energy Landscape (FEL) analysis

FELs of conformations corresponding to the different MD and MDeNM simulations were calculated within the plane defined by the distances d(L1,L2) and d(L1,L3). The most populated state was used as a reference for calculating free energy differences. The free energy difference (ΔG_α) of a given state α was determined by considering the probability of the occurrence of the two states $P(q_\alpha)$ *and* $P_{max}(q)$ given by the equation:

$$\Delta G_\alpha = -k_B T ln\left[\frac{P(q_\alpha)}{P_{max}(q)}\right] \qquad (A.1)$$

where k_B is the Boltzmann constant, T is the temperature of the simulation, $P(q_\alpha)$ *is* an estimate of the probability density function obtained from the bi-dimensional histogram of the conformations distribution in the plane of d(L1,L2) and d(L1,L3) during the simulation. $P_{max}(q)$ is the probability of the most populated state.

Acknowledgements

This project has been funded by the French ANR (project ToxME), Paris Univ., Inserm Institute and the French-Hungarian bilateral program NKFIH 2019-2.1.11-TÉT-2020-00096. The MD and MDeNM simulations were carried out using the Hungarian supercomputing facility KIFÜ. We thank Dr. G. Clavier (PPSM CNRS, ENS Paris-Saclay, France) for helpful discussion.

Author contributions

Conceptualization: E.B., D.P., M.A.M; investigation: all; methodology: all; analyses: B.D., D.T; original draft preparation: B.D., D.T ; editing manuscript: E.B., D.P., A. N., M.A.M; All authors have read and agreed to the published version of the manuscript.

Competing Interests

The authors declare no competing interests.

5. References

ALLALI-HASSANI, A., PAN, P. W., DOMBROVSKI, L., NAJMANOVICH, R., TEMPEL, W., DONG, A., LOPPNAU, P., MARTIN, F., THORNTON, J., EDWARDS, A. M., BOCHKAREV, A., PLOTNIKOV, A. N., VEDADI, M. & ARROWSMITH, C. H. 2007. Structural and chemical profiling of the human cytosolic sulfotransferases. *PLoS Biol,* 5, e97. doi:10.1371/journal.pbio.0050097

BERMAN, H. M., WESTBROOK, J., FENG, Z., GILLILAND, G., BHAT, T. N., WEISSIG, H., SHINDYALOV, I. N. & BOURNE, P. E. 2000. The Protein Data Bank. *Nucleic Acids Res,* 28, 235-42. doi:10.1093/nar/28.1.235

BOJAROVA, P. & WILLIAMS, S. J. 2008. Sulfotransferases, sulfatases and formylglycine-generating enzymes: a sulfation fascination. *Curr Opin Chem Biol,* 12, 573-81. doi:10.1016/j.cbpa.2008.06.018

BROOKS, B. R., BROOKS, C. L., 3RD, MACKERELL, A. D., JR., NILSSON, L., PETRELLA, R. J., ROUX, B., WON, Y., ARCHONTIS, G., BARTELS, C., BORESCH, S., CAFLISCH, A., CAVES, L., CUI, Q., DINNER, A. R., FEIG, M., FISCHER, S., GAO, J., HODOSCEK, M., IM, W., KUCZERA, K., LAZARIDIS, T., MA, J., OVCHINNIKOV, V., PACI, E., PASTOR, R. W., POST, C. B., PU, J. Z., SCHAEFER, M., TIDOR, B., VENABLE, R. M., WOODCOCK, H. L., WU, X., YANG, W., YORK, D. M. & KARPLUS, M. 2009. CHARMM: the biomolecular simulation program. *J Comput Chem,* 30, 1545-614. doi:10.1002/jcc.21287

CHAPMAN, E., BEST, M. D., HANSON, S. R. & WONG, C. H. 2004. Sulfotransferases: structure, mechanism, biological activity, inhibition, and synthetic utility. *Angew Chem Int Ed Engl,* 43, 3526-48. doi:10.1002/anie.200300631

COOK, I., WANG, T., ALMO, S. C., KIM, J., FALANY, C. N. & LEYH, T. S. 2013a. The gate that governs sulfotransferase selectivity. *Biochemistry,* 52, 415-24. doi:10.1021/bi301492j

COOK, I., WANG, T., FALANY, C. N. & LEYH, T. S. 2012. A nucleotide-gated molecular pore selects sulfotransferase substrates. *Biochemistry,* 51, 5674-83. doi:10.1021/bi300631g

COOK, I., WANG, T., FALANY, C. N. & LEYH, T. S. 2013b. High accuracy in silico sulfotransferase models. *J Biol Chem,* 288, 34494-501. doi:10.1074/jbc.M113.510974

COOK, I., WANG, T., FALANY, C. N. & LEYH, T. S. 2015a. The allosteric binding sites of sulfotransferase 1A1. *Drug Metab Dispos,* 43, 418-23. doi:10.1124/dmd.114.061887

COOK, I., WANG, T., GIRVIN, M. & LEYH, T. S. 2016. The structure of the catechin-binding site of human sulfotransferase 1A1. *Proc Natl Acad Sci U S A,* 113, 14312-14317. doi:10.1073/pnas.1613913113

COOK, I., WANG, T. & LEYH, T. S. 2015b. Sulfotransferase 1A1 Substrate Selectivity: A Molecular Clamp Mechanism. *Biochemistry,* 54, 6114-22. doi:10.1021/acs.biochem.5b00406

COSTA, M. G., BATISTA, P. R., BISCH, P. M. & PERAHIA, D. 2015. Exploring free energy landscapes of large conformational changes: molecular dynamics with excited normal modes. *J Chem Theory Comput,* 11, 2755-67. doi:10.1021/acs.jctc.5b00003

COUGHTRIE, M. W. & JOHNSTON, L. E. 2001. Interactions between dietary chemicals and human sulfotransferases-molecular mechanisms and clinical significance. *Drug Metab Dispos,* 29, 522-8.

DAJANI, R., HOOD, A. M. & COUGHTRIE, M. W. 1998. A single amino acid, glu146, governs the substrate specificity of a human dopamine sulfotransferase, SULT1A3. *Mol Pharmacol,* 54, 942-8. doi:10.1124/mol.54.6.942

DANIELS, J. & KADLUBAR, S. 2013. Sulfotransferase genetic variation: from cancer risk to treatment response. *Drug Metab Rev,* 45, 415-22. doi:10.3109/03602532.2013.835621

DONG, D., AKO, R. & WU, B. 2012. Crystal structures of human sulfotransferases: insights into the mechanisms of action and substrate selectivity. *Expert Opin Drug Metab Toxicol,* 8, 635-46. doi:10.1517/17425255.2012.677027

DUDAS, B., MERZEL, F., JANG, H., NUSSINOV, R., PERAHIA, D. & BALOG, E. 2020. Nucleotide-Specific Autoinhibition of Full-Length K-Ras4B Identified by Extensive Conformational Sampling. *Front Mol Biosci,* 7, 145. doi:10.3389/fmolb.2020.00145

DUDAS, B., PERAHIA, D. & BALOG, E. 2021. Revealing the activation mechanism of autoinhibited RalF by integrated simulation and experimental approaches. *Sci Rep,* doi:10.1038/s41598-021-89169-5

EDAVANA, V. K., YU, X., DHAKAL, I. B., WILLIAMS, S., NING, B., COOK, I. T., CALDWELL, D., FALANY, C. N. & KADLUBAR, S. 2011. Sulfation of fulvestrant by human liver cytosols and recombinant SULT1A1 and SULT1E1. *Pharmgenomics Pers Med,* 4, 137-145. doi:10.2147/PGPM.S25418

FAGNEN, C., BANNWARTH, L., OUBELLA, I., FOREST, E., DE ZORZI, R., DE ARAUJO, A., MHOUMADI, Y., BENDAHHOU, S., PERAHIA, D. & VENIEN-BRYAN, C. 2020. New Structural insights into Kir channel gating from molecular simulations, HDX-MS and functional studies. *Sci Rep,* 10, 8392. doi:10.1038/s41598-020-65246-z

FALANY, J. L., PILLOFF, D. E., LEYH, T. S. & FALANY, C. N. 2006. Sulfation of raloxifene and 4-hydroxytamoxifen by human cytosolic sulfotransferases. *Drug Metab Dispos,* 34, 361-8. doi:10.1124/dmd.105.006551

FRISCH, M. J., TRUCKS, G. W., SCHLEGEL, H. B., SCUSERIA, G. E., ROBB, M. A., CHEESEMAN, J. R., SCALMANI, G., BARONE, V., PETERSSON, G. A., NAKATSUJI, H., LI, X., CARICATO, M., MARENICH, A. V., BLOINO, J., JANESKO, B. G., GOMPERTS, R., MENNUCCI, B., HRATCHIAN, H. P., ORTIZ, J. V., IZMAYLOV, A. F., SONNENBERG, J. L., WILLIAMS, DING, F., LIPPARINI, F., EGIDI, F., GOINGS, J., PENG, B., PETRONE, A., HENDERSON, T., RANASINGHE, D., ZAKRZEWSKI, V. G., GAO, J., REGA, N., ZHENG, G., LIANG, W., HADA, M., EHARA, M., TOYOTA, K., FUKUDA, R., HASEGAWA, J., ISHIDA, M., NAKAJIMA, T., HONDA, Y., KITAO, O., NAKAI, H., VREVEN, T., THROSSELL, K., MONTGOMERY JR., J. A., PERALTA, J. E., OGLIARO, F., BEARPARK, M. J., HEYD, J. J., BROTHERS, E. N., KUDIN, K. N., STAROVEROV, V. N., KEITH, T. A., KOBAYASHI, R., NORMAND, J., RAGHAVACHARI, K., RENDELL, A. P., BURANT, J. C., IYENGAR, S. S., TOMASI, J., COSSI, M., MILLAM, J. M., KLENE, M., ADAMO, C., CAMMI, R., OCHTERSKI, J. W., MARTIN, R. L., MOROKUMA, K., FARKAS, O., FORESMAN, J. B. & FOX, D. J. 2016. Gaussian 16 Rev. C.01. Wallingford, CT.

GAMAGE, N., BARNETT, A., HEMPEL, N., DUGGLEBY, R. G., WINDMILL, K. F., MARTIN, J. L. & MCMANUS, M. E. 2006. Human sulfotransferases and their role in chemical metabolism. *Toxicol Sci,* 90, 5-22. doi:10.1093/toxsci/kfj061

GAMAGE, N. U., DUGGLEBY, R. G., BARNETT, A. C., TRESILLIAN, M., LATHAM, C. F., LIYOU, N. E., MCMANUS, M. E. & MARTIN, J. L. 2003. Structure of a human carcinogen-converting enzyme, SULT1A1. Structural and kinetic implications of substrate inhibition. *J Biol Chem,* 278, 7655-62. doi:10.1074/jbc.M207246200

GAMAGE, N. U., TSVETANOV, S., DUGGLEBY, R. G., MCMANUS, M. E. & MARTIN, J. L. 2005. The structure of human SULT1A1 crystallized with estradiol. An insight into active site plasticity and substrate inhibition with multi-ring substrates. *J Biol Chem,* 280, 41482-6. doi:10.1074/jbc.M508289200

GOMES, A. A. S., CARDOSO, F. F., SOUZA, M. F., OLIVEIRA, C. L. P., PERAHIA, D., MAGRO, A. J. & FONTES, M. R. M. 2020. The allosteric activation mechanism of a phospholipase A2-like toxin from Bothrops jararacussu venom: a dynamic description. *Sci Rep,* 10, 16252. doi:10.1038/s41598-020-73134-9

GUENGERICH, F. P., WILKEY, C. J., GLASS, S. M. & REDDISH, M. J. 2019. Conformational selection dominates binding of steroids to human cytochrome P450 17A1. *J Biol Chem,* 294, 10028-10041. doi:10.1074/jbc.RA119.008860

HEMPEL, N., GAMAGE, N., MARTIN, J. L. & MCMANUS, M. E. 2007. Human cytosolic sulfotransferase SULT1A1. *Int J Biochem Cell Biol,* 39, 685-9. doi:10.1016/j.biocel.2006.10.002

HEYER, L. J., KRUGLYAK, S. & YOOSEPH, S. 1999. Exploring expression data: identification and analysis of coexpressed genes. *Genome Res,* 9, 1106-15. doi:10.1101/gr.9.11.1106

HUANG, J. & MACKERELL, A. D., JR. 2013. CHARMM36 all-atom additive protein force field: validation based on comparison to NMR data. *J Comput Chem,* 34, 2135-45. doi:10.1002/jcc.23354

HUMPHREY, W., DALKE, A. & SCHULTEN, K. 1996. VMD: visual molecular dynamics. *J Mol Graph,* 14, 33-8, 27-8. doi:10.1016/0263-7855(96)00018-5

JO, S., KIM, T., IYER, V. G. & IM, W. 2008. CHARMM-GUI: a web-based graphical user interface for CHARMM. *J Comput Chem,* 29, 1859-65. doi:10.1002/jcc.20945

LEE, J., CHENG, X., SWAILS, J. M., YEOM, M. S., EASTMAN, P. K., LEMKUL, J. A., WEI, S., BUCKNER, J., JEONG, J. C., QI, Y., JO, S., PANDE, V. S., CASE, D. A., BROOKS, C. L., 3RD, MACKERELL, A. D., JR., KLAUDA, J. B. & IM, W. 2016. CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations Using the CHARMM36 Additive Force Field. *J Chem Theory Comput,* 12, 405-13. doi:10.1021/acs.jctc.5b00935

LEE, K. A., FUDA, H., LEE, Y. C., NEGISHI, M., STROTT, C. A. & PEDERSEN, L. C. 2003. Crystal structure of human cholesterol sulfotransferase (SULT2B1b) in the presence of pregnenolone and 3'-phosphoadenosine 5'-

phosphate. Rationale for specificity differences between prototypical SULT2A1 and the SULT2BG1 isoforms. *J Biol Chem,* 278, 44593-9. doi:10.1074/jbc.M308312200

LOUET, M., LABBE, C. M., FAGNEN, C., AONO, C. M., HOMEM-DE-MELLO, P., VILLOUTREIX, B. O. & MITEVA, M. A. 2018. Insights into molecular mechanisms of drug metabolism dysfunction of human CYP2C9*30. *PLoS One,* 13, e0197249. doi:10.1371/journal.pone.0197249

LU, L. Y., CHIANG, H. P., CHEN, W. T. & YANG, Y. S. 2009. Dimerization is responsible for the structural stability of human sulfotransferase 1A1. *Drug Metab Dispos,* 37, 1083-8. doi:10.1124/dmd.108.025395

MARTINY, V. Y., CARBONELL, P., CHEVILLARD, F., MOROY, G., NICOT, A. B., VAYER, P., VILLOUTREIX, B. O. & MITEVA, M. A. 2015. Integrated structure- and ligand-based in silico approach to predict inhibition of cytochrome P450 2D6. *Bioinformatics,* 31, 3930-7. doi:10.1093/bioinformatics/btv486

MARTINY, V. Y., CARBONELL, P., LAGORCE, D., VILLOUTREIX, B. O., MOROY, G. & MITEVA, M. A. 2013. In silico mechanistic profiling to probe small molecule binding to sulfotransferases. *PLoS One,* 8, e73587. doi:10.1371/journal.pone.0073587

MARTINY, V. Y. & MITEVA, M. A. 2013. Advances in molecular modeling of human cytochrome P450 polymorphism. *J Mol Biol,* 425, 3978-92. doi:10.1016/j.jmb.2013.07.010

MOROY, G., SPERANDIO, O., RIELLAND, S., KHEMKA, S., DRUART, K., GOYAL, D., PERAHIA, D. & MITEVA, M. A. 2015. Sampling of conformational ensemble for virtual screening using molecular dynamics simulations and normal mode analysis. *Future Med Chem,* 7, 2317-31. doi:10.4155/fmc.15.150

MORRIS, G. M., HUEY, R., LINDSTROM, W., SANNER, M. F., BELEW, R. K., GOODSELL, D. S. & OLSON, A. J. 2009. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J Comput Chem,* 30, 2785-91. doi:10.1002/jcc.21256

MORTIER, J., RAKERS, C., BERMUDEZ, M., MURGUEITIO, M. S., RINIKER, S. & WOLBER, G. 2015. The impact of molecular dynamics on drug design: applications for the characterization of ligand-macromolecule complexes. *Drug Discov Today,* 20, 686-702. doi:10.1016/j.drudis.2015.01.003

PAITZ, R. T. & BOWDEN, R. M. 2013. Sulfonation of maternal steroids is a conserved metabolic pathway in vertebrates. *Integr Comp Biol,* 53, 895-901. doi:10.1093/icb/ict027

PANTALEAO, S. Q., PHILOT, E. A., DE RESENDE-LARA, P. T., LIMA, A. N., PERAHIA, D., MITEVA, M. A., SCOTT, A. L. & HONORIO, K. M. 2018. Structural Dynamics of DPP-4 and Its Influence on the Projection of Bioactive Ligands. *Molecules,* 23 doi:10.3390/molecules23020490

PEDERSEN, L. C., PETROTCHENKO, E., SHEVTSOV, S. & NEGISHI, M. 2002. Crystal structure of the human estrogen sulfotransferase-PAPS complex: evidence for catalytic role of Ser137 in the sulfuryl transfer reaction. *J Biol Chem,* 277, 17928-32. doi:10.1074/jbc.M111651200

PHILLIPS, J. C., HARDY, D. J., MAIA, J. D. C., STONE, J. E., RIBEIRO, J. V., BERNARDI, R. C., BUCH, R., FIORIN, G., HENIN, J., JIANG, W., MCGREEVY, R., MELO, M. C. R., RADAK, B. K., SKEEL, R. D., SINGHAROY, A., WANG, Y., ROUX, B., AKSIMENTIEV, A., LUTHEY-SCHULTEN, Z., KALE, L. V., SCHULTEN, K., CHIPOT, C. & TAJKHORSHID, E. 2020. Scalable molecular dynamics on CPU and GPU architectures with NAMD. *J Chem Phys,* 153, 044130. doi:10.1063/5.0014475

PRATT, W. B. & TAYLOR, P. 1990. *Principles of drug action : the basis of pharmacology,* New York, NY, USA, Churchill Livingstone.

RAKERS, C., SCHUMACHER, F., MEINL, W., GLATT, H., KLEUSER, B. & WOLBER, G. 2016. In Silico Prediction of Human Sulfotransferase 1E1 Activity Guided by Pharmacophores from Molecular Dynamics Simulations. *J Biol Chem,* 291, 58-71. doi:10.1074/jbc.M115.685610

SHIMADA, T. 2006. Xenobiotic-metabolizing enzymes involved in activation and detoxification of carcinogenic polycyclic aromatic hydrocarbons. *Drug Metab Pharmacokinet,* 21, 257-76. doi:10.2133/dmpk.21.257

SONDERGAARD, C. R., OLSSON, M. H., ROSTKOWSKI, M. & JENSEN, J. H. 2011. Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of pKa Values. *J Chem Theory Comput,* 7, 2284-95. doi:10.1021/ct200133y

SREJBER, M., NAVRATILOVA, V., PALONCYOVA, M., BAZGIER, V., BERKA, K., ANZENBACHER, P. & OTYEPKA, M. 2018. Membrane-attached mammalian cytochromes P450: An overview of the membrane's effects on structure, drug binding, and interactions with redox partners. *J Inorg Biochem,* 183, 117-136. doi:10.1016/j.jinorgbio.2018.03.002

SUN, H. & SCOTT, D. O. 2010. Structure-based drug metabolism predictions for drug design. *Chem Biol Drug Des,* 75, 3-17. doi:10.1111/j.1747-0285.2009.00899.x

TESTA, B., PEDRETTI, A. & VISTOLI, G. 2012. Reactions and enzymes in the metabolism of drugs and other xenobiotics. *Drug Discov Today,* 17, 549-60. doi:10.1016/j.drudis.2012.01.017

TIBBS, Z. E., ROHN-GLOWACKI, K. J., CRITTENDEN, F., GUIDRY, A. L. & FALANY, C. N. 2015. Structural plasticity in the human cytosolic sulfotransferase dimer and its role in substrate selectivity and catalysis. *Drug Metab Pharmacokinet,* 30, 3-20. doi:10.1016/j.dmpk.2014.10.004

TROTT, O. & OLSON, A. J. 2010. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem,* 31, 455-61. doi:10.1002/jcc.21334

VANOMMESLAEGHE, K., HATCHER, E., ACHARYA, C., KUNDU, S., ZHONG, S., SHIM, J., DARIAN, E., GUVENCH, O., LOPES, P., VOROBYOV, I. & MACKERELL, A. D., JR. 2010. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J Comput Chem,* 31, 671-90. doi:10.1002/jcc.21367

WANG, T., COOK, I. & LEYH, T. S. 2016. Design and Interpretation of Human Sulfotransferase 1A1 Assays. *Drug Metab Dispos,* 44, 481-4. doi:10.1124/dmd.115.068205

WANG, T., COOK, I. & LEYH, T. S. 2017. The NSAID allosteric site of human cytosolic sulfotransferases. *J Biol Chem,* 292, 20305-20312. doi:10.1074/jbc.M117.817387

WOODCOCK, H. L., ZHENG, W., GHYSELS, A., SHAO, Y., KONG, J. & BROOKS, B. R. 2008. Vibrational subsystem analysis: A method for probing free energies and correlations in the harmonic limit. *J Chem Phys,* 129, 214109. doi:10.1063/1.3013558

ZHU, J., QI, R., LIU, Y., ZHAO, L. & HAN, W. 2019. Mechanistic Insights into the Effect of Ligands on Structural Stability and Selectivity of Sulfotransferase 2A1 (SULT2A1). *ACS Omega,* 4, 22021-22034. doi:10.1021/acsomega.9b03136

B. UGT1A1

The primary sugar conjugation route in humans is glucuronidation (conjugation with α-D-glucuronic acid). Besides sulfonation, glucuronidation is the other major phase II reaction type in humans, the reaction is catalyzed by uridine 5'-diphosphate-glucuronosyltransferases (UGTs). The formation of glucuronide metabolites is quantitatively the most important form of conjugation for drugs as well as endogenous compounds. The isoenzyme UGT1A1 is of particular importance considering its broad substrate specificity and exclusive role in the glucuronidation, and therefore the detoxification of the endogenous heme breakdown by-product, bilirubin.

The following chapter on the prediction of UDP-glucuronosyltransferase inhibition introduces the combination of ligand- and structure-based information to train machine learning classification models. In 2013, the host laboratory was the first to integrate ligand- and structure-based information for the prediction of DME inhibition. The first models were trained by Martiny et al. on the SULT isoforms 1A1, 1A3, and 1E1, and at the same time by Cook et al. for the isoforms 1A1 and 2A1. As part of my PhD work, the first machine learning prediction models of UGT inhibition are presented next.

The flexibility of the cofactor-bound (UDP-glucuronic acid) UGT1A1 is addressed by performing classical MD simulations on a homology model of the human enzyme. The generated conformational ensemble is used for ensemble docking simulations after conformational clustering. An original selection of the most important UGT1A1 conformations and ligand-based descriptors in terms of their ability to discriminate between actives and decoys is performed. The predicted binding affinities are combined with the selected ligand-based descriptors and classification models are trained on active and decoy compounds using different supervised machine learning classification approaches. Hyper-parameter optimization of the models is performed including the determination of the most relevant descriptors to be used. The optimized models are implemented in the DrugME software developed in our lab (the author of the thesis is co-author of the DrugME software under Inserm license) that can be helpful for the prediction of drug-drug interactions of new drug candidates.

Machine learning and structure-based modeling for the prediction of UDP-glucuronosyltransferase inhibition

Balint Dudas[1,2,#], Youcef Bagdad[1,#], Milan Picard[1,ϕ] , David Perahia[2], Maria A. Miteva[1,*]

[1]Inserm U1268 MCTR, CiTCoM UMR 8038 CNRS – Université Paris Cité, Paris, France
[2]Laboratoire de Biologie et Pharmacologie Appliquée (LBPA), UMR8113, Ecole Normale Supérieure Paris-Saclay, Gif-sur-Yvette, France
ϕPresent address: Molecular Medicine Department, CHU de Québec Research Center, Université Laval, Québec, Canada

#1st coauthors
*corresponding author: maria.mitev@inserm.fr

Published in iScience on 2022 Oct 6
doi: 10.1016/j.isci.2022.105290

Abstract

UDP-glucuronosyltransferases (UGTs) are responsible for 35% of the phase II drug metabolism. In this study, we focused on UGT1A1, which is a key UGT isoform. Strong inhibition of UGT1A1 may trigger adverse drug/herb-drug interactions, or result in disorders of endobiotic metabolism. Most of the current machine learning methods predicting inhibition of drug metabolizing enzymes neglect protein structure and dynamics, both being essential for the recognition of various substrates and inhibitors. We performed molecular dynamics simulations on a homology model of the human UGT1A1 structure containing both the cofactor- (UDP-glucuronic acid) and the substrate-binding domains to explore UGT conformational changes. Then, we created models for the prediction of UGT1A1 inhibitors by integrating information on UGT1A1 structure and dynamics, interactions with diverse ligands, and machine learning. These models can be helpful for further prediction of drug-drug interactions of drug candidates and safety treatments.

Keywords: UDP-glucuronosyltransferase, UGT, drug metabolizing enzymes, machine learning, molecular dynamics

1. Introduction

Drug metabolizing enzymes (DMEs) are involved in the metabolism of endogenous molecules, xenobiotics, and drugs (Testa et al., 2012). They play a key role in the detoxification of organisms by modifying toxic substances and drugs to facilitate their elimination (Rowland et al., 2013, Grant, 1991). In some cases, the metabolites are more toxic, thereby inducing severe side effects and adverse drug reactions (Shimada, 2006), or their inhibition can lead to drug–drug interactions (DDI) (Prueksaritanont et al., 2013). While phase I drug metabolism (or functionalization) involves mainly oxidation-reduction reactions, phase II metabolism (or conjugation) reactions catalyze the covalent addition of polar groups like sulfate, glutathione, glucuronic acid, or others to a broad range of substrates (Testa et al., 2012, Almazroo et al., 2017). Numerous compounds (e.g. bilirubin, steroids, paracetamol, etc.) are conjugated directly by phase II enzymes without a preceding phase I reaction (Testa et al., 2012, Kaivosaari et al., 2011). Although conjugation reactions generally create water-soluble and inactive metabolites, reactive conjugated metabolites have also been reported (Osborne et al., 1992, Bauer et al., 1995, Shimada, 2006).

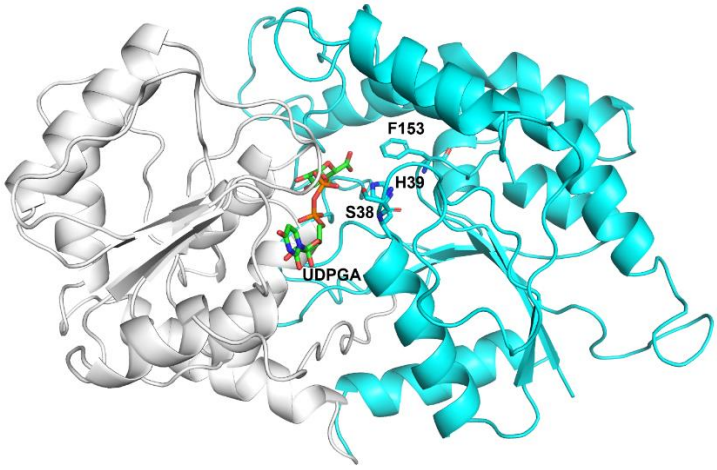

Figure B.1: Homology model of the human UGT1A1 (Locuson et al.) bound to the cofactor UDPGA containing both the cofactor-binding (in white) and the substrate-binding (in cyan) domains. The cofactor and key enzymatic residues are in licorice representation.

Uridine-Diphosphate (UDP)-glucuronosyl transferase (UGT) metabolism accounts for up to 35 % of all phase II DME reactions (Testa et al., 2012). UGT is a superfamily of phase II DMEs catalyzing the covalent addition of glucuronic acid to a wide range of substrates (Oda et al., 2015, Rowland et al., 2013) in the lumen of the endoplasmic reticulum (Meech and Mackenzie, 1997). Most human UGTs are physiologically highly expressed in the liver but are also present in other tissues like the intestine, the kidneys, the stomach, and the lungs (Ohno and Nakajin, 2009). Based on evolutionary divergence, mammalian UGTs can be divided into two families, UGT1 and UGT2. Human enzymes belonging to the UGT1 family all share an identical C-terminal domain which is responsible for the binding of the cofactor uridine-diphosphate glucuronic acid (UDPGA) and contains a Rossmann fold motif; and a characteristic N-terminal domain, containing highly variable regions, which is responsible for the substrate binding and accounts

for the selectivity of the different isoenzymes (FIGURE B.1) (Miners and Mackenzie, 1991, Tukey and Strassburg, 2000, Ritter et al., 1992). UGTs exhibit distinct but overlapping substrate specificity, and multiple UGT isoforms can be co-expressed in a given tissue (Court, 2005). The isoenzyme UGT1A1 is of particular importance accounting for 15 % of all UGT drug metabolism (Williams et al., 2004). It also plays an exclusive role in glucuronidation and, therefore, the detoxification of the endogenous heme breakdown by-product, bilirubin (Bosma, 2003).

Strong UGT1A1 inhibition may trigger adverse drug/herb-drug interactions, or can result in metabolic disorders of the endobiotic metabolism (Lv et al., 2019, Li et al., 2019, Liu et al., 2019). Numerous drugs, including virus protease inhibitors, tyrosine kinase inhibitors, and antifungal agents, have been reported to induce unconjugated hyperbilirubinemia or increase the concentration of cytotoxic agents through UGT1A1 inhibition in clinic (Steventon, 2020, Lv et al., 2019, Goon et al., 2016). Therefore, the European Medicines Agency (EMA) and the United States Food and Drug Administration (FDA) recommend testing for possible UGT1A1 inhibitor status over the course of drug development (Lv et al., 2019, Prueksaritanont et al., 2013) to avoid possible DDI.

Here, we focus on the prediction of inhibitors of UGT1A1. The prediction of DMEs inhibition is a challenging task (Kato, 2020) due to their promiscuous nature. Thus, it is important to consider protein structure and dynamics of DMEs, both being essential for the recognition of the various substrates and inhibitors. Structural information is also important to understand the molecular mechanism of the UGT catalyzed glucuronidation and its inhibition in all their complexity. Up to now, no experimental structures of the UGT1 family have been resolved. There are two crystal structures available only for the cofactor-binding domain of UGT2B7 (Miley et al., 2007) and UGT2B15 (Zhang et al., 2020). Multiple homology models of the human UGT1A1 have been published using the former structure together with plant and bacterial homologs (Locuson and Tracy, 2007, Li and Wu, 2007, Laakkonen and Finel, 2010). We have exploited such information to build prediction models of UGT inhibition using structure-based, and machine learning (ML) approaches. Previously, we have developed similar models to predict the inhibition of cytochrome P450 (Martiny et al., 2015, Goldwaser et al., 2022) and phase II sulfotransferases (Martiny et al., 2013). Machine learning approaches have become fundamental in all stages of drug discovery and development (Mao et al., 2021, Carracedo-Reboredo et al., 2021). During the last decade, several ML models have been developed to predict the site of metabolism for UGT-catalyzed reactions (Hwang et al., 2020, Peng et al., 2014, Cai et al., 2019, Sorich et al., 2008, Sorich et al., 2004). To the best of our knowledge, no predictive models have been reported to date for UGT inhibition.

In the present study, we integrated structural and ligand-based information in different machine learning approaches to generate predictive models of UGT inhibition. We ran molecular dynamics simulations on the human UGT1A1 structure containing both the cofactor- and the substrate-binding domains to consider conformational changes in its active site, critical for the accommodation of the diverse substrates and inhibitors, and performed docking simulations with a collection of experimentally validated UGT1A1 ligands to gain information

on enzyme-inhibitor interactions. We performed a rational selection of ligand-based descriptors and successfully trained ML models for the prediction of UGT1A1 inhibitors with around 90 % accuracy.

2. Results and Discussion

Our study combines structure-based modeling and machine learning to build models for predicting UGT1A1 inhibition. The workflow is shown in FIGURE B.2. For the dataset preparation, we collected known ligands of UGT1A1, inhibitors and substrates, from the ChEMBL, DrugBank, and PubChem databases. We performed curation of the collected compounds (see *Method* Details) and finally, 89 actives (listed in the Supplemental Information (SI) and 450 decoys were retained for docking and ML datasets. Approximately 5 times more decoys than actives were used due to the lack of experimentally validated inactive molecules in the dataset. The training and external test sets were constructed by randomly dividing both the final actives and decoys, according to a ratio of 70 % and 30 %, respectively.

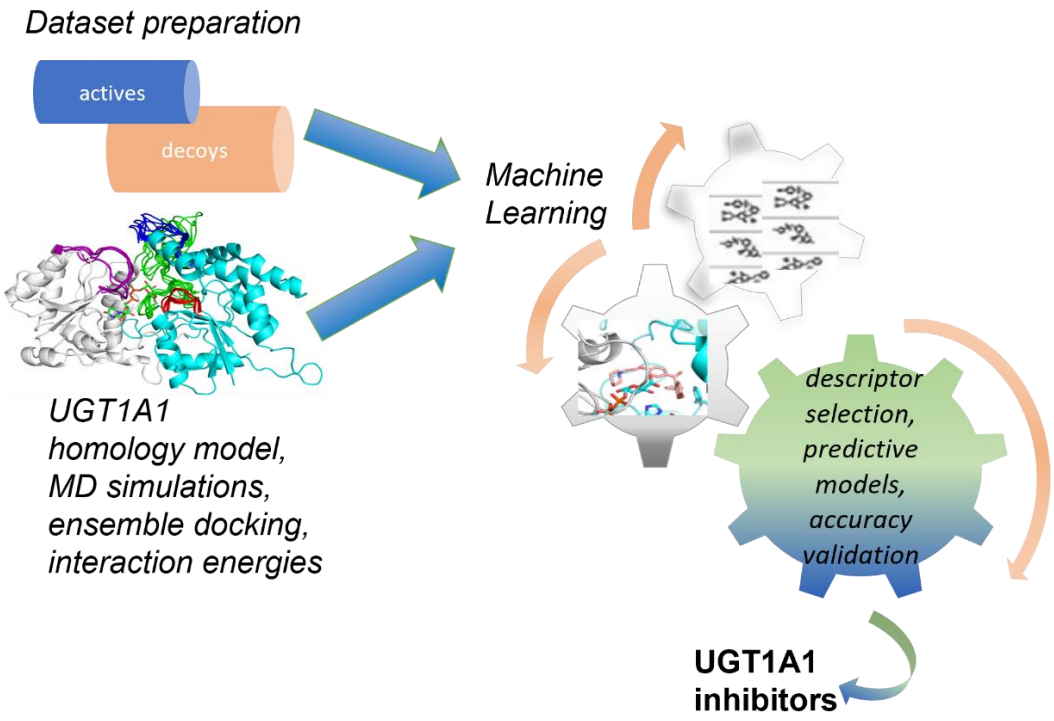

Figure B.2: Workflow of the models' development including datasets preparation, UGT1A1 homology modeling, molecular dynamics simulations, docking-scoring and machine learning to train different models for the prediction of UGT1A1 inhibitors.

We performed MD simulations of UGT1A1 to address its conformational flexibility and the large substrate spectrum of the enzyme. The homology model of UGT1A1 with the bound cofactor UDPGA built by Locuson et al. (Locuson and Tracy, 2007) was used as a starting structure for our MD simulations. Selected MD conformations were used for a subsequent ensemble docking step with the active UGT1A1 ligands. Finally, ML models were created to predict molecules inhibiting UGT1A1, incorporating interactions with six UGT1A1 conformations selected after the ensemble docking.

2.1. Molecular dynamics simulations

To address the degree of structural flexibility and the conformational adaptation of the binding pocket in light of the structural variety of the active compounds, we performed three 100-ns long MD simulations in the presence of the cofactor. Root Mean Square Deviation (RMSD) was calculated over time with respect to the starting structure to monitor conformational evolution (FIGURE B.3A).

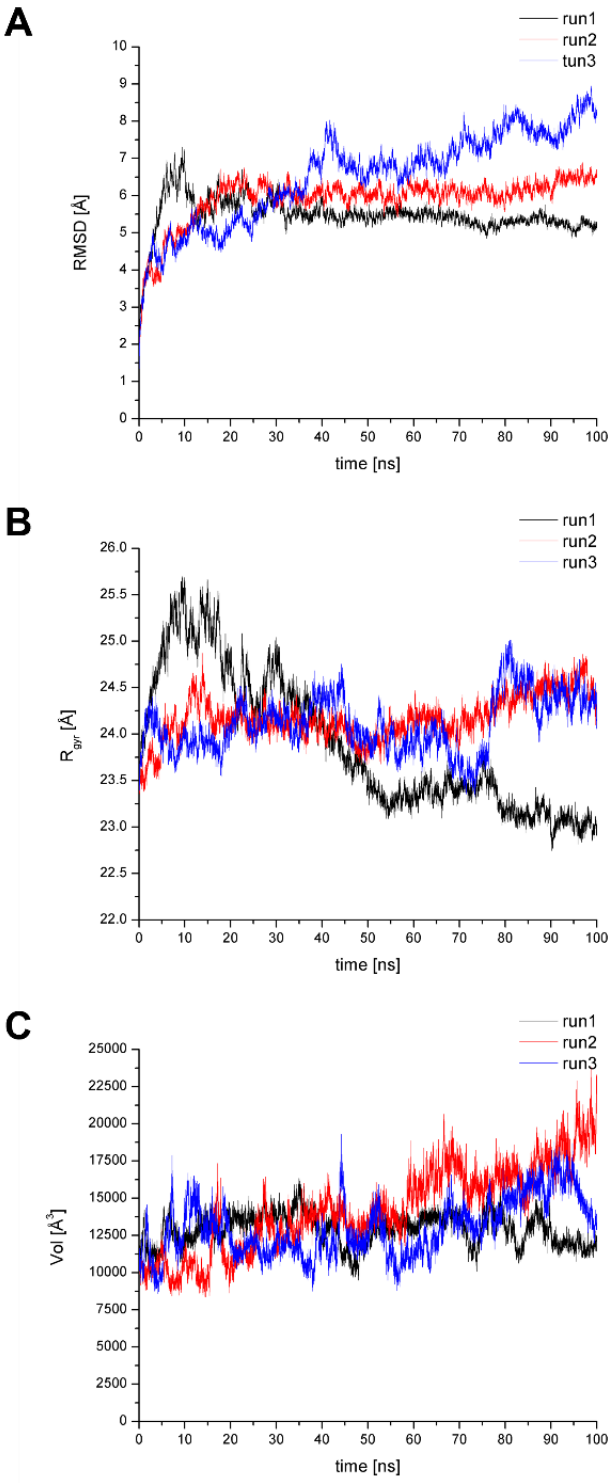

Figure B.3: Monitored parameters during the three 100-ns long MD simulations. ***(A)*** *Root Mean Square Deviation (RMSD) of the backbone atoms with respect to the starting structure.* ***(B)*** *Radius of gyration (Rgyr).* ***(C)*** *the volume of the substrate binding pocket.*

All three runs quickly diverted from the initial conformation reaching relatively high differences (up to 9 Å). The radius of gyration was calculated to examine further the behavior of the enzyme dynamics (FIGURE B.3B). Alteration in the radius of gyration generally corresponds to the overall changes in compactness. For the first MD run, we observed variations as high as 2.5 Å, which, together with a visual inspection of the trajectory, revealed larger interdomain opening-closing motions (dissociation/approaching of the domain tips). To further investigate the underlying conformational changes, especially focusing on the catalytically important pocket regions, we monitored the variations in the substrate-binding pocket volume (FIGURE B.3C, for its definition, see the list of residues in SI. In some conformations, its volume reached 1.5 to 2 times the size of the starting structure. The large variations in the substrate-binding pocket volume and the opening towards the lumen can facilitate access to the catalytic site and accommodate the diverse substrates. RMSD-based clustering of the MD trajectories enabled the extraction of 57 enzyme conformations with diverse binding pockets (see Method Details).

2.2. Ensemble docking and MD-derived structures best retrieving the UGT1A1 binders

In order to select the protein conformations best distinguishing between active and inactive compounds, we performed virtual screening of the active and decoy molecules of the training set using docking-scoring with AutoDock Vina (Trott and Olson, 2010) into the 57 centroid conformations of UGT1A1. Enrichment curves representing the percentage of actives retrieved at a percentage of screened actives and decoys were calculated by retaining the best score of interaction energies (IE) computed by docking-scoring for each compound in each protein conformation. The area under the receiver operating characteristic curve (AUC) revealed six best UGT1A1 conformations (see SI Figure S1): MD6, MD7, MD47, MD52, MD53, and MD54. The computed IE scores for these six UGT1A1 conformations were then used as protein-ligand interactions-based descriptors for the ML modeling.

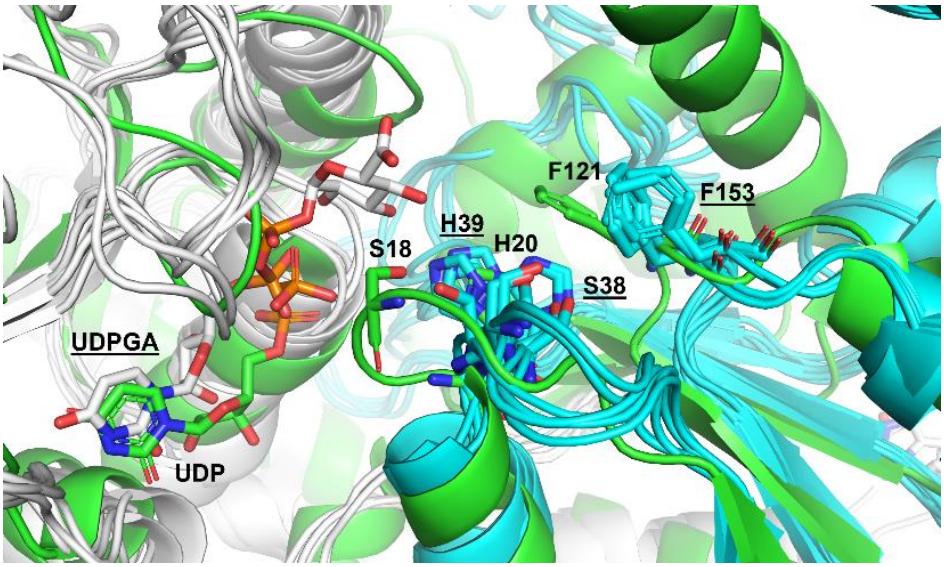

Figure B.4: The orientations of key enzymatic residues in the homologous plant flavonoid 3-O-glycosyltransferase (in green) and the six UGT1A1 conformations used in the ML models (cofactor-binding domains in white, substrate-binding domains in cyan). Residue labels belonging to the UGT1A1 conformations are underlined while labels corresponding to the plant homologue are not.

Interestingly, we found that even though there is a large flexibility of the substrate binding-pocket, in these six conformations, key residues for the catalytic reaction remained less flexible, and kept their orientation within the binding pocket. A key element of the enzymatic reaction of UGT1A1 is the deprotonation of the substrate by residue H39 for a nucleophilic attack (Li and Wu, 2007, Miley et al., 2007, Patana et al., 2008). This histidine, together with other catalytic residues, S38 and F153 (Offen et al., 2006), kept their backbone part rigid while their sidechains showed rotational flexibility, which enables some degree of freedom to adapt to the binding of the different ligands while maintaining the necessary catalytic configuration (FIGURE B.4). The position and orientation of these catalytic residues are very similar to what is found in the plant flavonoid 3-O-glycosyltransferase VvGT1 (PDB ID: 2C9Z) (Offen et al., 2006), where the corresponding residues are H20, S18, and F121, respectively. Other regions, especially loop segments at the entrance and the edge of the substrate-binding pocket, show considerable fluctuation among the 6 conformations, including residues 34-39, 99-109, 175-210, and 307-316 (FIGURE B.5). The flexibility of these loops also promotes the admission and the accommodation of the diverse ligands and further emphasizes the importance of considering enzyme dynamics in docking experiments, and therefore, in enzyme inhibition prediction studies.

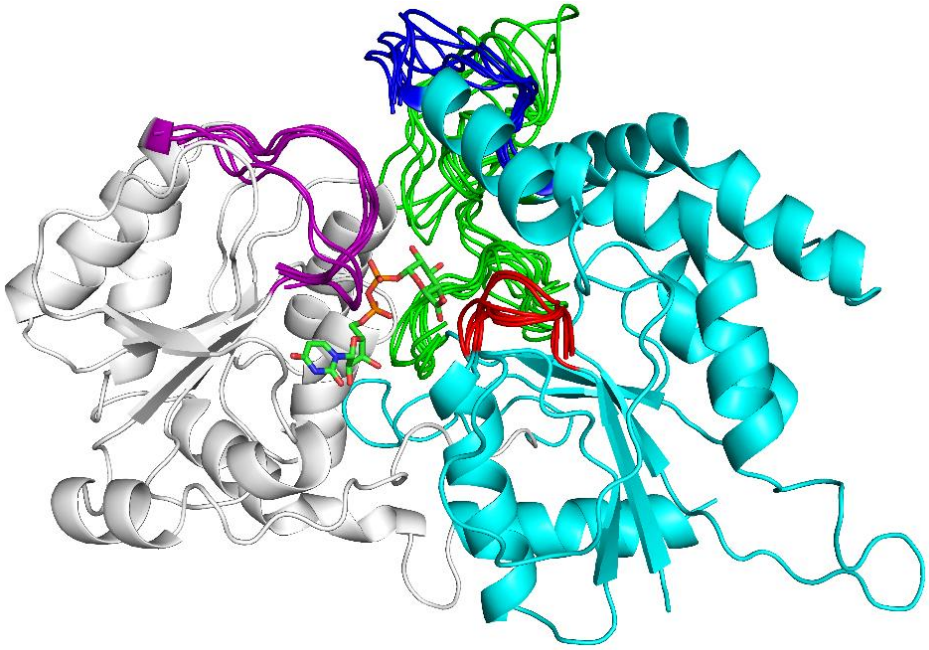

Figure B.5: Flexible loop regions of the six MD conformations of UGT1A1 at the entrance and the edge of the substrate-binding pocket, residues 34-39 (in red), 99-109 (in blue), 175-210 (in green), and 307-316 (in purple). The cofactor-binding domain is in white, the substrate-binding domain in cyan, the cofactor is in licorice representation.

2.3. Descriptor calculation and machine learning modeling

Then, we developed classification ML models for the prediction of UGT1A1 inhibitors. Physicochemical molecular descriptors of the training set's molecules were calculated using the MOE software . Initially, we calculated 354 2D and 3D MOE descriptors. Highly correlated descriptors with an absolute value of the Pearson correlation coefficient greater than or equal

to 0.85 and descriptors with near null variance were removed. This selection resulted in a total of 162 descriptors. The IE scores of the compounds of the training set calculated for the six selected UGT1A1 conformations were added as structure-based descriptors accounting for the protein-ligand interactions. To avoid overfitting and decrease the calculation time, we selected the best descriptors based on their relative importance in predicting the interaction with UGT1A1.

The selection comprised of building a number of Random Forest (RF) models on the training dataset and selecting the subset of descriptors with the highest Gini importance (Kantardzic, 2019). The Gini impurity index is a measure of the probability of incorrectly classifying a randomly selected element in a dataset. Thus, we performed 1000 RF runs with the 162 MOE and 6 IE descriptors with the default values of ntree, mtry, and sampsize (SI Table S1) to calculate the mean importance of the 168 descriptors, according to the diminution of the Gini criterion (see Method Details and SI Figure S2). The first 25 descriptors (including 4 IEs) were most important for the model performance (see SI Table S2). Then, the importance decreased slowly, and we decided to consider all the descriptors showing an importance greater than 0.5, including thus a total of 56 MOE and the 6 IEs.

The most important descriptors are related to polarity, lipophilicity, and charges. Principal component analysis (PCA) was performed on the 56 best MOE descriptors, and the training and the external test sets are shown in the subspace spanned by the first two PCs in FIGURE B.6. Overall, the training and test sets' compounds covered similar chemical space. Thus, our models are applicable within a domain given by the "soft" drug-like filter thresholds (see Method Details). Interestingly, even though our negative dataset contains decoys instead of real non-inhibitors, FIGURE B.6C shows that an important part of the actives is in different chemical space.

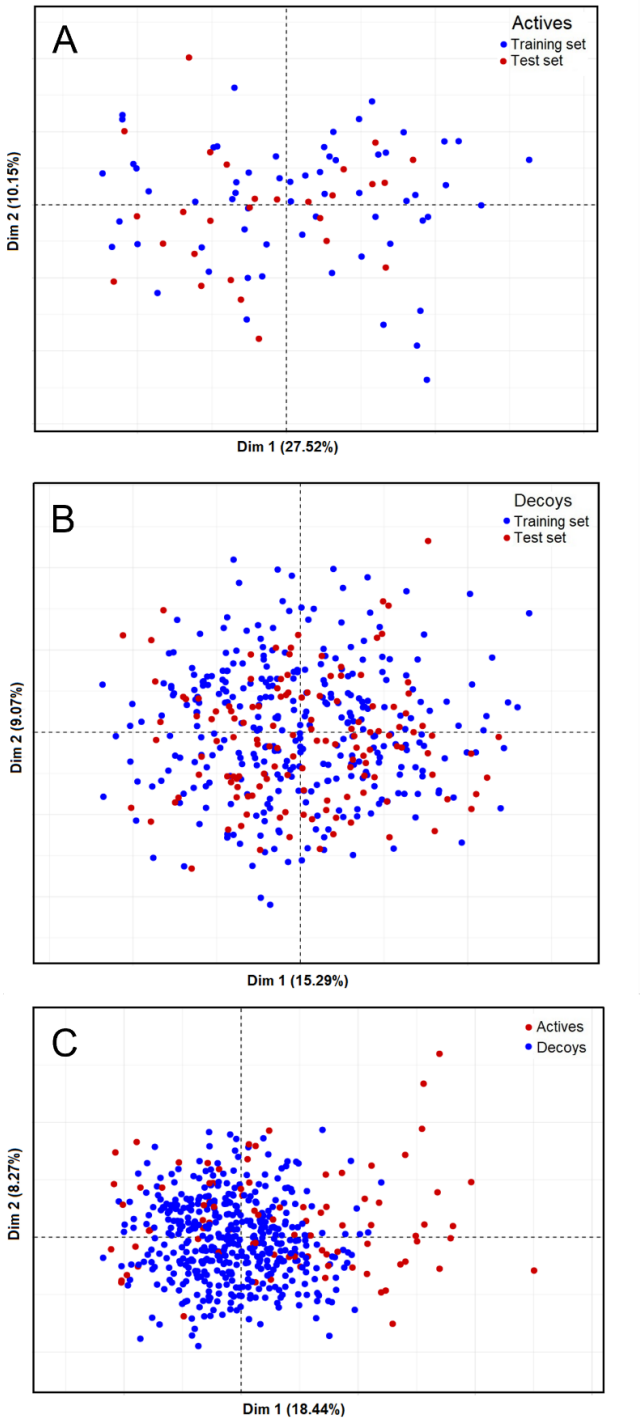

Figure B.6: Chemical space of the training and external test sets as described by the principal component analysis (PCA). The first two components, Dim 1 and Dim 2, and their representation in % of the total variance are indicated. ***A)*** *PCA of the actives of training vs. test sets.* ***B)*** *PCA of the decoys of training vs. test sets.* ***C)*** *PCA of all actives vs. decoys.*

2.4. Performance of the ML models in predicting binders of UGT1A1

Firstly, we created initial RF and Support Vector Machine (SVM) models using only MOE descriptors without parameter optimization. These preliminary RF and SVM models showed unsatisfactory performance (shown in SI Tables S3 and S4), in particular in terms of sensitivity (between 50% and 67% on the cross-validation (CV). Taking into account that our non-inhibitor molecules are decoys, the sensitivity performance is the most reliable assessment evaluation.

Due to the strong imbalance in our dataset between the number of active molecules and decoys, the optimization of hyper parameters (see SI Figure S3 and Method Details), including sample size and weight for RF and SVM, respectively, is critical to build robust predictive models.

Next, we built RF and SVM models: i) with the 168 descriptors, including 6 IEs; ii) with the best 62 descriptors, including 6 IEs, by optimizing the hyper parameters (see in Method Details and SI Figure S3 and Table S1). Cross-validation was applied for the RF and SVM modeling. The performances of the best RF and SVM models created and applied on the training and external validation test sets are summarized in Table B.1 and Table B.2, respectively. The area under the receiver operating characteristic curve (AUC), balanced accuracy (BA), sensitivity, specificity and Matthew's correlation coefficient were calculated. The AUC and BA values showed that all the RF and SVM models have excellent predictive powers for the discrimination of the UGT1A1 active molecules. The RF model with '56 MOE + 6 IE' descriptors showed better sensitivity on the cross-validation compared to the sensitivity of the '162 MOE + 6 IE' model.

Table B.1: Performances of the optimized RF models with MOE and IE descriptors on the training set (cross-validation CV) and the external validation test set.

Descriptors	Dataset	AUC %	BA %	Sensitivity %	Specificity %	MCC %
56 MOE + 6 IE	Internal CV	91.2	91.3	91.1	91.4	74.5
	External	93.7	93.7	92.6	94.8	81.8
162 MOE + 6 IE	Internal CV	90.6	90.2	88.0	92.4	74.1
	External	94.4	94.4	92.6	96.3	85.3

Table B.2: Performances of the optimized SVM models with MOE and IE descriptors on the training set (cross-validation CV) and the external validation test set.

Descriptors	Dataset	AUC %	BA %	Sensitivity %	Specificity %	MCC %
56 MOE + 6 IE	Internal CV	91.3	91.0	90.2	91.8	74.3
	External	90.7	90.7	88.9	92.5	74.5
162 MOE + 6 IE	Internal CV	90.2	88.9	88.1	89.7	68.7
	External	92.6	92.6	92.6	92.5	77.1

Similarly, the SVM models showed excellent performance, and the SVM model with '56 MOE + 6 IE' descriptors showed improved sensitivity and MCC on the cross-validation compared to those of the '162 MOE + 6 IE' model. Thus, our rational selection of the best 62 descriptors using the Gini index slightly improves the performance of the predictive models by diminishing the noise of the less discriminating descriptors; moreover, it also decreases the computational prediction time. The performance on the external test set was excellent, slightly better than the internal CV performance. Although the diversity was ensured between the molecules of the

training and the external test sets with a maximal chemical similarity of 0.80, the better performance on the external dataset may be due to the random choice of the molecules for the external set. Some over-performances on external datasets have also been observed in other ML modeling studies (Green et al., 2021, Goldwaser et al., 2022).

2.5. Binding positions of UGT1A1 ligands

Various small and bulky compounds are known to be metabolized by UGT1A1. The predicted binding positions of three different substrates of UGT1A1 as docked into the MD47 structure are shown in FIGURE B.7. Bilirubin was present in our training set. Quercetin and raloxifene, being in the external test sets, were successfully predicted to be binders of UGT1A1 by the two RF and the two SVM models, '56 MOE + 6 IE' and '162 MOE + 6 IE'. The poses were selected based on the best predicted IEs among the six different UGT1A1 MD conformations. The top scored poses of quercetin (IE = -8.0 kcal/mol), raloxifene (IE = -10.2 kcal/mol), and bilirubin (IE = -10.0 kcal/mol) are shown in FIGURE B.7.

Interestingly, the binding pose of quercetin corresponds to that of the crystal structure of quercetin bound to the plant flavonoid 3-O-glycosyltransferase VvGT1 (PDB IDs 2C9Z) (Offen et al., 2006). Based on the docking pose, we predicted that the binding of quercetin to UGT1A1 involves hydrogen bonds with H39 and S38, and aromatic interactions with F153, as in the crystal structure 2C9Z (H20, S18, F121). Similarly, raloxifene (its 6-O-glucuronidation site) (Guo et al., 2022) is in hydrogen bonding with H39 and S38 and in aromatic interactions with F153 in UGT1A1.

In the predicted pose of bilirubin, the two carboxylic groups that should be metabolized are in hydrogen bonding with the cofactor sugar group and the catalytic H39, respectively. The (-CH2-)2 side chain of the first propionic group is in hydrophobic contact with F153. The second propionic group is in an intramolecular hydrogen bonding, as in the solution structure of bilirubin (Nogales and Lightner, 1995), and similarly to bilirubin bound to other proteins (e.g. see in PDB structures IDs: 4I3D , 2VUE). V109 stabilizes the pyrrole group, and the two pyrroline cycles are stabilized by L175 and F181, and by P194 and F217, respectively. The docking pose of bilirubin suggests that it adopts a conformation similar to its structure in solution (Nogales and Lightner, 1995).

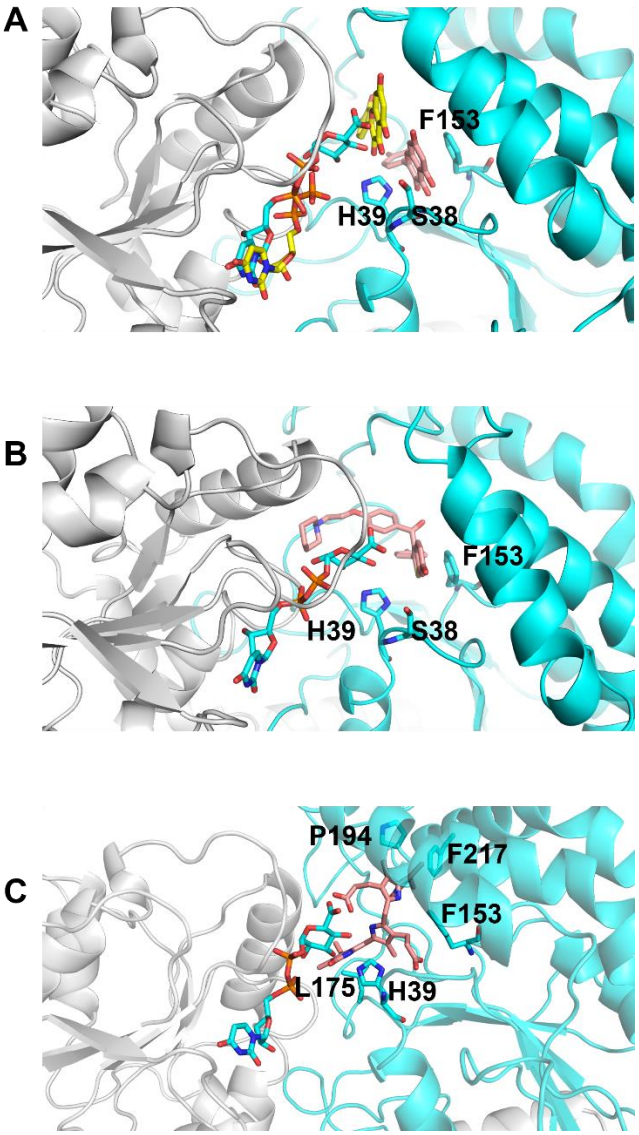

Figure B.7: Docking conformations of three substrates of UGT1A1 docked into MD47 (in cartoon, cofactor-binding domain in white, substrate-binding domain in cyan); UDPGA and key residues are shown in cyan sticks. A) The top scored pose of quercetin is shown in salmon stick. UDP and quercetin of the PDB structure of the homologous plant flavonoid 3-O-glycosyltransferase (2C9Z) are shown in yellow sticks. B) The top scored pose of raloxifene (in salmon stick). C) The top scored pose of bilirubin (in salmon stick).

To further investigate the predicted binding modes of raloxifene, bilirubin, and quercetin, additional 50-ns long MD simulations were performed starting from the docking complexes shown in FIGURE B.7. In the case of raloxifene, its 6-O-glucuronidation site remained in close contact with the catalytic S38, its benzothiophene part maintained a stable contact with F153 while its piperidine tail displayed more flexibility. At the beginning of the dynamics the piperidine tail of raloxifene tightly covered the cofactor whereas with time, it lifted slightly in direction of a sub-pocket of Q107, P194, F181 and F217 (see SI Figure S4A). Bilirubin also remained in the vicinity of F153 throughout the simulation maintaining their hydrophobic contacts. One of the carboxylic groups kept its orientation towards the sugar ring of the cofactor even though their distance slightly increased. Bilirubin also showed a similar movement towards this sub-pocket being in contacts with F181 and F217. Interestingly, the

above discussed configuration of bilirubin within the substrate-binding pocket, similar to what was observed in solution and bound to other proteins, was preserved throughout the entire simulation despite the large flexibility capability of bilirubin (see SI Figure S4B). Quercetin stayed in a stable position during the first 15 ns, primarily stabilized by the aromatic interactions with F153, after which it started shifting into the same sub-pocket of F181 and F217, distancing itself from the catalytic site (see SI Figure S4C).

3. Conclusions

In this study, we integrated structure-based modeling and machine learning techniques to build the first prediction models of UGT1A1 inhibition. We performed molecular dynamics simulations of the enzyme in the presence of the cofactor to gain insight into the structural variability of the catalytic site. We observed large conformational variability, which is crucial for accommodating the diverse substrates and inhibitors. RMSD-based clustering of the MD trajectories enabled us to extract a set of diverse enzyme conformations. Ensemble docking of experimentally validated active compounds and decoys identified 6 enzyme conformations that can efficiently differentiate between active and non-active compounds.

We found that while loop regions in the substrate-binding cavity exhibit large flexibility, the catalytically essential residues maintain their relative positions among the identified conformations. The docking of quercetin suggested that its catalytic pose within the substrate-binding pocket matches that experimentally found for the plant flavonoid 3-O-glycosyltransferase VvGT1 but quercetin moved a lot during the MD simulations. Bilirubin was stabilized by a hydrogen bond of one of its carboxylic groups that should be metabolized with the cofactor sugar group and hydrophobic contacts with F153.

We found that the contacts of the glucuronidation site of raloxifene with the catalytic residue S38, as well as its hydrophobic contacts with F153, remained stable during the simulations. The MD simulations with bound substrates suggested an additional sub-pocket in the area of F181 and F217 that could also be important for the wide substrate recognition and binding. Finally, we created ML models using RF and SVM techniques, integrating a rational selection of ligand-based descriptors together with information on the enzyme-ligand interactions.

The excellent performance of around 90 % accuracy and sensitivity obtained with the selected 56 MOE and 6 IE descriptors suggests that our models can be employed to identify new UGT1A1 inhibitors. To the best of our knowledge, the ML models reported here are the first for predicting UGT inhibition. They can be helpful for further prediction of drug-drug interactions of new drug candidates and safety treatments while also providing structural information on the enzyme-ligand interactions.

4. STAR Methods

4.1. Key Resources Table

Table B.3

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
ChEMBL database	ChEMBL Database - EMBL-EBI	https://www.ebi.ac.uk
DrugBank database	OMx Personal Health Analytic	https://go.drugbank.com
PubChem database	National Center for Biotechnology Information	https://pubchem.ncbi.nlm.nih.gov
Maybridge® HitFinder™ chemical library	ThermoFisher Sci.	http://www.maybridge.com
Software and algorithms		
CHARMM	Chemistry at HARvard Macromolecular Mechanics	https://www.charmm.org
CHARMM-GUI	Lehigh University	https://www.charmm-gui.org/
NAMD	Scalable Molecular Dynamics, University of Illinois	https://www.ks.uiuc.edu/Research/namd/
VMD v.1.3.9	Visual Molecular Dynamics, University of Illinois	http://www.ks.uiuc.edu
PropKa On-line	Alessandro Pedretti & Giulio Vistoli	*https://www.ddl.unimi.it*
AutoDockTools v.1.5.6	The Scripps Research Institute, CCSB	https://autodock.scripps.edu/
AutoDock Vina 1.1.2	The Scripps Research Institute, CCSB	https://vina.scripps.edu/
Pipeline Pilot - BIOVIA - Dassault Systèmes®, v.20.1	Discngine	https://www.3ds.com ›
CORINA Classic v.4.3	Molecular Networks	www.mn-am.com
MOE	Chemical Computing Group	https://www.chemcomp.com
FAF-Drugs4 server	RPBS platform	https://fafdrugs4.rpbs.univ-paris-diderot.fr/
R software v.3.5	The R Project for Statistical Computing	https://www.r-project.org

4.2. Detailed Methods

4.2.1. Protein structure preparation

The homology model of Locuson et al. (Locuson and Tracy, 2007) of the human UGT1A1 bound to the cofactor UDPGA was used as starting structure. UDPGA parameters were determined using CHARMM General Force Field (CGenFF) 2.5 (Vanommeslaeghe et al., 2010). The pKa values of the protein titratable groups were calculated with PROPKA (Sondergaard et al., 2011), and the protonation states were assigned at pH 7.0. The structure was solvated by CHARMM-GUI (Jo et al., 2008, Lee et al., 2016), and placed in a rectangular water box of TIP3 water molecules extending 15 Å in all directions from the protein surface (120 Å x 120 Å x 120 Å); the NaCl concentration was set to 0.15 M. The system was energy minimized using the steepest descent (SD) algorithm with harmonic restraints applied to the heavy atoms decreasing every 100 steps and adapting the values 50, 10, 1, and 0.1 kcal/mol/Å^2. Further minimization was performed without harmonic restraints in the form of successive cycles of SD and Adopted Basis Newton-Raphson (ABNR) minimizations until an RMS energy gradient tolerance of 0.01 kcal/mol/Å was met. Energy minimization was performed with CHARMM using the additive all-atom CHARMM force field C36m (Huang and MacKerell, 2013).

4.2.2. Molecular dynamics simulations

The system was equilibrated at 300 K for 100 ps in an NVT, then for 5 ns in an NpT ensemble at 1 atm pressure. Equilibration was performed with NAMD (Phillips et al., 2020) with the same force field, C36m. Langevin dynamics was used with a damping coefficient of 1 ps^{-1} for the constant temperature control. The Nose-Hoover method was used for the constant pressure control, with a piston oscillation period of 50 ps and a piston oscillation decay of 25 fs. The integration time step was 1 fs. The dielectric constant was set to 1 for energy evaluation. The particle mesh Ewald (PME) method was used to calculate electrostatic interactions with a grid spacing of 1 Å or less, having the order of 6. The real space summation was truncated at 12.0 Å, and the width of the Gaussian distribution was set to 0.34 Å^{-1}. Van der Waals interactions were reduced to zero by 'switch' truncation operating between 10.0 and 12.0 Å.

MD production runs were performed with NAMD. Three parallel 100-ns long MD simulations were run for the cofactor-bound UGT1A1 starting from the equilibrated conformation, with different random initial velocity distributions according to the Maxwell-Boltzmann distribution at 300 K. The integration time step was 2 fs; other parameters were identical to the 5 ns NpT equilibration run. The coordinates were saved every 5 ps, generating a total of 60 000 conformations.

Additional 50-ns long MD simulations were performed of the cofactor-bound UGT1A1 in the presence of different substrates, bilirubin, quercetin, and raloxifene, starting from the complexes retrieved from the docking simulations. The parameters of the substrates were determined by CGenFF 2.5. The same MD protocol was then applied as detailed above for the cofactor-bound UGT1A1 without bound substrate.

4.2.3. Clustering of the protein conformations

The Quality Threshold (QT) algorithm (Heyer et al., 1999), as implemented in VMD (Humphrey et al., 1996), was used to perform conformational clustering of the MD generated conformations. A distance function defined as the RMSD difference, calculated for the heavy atoms of the substrate-binding pocket, was used for clustering with a minimal distance of 1.5 Å. The centroid conformations of the 57 most populated clusters (numbered by the rank of population, i.e. the most populated cluster has the centroid conformation called MD1) covering 90 % of all the generated conformations were used in the subsequent docking simulations.

4.2.4. Dataset preparation

We collected 113 known ligands of UGT1A1, inhibitors and substrates, from the ChEMBL (ebi.ac.uk/chembl), DrugBank (go.drugbank.com), and PubChem (pubchem.ncbi.nlm.nih.gov) databases. Substrates were also included as they could cause concentration-dependent enzyme inhibition, a commonly observed phenomenon for metabolic enzymes (Wu, 2011). Among the collected 113 actives, 10 compounds had activity between 20 µM and 50 µM, and only 4 compounds had activity above 50 µM. To increase the applicability domain of our models, we thus decided to retain the compounds with activity below 50 µM (IC50). The four compounds showing very low activity (>50 µM) were not included into the dataset.

Decoys (putatively inactive molecules) for docking and ML model validation were taken from the diverse chemical compound collection Maybridge® HitFinder™ (maybridge.com), prepared as detailed in (Martiny et al., 2013). In order to build predictive models with applicability that covers drug-like molecules while maintaining chemical diversity, for all actives and decoys, we performed: i) filtering using the FAF-Drugs4 server (Lagorce et al., 2017) and an in-house developed 'soft' drug-like filter (molecular weight ≤ 1000 Da, number of H-bond donors ≤ 8, number of H-bond acceptors ≤ 12, number of rotatable bonds ≤ 20, logP between -7 and 10, and number of heteroatoms ≤ 15) without removing toxic/reactive/PAINS (Pan Assay Interference) compounds; ii) diversity clustering using FCFP_4 with a Tanimoto similarity criterion of 0.8 as implemented in Pipeline Pilot v.20.1 . The 3D structures of the compounds were generated using CORINA Classic v.4.3 , and the compounds were protonated at pH 7.0 using the FAF-Drugs4 server.

4.2.5. Ensemble docking

We performed docking simulations of the final dataset compounds into the centroid protein conformations of the 57 most populated clusters using the software AutoDock Vina 1.1.2 (Trott and Olson, 2010), which employs gradient-based conformational docking and an empirical scoring function predicting protein-ligand interaction energy (IE, in kcal/mol). The protein conformations and the ligands were pre-processed with AutoDockTools (Morris et al., 2009), the solvent molecules were removed, non-polar hydrogens were merged, and Gasteiger charges were assigned. A grid box of 24 Å x 20 Å x 22 Å was used with 1 Å spacing in the substrate-binding cavity. The maximum number of binding modes was set to 10, and the exhaustiveness of the global search to 8. The protein was kept rigid while the ligands were handled flexibly.

4.2.6. Machine learning classification modeling

Random forest (RF) classification (Breiman, 2001) was performed using the Random Forest R library (Liaw and Wiener, 2002) of the statistical software package R. Multiple decision trees were built with bootstrap samples from the training data. A small subset of descriptors was randomly selected to make decisions at each tree node to introduce diversity between the trees of the RF. The classification was obtained by taking the results of all the trees through a majority vote. To find the optimal size of the forest, '*ntree*' (number of trees), and the optimal number of descriptors, '*mtry*' (number of selected descriptors), for each model, we ran RF calculations performing an exhaustive nested-loop search of the *ntree* (128-1024) and *mtry* (5-50) parameters. As the dataset is imbalanced, the parameter '*sampsize*' (numbers of actives and decoys) was also optimized. We selected the combinations of *ntree*, *mtry,* and *sampsize* parameters for each model that yielded the best internal balanced accuracy (BA) while retaining the lowest acceptable *ntree* (see SI Table S1 and Figure S3). Five-fold cross-validation (CV) procedure was repeated ten times.

Support vector machine (SVM) approaches are based on the minimization principle from statistical learning theory and place data into hyperspaces through different kernel functions for its separation into datasets for classification or regression modeling (Cortes and Vapnik, 1995). For the nonlinearly separable cases, the kernel function allows SVM to transfer the data points into a higher-dimensional space where linear separation is possible. To build classification models, we also used the SVM algorithms implemented in the R package with the Caret library (Kuhn, 2008). The descriptors were centered around a mean of 0 and scaled to have a variance of 1. The radial basis function kernel (SVM-Rad) was used. The '*cost*' parameter was optimized in the range of 2^0 to 2^{18} through a five-fold cross-validation procedure that was repeated ten times. The best combinations of the hyperparameter *cost*, scaling function *gamma* (optimized in the range of 2^{-14} to 2^0), and '*weight*' are shown in SI Table S1 and Figure S3).

4.2.7. Assessment of the quality of the models

Different statistical quantities were evaluated to assess the predictive ability of the models. Sensitivity, or the true positive rate, is the fraction of true positives among all positively classified instances (Equation B.1), specificity is the true negative rate (Equation B.2), and balanced accuracy (BA) is an overall performance estimator used in the case of imbalanced datasets (Equation B.3). The area under the receiver operating characteristic curve (AUC) was also calculated. The AUC ranges from 0 to 1. Values of 0.8 or greater generally indicate good to excellent performance of a predictive model. The Matthew's correlation coefficient (MCC, Equation B.4) was calculated to measure the quality of binary classifications according to the following formulas:

$$Sensitivity = \frac{TP}{TP+FN} \qquad (B.1)$$

$$Specificity = \frac{TN}{TN+FP} \qquad (B.2)$$

$$Balanced\ Accuracy = \frac{Sensitivity+Specificity}{2} \quad (B.3)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}} \quad (B.4)$$

where *TP* and *TN* are the true positive and true negative, and *FP* and *FN* the false positive and false negative instances, respectively.

Limitations of Study

Our predictive models were developed based on a homology model of the human UGT1A1 structure containing the cofactor- and the substrate-binding domains. A future X-ray structure of the human UGT1A1 containing the two domains would be helpful to improve the performance of the models for the prediction of UGT1A1 inhibitors.

Acknowledgements

The authors thank Université Paris Cité, Inserm Instute, and Ecole Normale Supérieure Paris-Saclay for supporting this research. We thank S. Timouma for helpful discussion.

Author contributions

All authors contributed to the study conception and design. Material preparation and data collection were performed by B. D. and Y.B. All authors contributed to the analysis. The first draft of the manuscript was written by B. D. and Y. B. All authors read and approved the final manuscript.

Competing Interests

The authors declare no competing interests.

5. References

BIOVIA Pipeline Pilot, Release 2020, v.20.1. 20.1 ed.

CORINA Classic, Release 2019, v.4.3. Molecular Networks GmbH and Altamira, LLC.

Molecular Operating Environment (MOE), Release 2016. Chemical Computing Group Inc., 1010 Sherbooke St. West, Montreal, QC, Canada, H3A 2R7.

ALMAZROO, O. A., MIAH, M. K. & VENKATARAMANAN, R. 2017. Drug Metabolism in the Liver. *Clin Liver Dis,* 21, 1-20. doi:10.1016/j.cld.2016.08.001

BAUER, T. M., RITZ, R., HABERTHUR, C., HA, H. R., HUNKELER, W., SLEIGHT, A. J., SCOLLO-LAVIZZARI, G. & HAEFELI, W. E. 1995. Prolonged sedation due to accumulation of conjugated metabolites of midazolam. *Lancet,* 346, 145-7. doi:10.1016/s0140-6736(95)91209-6

BOSMA, P. J. 2003. Inherited disorders of bilirubin metabolism. *J Hepatol,* 38, 107-17. doi:10.1016/s0168-8278(02)00359-8

BREIMAN, L. 2001. Random Forests. *Machine Learning,* 45, 5-32. doi:10.1023/a:1010933404324

CAI, Y., YANG, H., LI, W., LIU, G., LEE, P. W. & TANG, Y. 2019. Computational Prediction of Site of Metabolism for UGT-Catalyzed Reactions. *J Chem Inf Model,* 59, 1085-1095. doi:10.1021/acs.jcim.8b00851

CARRACEDO-REBOREDO, P., LINARES-BLANCO, J., RODRIGUEZ-FERNANDEZ, N., CEDRON, F., NOVOA, F. J., CARBALLAL, A., MAOJO, V., PAZOS, A. & FERNANDEZ-LOZANO, C. 2021. A review on machine learning approaches and trends in drug discovery. *Comput Struct Biotechnol J,* 19, 4538-4558. doi:10.1016/j.csbj.2021.08.011

CORTES, C. & VAPNIK, V. 1995. Support-Vector Networks. *Machine Learning,* 20, 273-297. doi:10.1023/a:1022627411411

COURT, M. H. 2005. Isoform-selective probe substrates for in vitro studies of human UDP-glucuronosyltransferases. *Methods Enzymol,* 400, 104-16. doi:10.1016/S0076-6879(05)00007-8

GOLDWASER, E., LAURENT, C., LAGARDE, N., FABREGA, S., NAY, L., VILLOUTREIX, B. O., JELSCH, C., NICOT, A. B., LORIOT, M. A. & MITEVA, M. A. 2022. Machine learning-driven identification of drugs inhibiting cytochrome P450 2C9. *PLoS Comput Biol,* 18, e1009820. doi:10.1371/journal.pcbi.1009820

GOON, C. P., WANG, L. Z., WONG, F. C., THUYA, W. L., HO, P. C. & GOH, B. C. 2016. UGT1A1 Mediated Drug Interactions and its Clinical Relevance. *Curr Drug Metab,* 17, 100-6. doi:10.2174/1389200216666151103121253

GRANT, D. M. 1991. Detoxification pathways in the liver. *J Inherit Metab Dis,* 14, 421-30. doi:10.1007/BF01797915

GREEN, A. J., MOHLENKAMP, M. J., DAS, J., CHAUDHARI, M., TRUONG, L., TANGUAY, R. L. & REIF, D. M. 2021. Leveraging high-throughput screening data, deep neural networks, and conditional generative adversarial networks to advance predictive toxicology. *PLoS Comput Biol,* 17, e1009135. doi:10.1371/journal.pcbi.1009135

GUO, Y., SHAH, A., OH, E., CHOWDHURY, S. K. & ZHU, X. 2022. Determination of Acyl-, O-, and N-Glucuronide Using Chemical Derivatization Coupled with Liquid Chromatography-High-Resolution Mass Spectrometry. *Drug Metab Dispos,* 50, 716-724. doi:10.1124/dmd.122.000832

HEYER, L. J., KRUGLYAK, S. & YOOSEPH, S. 1999. Exploring expression data: identification and analysis of coexpressed genes. *Genome Res,* 9, 1106-15. doi:10.1101/gr.9.11.1106

HUANG, J. & MACKERELL, A. D., JR. 2013. CHARMM36 all-atom additive protein force field: validation based on comparison to NMR data. *J Comput Chem,* 34, 2135-45. doi:10.1002/jcc.23354

HUMPHREY, W., DALKE, A. & SCHULTEN, K. 1996. VMD: visual molecular dynamics. *J Mol Graph,* 14, 33-8, 27-8. doi:10.1016/0263-7855(96)00018-5

HWANG, S., SHIN, H. K., SHIN, S. E., SEO, M., JEON, H. N., YIM, D. E., KIM, D. H. & NO, K. T. 2020. PreMetabo: An in silico phase I and II drug metabolism prediction platform. *Drug Metab Pharmacokinet,* 35, 361-367. doi:10.1016/j.dmpk.2020.05.007

JO, S., KIM, T., IYER, V. G. & IM, W. 2008. CHARMM-GUI: a web-based graphical user interface for CHARMM. *J Comput Chem,* 29, 1859-65. doi:10.1002/jcc.20945

KAIVOSAARI, S., FINEL, M. & KOSKINEN, M. 2011. N-glucuronidation of drugs and other xenobiotics by human and animal UDP-glucuronosyltransferases. *Xenobiotica,* 41, 652-69. doi:10.3109/00498254.2011.563327

KANTARDZIC, M. 2019. *Data Mining: Concepts, Models, Methods, and Algorithms*, Wiley-IEEE Press.

KATO, H. 2020. Computational prediction of cytochrome P450 inhibition and induction. *Drug Metab Pharmacokinet,* 35, 30-44. doi:10.1016/j.dmpk.2019.11.006

KUHN, M. 2008. Building Predictive Models inRUsing thecaretPackage. *Journal of Statistical Software,* 28 doi:10.18637/jss.v028.i05

LAAKKONEN, L. & FINEL, M. 2010. A molecular model of the human UDP-glucuronosyltransferase 1A1, its membrane orientation, and the interactions between different parts of the enzyme. *Mol Pharmacol,* 77, 931-9. doi:10.1124/mol.109.063289

LAGORCE, D., BOUSLAMA, L., BECOT, J., MITEVA, M. A. & VILLOUTREIX, B. O. 2017. FAF-Drugs4: free ADME-tox filtering computations for chemical biology and early stages drug discovery. *Bioinformatics,* 33, 3658-3660. doi:10.1093/bioinformatics/btx491

LEE, J., CHENG, X., SWAILS, J. M., YEOM, M. S., EASTMAN, P. K., LEMKUL, J. A., WEI, S., BUCKNER, J., JEONG, J. C., QI, Y., JO, S., PANDE, V. S., CASE, D. A., BROOKS, C. L., 3RD, MACKERELL, A. D., JR., KLAUDA, J. B. & IM, W. 2016. CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations Using the CHARMM36 Additive Force Field. *J Chem Theory Comput,* 12, 405-13. doi:10.1021/acs.jctc.5b00935

LI, C. & WU, Q. 2007. Adaptive evolution of multiple-variable exons and structural diversity of drug-metabolizing enzymes. *BMC Evol Biol,* 7, 69. doi:10.1186/1471-2148-7-69

LI, J., OLALEYE, O. E., YU, X., JIA, W., YANG, J., LU, C., LIU, S., YU, J., DUAN, X., WANG, Y., DONG, K., HE, R., CHENG, C. & LI, C. 2019. High degree of pharmacokinetic compatibility exists between the five-herb medicine XueBiJing and antibiotics comedicated in sepsis care. *Acta Pharm Sin B,* 9, 1035-1049. doi:10.1016/j.apsb.2019.06.003

LIAW, A. & WIENER, M. 2002. Classification and Regression by randomForest. *R News,* 2, 18-22.

LIU, X. Y., LV, X., WANG, P., AI, C. Z., ZHOU, Q. H., FINEL, M., FAN, B., CAO, Y. F., TANG, H. & GE, G. B. 2019. Inhibition of UGT1A1 by natural and synthetic flavonoids. *Int J Biol Macromol,* 126, 653-661. doi:10.1016/j.ijbiomac.2018.12.171

LOCUSON, C. W. & TRACY, T. S. 2007. Comparative modelling of the human UDP-glucuronosyltransferases: insights into structure and mechanism. *Xenobiotica,* 37, 155-68. doi:10.1080/00498250601129109

LV, X., XIA, Y., FINEL, M., WU, J., GE, G. & YANG, L. 2019. Recent progress and challenges in screening and characterization of UGT1A1 inhibitors. *Acta Pharm Sin B,* 9, 258-278. doi:10.1016/j.apsb.2018.09.005

MAO, J., AKHTAR, J., ZHANG, X., SUN, L., GUAN, S., LI, X., CHEN, G., LIU, J., JEON, H. N., KIM, M. S., NO, K. T. & WANG, G. 2021. Comprehensive strategies of machine-learning-based quantitative structure-activity relationship models. *iScience,* 24, 103052. doi:10.1016/j.isci.2021.103052

MARTINY, V. Y., CARBONELL, P., CHEVILLARD, F., MOROY, G., NICOT, A. B., VAYER, P., VILLOUTREIX, B. O. & MITEVA, M. A. 2015. Integrated structure- and ligand-based in silico approach to predict inhibition of cytochrome P450 2D6. *Bioinformatics,* 31, 3930-7. doi:10.1093/bioinformatics/btv486

MARTINY, V. Y., CARBONELL, P., LAGORCE, D., VILLOUTREIX, B. O., MOROY, G. & MITEVA, M. A. 2013. In silico mechanistic profiling to probe small molecule binding to sulfotransferases. *PLoS One,* 8, e73587. doi:10.1371/journal.pone.0073587

MEECH, R. & MACKENZIE, P. I. 1997. Structure and function of uridine diphosphate glucuronosyltransferases. *Clin Exp Pharmacol Physiol,* 24, 907-15. doi:10.1111/j.1440-1681.1997.tb02718.x

MILEY, M. J., ZIELINSKA, A. K., KEENAN, J. E., BRATTON, S. M., RADOMINSKA-PANDYA, A. & REDINBO, M. R. 2007. Crystal structure of the cofactor-binding domain of the human phase II drug-metabolism enzyme UDP-glucuronosyltransferase 2B7. *J Mol Biol,* 369, 498-511. doi:10.1016/j.jmb.2007.03.066

MINERS, J. O. & MACKENZIE, P. I. 1991. Drug glucuronidation in humans. *Pharmacol Ther,* 51, 347-69. doi:10.1016/0163-7258(91)90065-t

MORRIS, G. M., HUEY, R., LINDSTROM, W., SANNER, M. F., BELEW, R. K., GOODSELL, D. S. & OLSON, A. J. 2009. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J Comput Chem,* 30, 2785-91. doi:10.1002/jcc.21256

NOGALES, D. & LIGHTNER, D. A. 1995. On the structure of bilirubin in solution. 13C[1H] heteronuclear Overhauser effect NMR analyses in aqueous buffer and organic solvents. *J Biol Chem,* 270, 73-7. doi:10.1074/jbc.270.1.73

ODA, S., FUKAMI, T., YOKOI, T. & NAKAJIMA, M. 2015. A comprehensive review of UDP-glucuronosyltransferase and esterases for drug development. *Drug Metab Pharmacokinet,* 30, 30-51. doi:10.1016/j.dmpk.2014.12.001

OFFEN, W., MARTINEZ-FLEITES, C., YANG, M., KIAT-LIM, E., DAVIS, B. G., TARLING, C. A., FORD, C. M., BOWLES, D. J. & DAVIES, G. J. 2006. Structure of a flavonoid glucosyltransferase reveals the basis for plant natural product modification. *EMBO J,* 25, 1396-405. doi:10.1038/sj.emboj.7600970

OHNO, S. & NAKAJIN, S. 2009. Determination of mRNA expression of human UDP-glucuronosyltransferases and application for localization in various human tissues by real-time reverse transcriptase-polymerase chain reaction. *Drug Metab Dispos,* 37, 32-40. doi:10.1124/dmd.108.023598

OSBORNE, R., THOMPSON, P., JOEL, S., TREW, D., PATEL, N. & SLEVIN, M. 1992. The analgesic activity of morphine-6-glucuronide. *Br J Clin Pharmacol,* 34, 130-8. doi:10.1111/j.1365-2125.1992.tb04121.x

PATANA, A. S., KURKELA, M., FINEL, M. & GOLDMAN, A. 2008. Mutation analysis in UGT1A9 suggests a relationship between substrate and catalytic residues in UDP-glucuronosyltransferases. *Protein Eng Des Sel,* 21, 537-43. doi:10.1093/protein/gzn030

PENG, J., LU, J., SHEN, Q., ZHENG, M., LUO, X., ZHU, W., JIANG, H. & CHEN, K. 2014. In silico site of metabolism prediction for human UGT-catalyzed reactions. *Bioinformatics,* 30, 398-405. doi:10.1093/bioinformatics/btt681

PHILLIPS, J. C., HARDY, D. J., MAIA, J. D. C., STONE, J. E., RIBEIRO, J. V., BERNARDI, R. C., BUCH, R., FIORIN, G., HENIN, J., JIANG, W., MCGREEVY, R., MELO, M. C. R., RADAK, B. K., SKEEL, R. D., SINGHAROY, A., WANG, Y., ROUX, B., AKSIMENTIEV, A., LUTHEY-SCHULTEN, Z., KALE, L. V., SCHULTEN, K., CHIPOT, C. & TAJKHORSHID, E. 2020. Scalable molecular dynamics on CPU and GPU architectures with NAMD. *J Chem Phys,* 153, 044130. doi:10.1063/5.0014475

PRUEKSARITANONT, T., CHU, X., GIBSON, C., CUI, D., YEE, K. L., BALLARD, J., CABALU, T. & HOCHMAN, J. 2013. Drug-drug interaction studies: regulatory guidance and an industry perspective. *AAPS J,* 15, 629-45. doi:10.1208/s12248-013-9470-x

RITTER, J. K., CHEN, F., SHEEN, Y. Y., TRAN, H. M., KIMURA, S., YEATMAN, M. T. & OWENS, I. S. 1992. A novel complex locus UGT1 encodes human bilirubin, phenol, and other UDP-glucuronosyltransferase isozymes with identical carboxyl termini. *J Biol Chem,* 267, 3257-61.

ROWLAND, A., MINERS, J. O. & MACKENZIE, P. I. 2013. The UDP-glucuronosyltransferases: their role in drug metabolism and detoxification. *Int J Biochem Cell Biol,* 45, 1121-32. doi:10.1016/j.biocel.2013.02.019

SHIMADA, T. 2006. Xenobiotic-metabolizing enzymes involved in activation and detoxification of carcinogenic polycyclic aromatic hydrocarbons. *Drug Metab Pharmacokinet,* 21, 257-76. doi:10.2133/dmpk.21.257

SONDERGAARD, C. R., OLSSON, M. H., ROSTKOWSKI, M. & JENSEN, J. H. 2011. Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of pKa Values. *J Chem Theory Comput,* 7, 2284-95. doi:10.1021/ct200133y

SORICH, M. J., MCKINNON, R. A., MINERS, J. O., WINKLER, D. A. & SMITH, P. A. 2004. Rapid prediction of chemical metabolism by human UDP-glucuronosyltransferase isoforms using quantum chemical descriptors derived with the electronegativity equalization method. *J Med Chem,* 47, 5311-7. doi:10.1021/jm0495529

SORICH, M. J., SMITH, P. A., MINERS, J. O., MACKENZIE, P. I. & MCKINNON, R. A. 2008. Recent advances in the in silico modelling of UDP glucuronosyltransferase substrates. *Curr Drug Metab,* 9, 60-9. doi:10.2174/138920008783331167

STEVENTON, G. 2020. Uridine diphosphate glucuronosyltransferase 1A1. *Xenobiotica,* 50, 64-76. doi:10.1080/00498254.2019.1617910

TESTA, B., PEDRETTI, A. & VISTOLI, G. 2012. Reactions and enzymes in the metabolism of drugs and other xenobiotics. *Drug Discov Today,* 17, 549-60. doi:10.1016/j.drudis.2012.01.017

TROTT, O. & OLSON, A. J. 2010. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem,* 31, 455-61. doi:10.1002/jcc.21334

TUKEY, R. H. & STRASSBURG, C. P. 2000. Human UDP-glucuronosyltransferases: metabolism, expression, and disease. *Annu Rev Pharmacol Toxicol,* 40, 581-616. doi:10.1146/annurev.pharmtox.40.1.581

VANOMMESLAEGHE, K., HATCHER, E., ACHARYA, C., KUNDU, S., ZHONG, S., SHIM, J., DARIAN, E., GUVENCH, O., LOPES, P., VOROBYOV, I. & MACKERELL, A. D., JR. 2010. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J Comput Chem,* 31, 671-90. doi:10.1002/jcc.21367

WILLIAMS, J. A., HYLAND, R., JONES, B. C., SMITH, D. A., HURST, S., GOOSEN, T. C., PETERKIN, V., KOUP, J. R. & BALL, S. E. 2004. Drug-drug interactions for UDP-glucuronosyltransferase substrates: a pharmacokinetic explanation for typically observed low exposure (AUCi/AUC) ratios. *Drug Metab Dispos,* 32, 1201-8. doi:10.1124/dmd.104.000794

WU, B. 2011. Substrate inhibition kinetics in drug metabolism reactions. *Drug Metab Rev,* 43, 440-56. doi:10.3109/03602532.2011.615320

ZHANG, L., ZHU, L., QU, W., WU, F., HU, M., XIE, W., LIU, Z. & WANG, C. 2020. Insight into tartrate inhibition patterns in vitro and in vivo based on cocrystal structure with UDP-glucuronosyltransferase 2B15. *Biochem Pharmacol,* 172, 113753. doi:10.1016/j.bcp.2019.113753

C. ABCG2

Drug transporters influence the disposition of a large number of drugs and drug candidates and are involved in drug-drug interactions (DDI). ATP-binding cassette (ABC) transporters harvest the energy of ATP hydrolysis in order to selectively translocate a variety of substrates across membranes. They are physiologically expressed in various tissues where they influence the absorption, distribution, and excretion of drugs. In particular, ABCG2 is a key player in preventing the absorption of toxic compounds from the gut, and it also plays an essential protective role at different tissue barriers like the maternal-fetus barrier, the blood-brain barrier, and the blood-testis barrier. ABCG2 transports a wide variety of drugs and also many phase II metabolites such as sulfate or glucuronide conjugates. The exact mechanisms of the ABC transporter-mediated substrate translocation are not fully understood.

The following chapter on the ABCG2 transport mechanism investigates the conformational transitions underlying its transport cycle and the substrate-transporter interactions along the translocation pathway. The timescale of a complete ABCG2 transport cycle falls in the range of a fraction of seconds or beyond, classical MD simulations fall short of providing a full atomic description of such cooperative events due to their time limitation.

As part of my PhD work, I have developed an enhanced MD simulation tool (kinetically excited targeted MD) and applied it to the two extreme conformational states of ABCG2 to elucidate possible conformational transition pathways. The translocation of the substrate is closely monitored, its interactions with transporter residues along its path are investigated. Its behavior is addressed in the second binding cavity about which little is known due to the lack of experimental structures. The different transport stages are further analyzed using classical MD and normal mode analysis.

In the near future, the generated transient conformations and predicted binding affinities towards them will serve as the basis for the creation of machine learning prediction models of ABCG2 substrates and inhibitors. The models will incorporate information on ligand-transporter interactions along the translocation pathway, and will be created in an attempt to overcome multidrug resistance and predict possible ABCG2-involved drug-drug interactions.

ABCG2/BCRP transport mechanism revealed through kinetically excited targeted molecular dynamics simulations

B. Dudas[a,b], X. Decleves[c,d], S. Cisternino[c,e], D. Perahia[b,*], M. A. Miteva[a,*]

[a]Inserm U1268 MCTR, CiTCoM UMR 8038 CNRS - Université Paris Cité, Paris, France
[b]Laboratoire de Biologie et Pharmacologie Appliquée, Ecole Normale Supérieure Paris-Saclay, Gif-sur-Yvette, France
[c]Inserm UMRS 1144, Optimisation Thérapeutique en Neuropsychopharmacologie - Université Paris Cité, Paris, France
[d]Biologie du Médicament et Toxicologie, Assistance Publique Hôpitaux de Paris, AP-HP, Hôpital Universitaire Cochin, Paris, France
[e]Service Pharmacie, Assistance Publique Hôpitaux de Paris, AP-HP, Hôpital Universitaire Necker-Enfants Malades, Paris, France

[*]corresponding authors: maria.mitev@inserm.fr, david.perahia@ens-paris-saclay.fr

Published in Computational Structural Biotechnology Journal on 2022 Jul 29
doi: 10.1016/j.csbj.2022.07.035

Abstract

ABCG2/BCRP is an ABC transporter that plays an important role in tissue protection by exporting endogenous substrates and xenobiotics. ABCG2 is of major interest due to its involvement in multidrug resistance (MDR), and understanding its complex efflux mechanism is essential to preventing MDR and drug-drug interactions (DDI). ABCG2 export is characterized by two major conformational transitions between inward- and outward-facing states, the structures of which have been resolved. Yet, the entire transport cycle has not been characterized to date. Our study bridges the gap between the two extreme conformations by studying connecting pathways. We developed an innovative approach to enhance molecular dynamics simulations, 'kinetically excited targeted molecular dynamics', and successfully simulated the transitions between inward- and outward-facing states in both directions and the transport of the endogenous substrate estrone 3-sulfate. We discovered an additional pocket between the two substrate-binding cavities and found that the presence of the substrate in the first cavity is essential to couple the movements between the nucleotide-binding and transmembrane domains. Our study shed new light on the complex efflux mechanism, and we provided transition pathways that can help to identify novel substrates and inhibitors of ABCG2 and probe new drug candidates for MDR and DDI.

Keywords: ABC transporters, BCRP, ABCG2, efflux mechanism, molecular dynamics simulations, drug-drug interactions

1. Introduction

ATP-binding cassette (ABC) transporters are molecular machineries that harvest energy from ATP hydrolysis to translocate substrates across membranes selectively (Thomas and Tampe, 2020). Some members of the ABCB, ABCC, and ABCG subfamilies are involved in drug transport and are responsible for unidirectional drug efflux. They are of major interest due to their involvement in the multidrug resistance (MDR) phenotype of tumor cells as well as the controlling of drug pharmacokinetics at several critical body interfaces, considering their physiological expression in cells like endothelial brain cells and enterocytes (Chapy et al., 2016, Cisternino et al., 2004). Furthermore, inhibition of ABC transporters and drug metabolizing enzymes (Sun and Scott, 2010, Martiny et al., 2015, Goldwaser et al., 2022) can lead to drug-drug interactions (DDI) and influence drug efficacy and safety (Brozik et al., 2011).

Human ABCG2, also known as BCRP (Breast Cancer Resistance Protein), belongs to the G-subfamily of ABC transporters and is physiologically expressed in tissue barriers like the blood-brain barrier (Fetsch et al., 2006, Robey et al., 2009, Thomas and Tampe, 2020, Maliepaard et al., 2001, Chapy et al., 2016). It plays an important role in tissue protection by selectively exporting numerous endogenous substrates and a broad variety of xenobiotics to extracellular spaces like the blood lumen at the blood-brain barrier (Suzuki et al., 2003, Imai et al., 2003, Mao and Unadkat, 2015). Similar to P-glycoprotein (ABCB1) and MRPs (ABCCs), ABCG2 has also been identified as a contributor to MDR in tumor cells (Diestra et al., 2002, Gillet and Gottesman, 2011, Gottesman et al., 2002, Mo and Zhang, 2012). ABCG2 can strongly influence the pharmacokinetic profile of a wide range of drugs due to its substrate poly-specificity. Interestingly, ABCG2 substrates comprise a broad spectrum of anticancer agents, sulfate and glucuronide conjugates of sterols and drugs that are common products of mammalian Phase II metabolism (Mo and Zhang, 2012). Therefore, drug agencies worldwide (e.g. the European Medicines Agency and the United States Food and Drug Administration) recommended testing for possible ABCG2 substrate or inhibitor status over the course of drug development (Toyoda et al., 2019, Hillgren et al., 2013, Prueksaritanont et al., 2013). It is crucial to understand the molecular mechanism of the underlying ABCG2 substrate export in all its complexity to better predict and prevent ABCG2-involved drug pharmacokinetic variability.

Conformational changes are driving forces for the substrate efflux in ABC transporters (Manolaridis et al., 2018, Oldham et al., 2008, Jones and George, 2004). Over recent years, thanks to breakthrough advances in single-particle cryogenic electron microscopy (cryo-EM), several transporter structures have been resolved at a nearly atomic resolution under different conditions (Taylor et al., 2017, Jackson et al., 2018, Manolaridis et al., 2018, Orlando and Liao, 2020, Kowal et al., 2021, Yu et al., 2021). These recent studies have identified two distinct conformational clusters of ABCG2, the transporter in the inward facing state (IFS) and the outward facing state (OFS). During the transport cycle, ABCG2 is thought to cycle between these two states (Orlando and Liao, 2020). ABCG2 functions as a homodimer, with each monomer consisting of a nucleotide-binding domain (NBD) and an integral transmembrane domain (TMD) (FIGURE C.1). NBDs contain highly conserved motifs shared among ABC

transporters and can bind two ATP molecules and coordinating Mg^{2+} ions at their dimer interface. TMDs are involved in substrate recognition by forming two substrate-binding cavities (FIGURE C.1A). Substrates have access to cavity 1 from both the cytosol and the lipid bilayer. As opposed to cavity 1, cavity 2 faces the extracellular space, and the two cavities are separated by the so-called leucine gate (also referred to as the leucine plug) (Manolaridis et al., 2018).

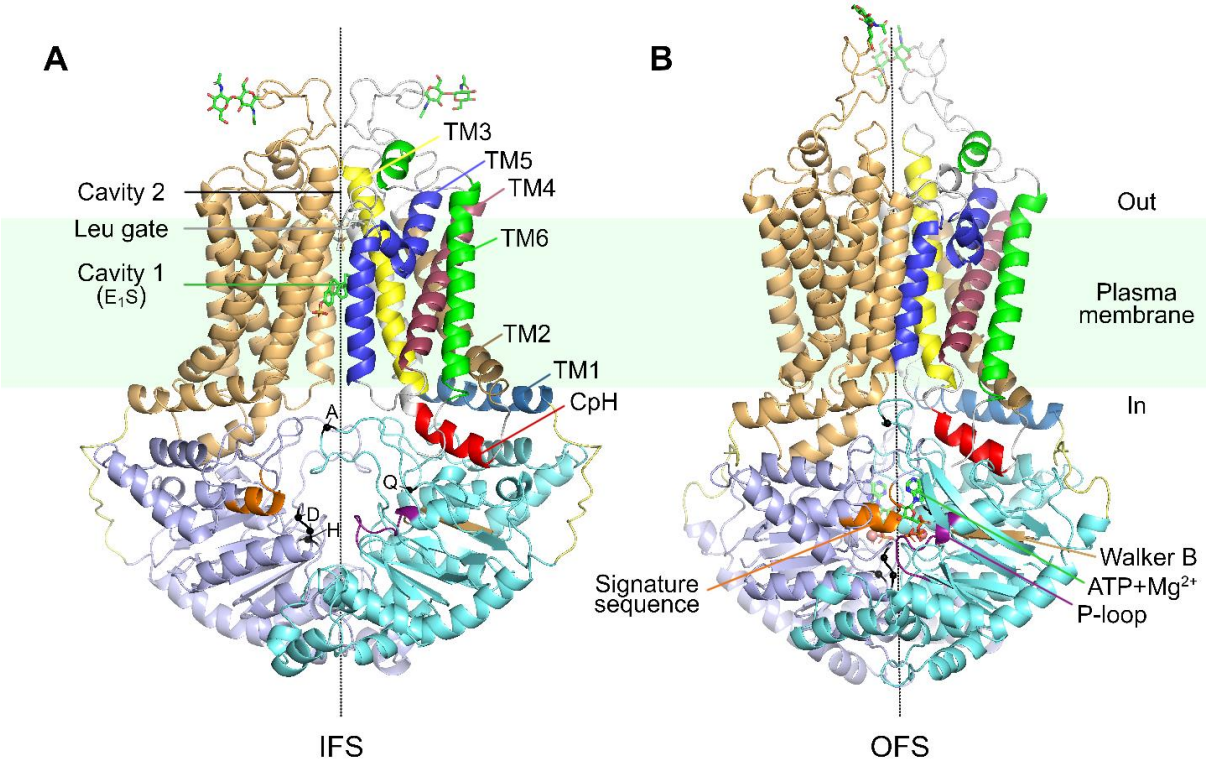

*Figure C.1: Experimental ABCG2 structure in (**A**) the E_1S substrate-bound IFS (PDB 6HCO) and (**B**) the ATP-Mg^{2+}-bound OFS (PDB 6HBU). The loop regions modelled here are shown for clarity. The rotational symmetry axis of the homodimer is indicated by a dashed line. Each monomer consists of a TMD and an NBD (e.g. TMD in light orange and NBD in light blue of one monomer). Conserved motifs within the NBDs are marked with letters (A-loop, Q-loop, D-loop, and H-loop). The "coupling helix" (CpH) of one monomer is highlighted in red, the different TM helices are highlighted in different colors. The linker segments connecting the individual NBDs and TMDs are in pale yellow. The ATPs, the substrate, the leucine gate, and the glycosyl groups are in licorice, the Mg^{2+} ions in sphere representations. Signature sequence, P-loop and Walker B motif are also colored orange, purple, and tan, respectively.*

To date, neither experimental structures with a substrate bound to cavity 2, nor transient structures along the translocation pathway and the transport cycle have been resolved. Therefore, the transporter's dynamics, playing a key role in the complex mechanism of drug efflux, needs to be elucidated. *In silico* approaches, in particular Molecular Dynamics (MD) simulations, are powerful tools in the exploration of related mechanisms (Mortier et al., 2015, Salinas et al., 2021, Ghode et al., 2020, Danilowicz et al., 2017). Yet, classical MD simulations fall short of providing a full atomic description of cooperative events at time scales beyond microseconds, let alone the timeframe of the transport cycle, for multi-domain systems. Although Nagy et al. investigated key interactions along the uric acid substrate-translocation pathway and its regulation by cholesterol with the help of metadynamics simulations (Nagy et

al., 2021), the entire ABCG2 transport cycle has not been thoroughly understood, and the dynamical behavior of the different transport stages has not been characterized to date.

To better understand the molecular mechanism of substrate export in all its complexity, here we explore the ABCG2 transition pathways of the transport cycle. Our study bridges the gap between the different transport states by employing an innovative simulation approach, starting from available experimental structures. We developed an enhanced MD simulation methodology to trace possible pathways between two terminal structures, termed 'kinetically excited targeted Molecular Dynamics', and successfully simulated transitions between the IFS and the OFS in both directions, along with the translocation of the physiological estrone 3-sulfate (E_1S) substrate. Furthermore, we characterized the dynamical behavior of ABCG2 in the different transport stages.

2. Results and Discussion

2.1. Structural models and kinetically excited targeted MD

We performed simulations starting from cryo-EM structures (Manolaridis et al., 2018) of the human homodimer of ABCG2 in its IFS and OFS (see Materials and Methods for details). The structure of ABCG2 contains highly conserved motifs shared among ABC transporters at their NBDs, such as the P-loop (Walker A motif), the Walker B motif, the signature sequence ('VSGGERKR'), and the A- and H-loops primarily responsible for ATP binding and hydrolysis, as well as the Q- and the D-loops responsible for NBD dimer formation or interdomain communication (Khunweeraphong and Kuchler, 2021). Two ATP molecules and coordinating Mg^{2+} ions have been found to bind symmetrically at the catalytic interface formed by the two NBDs, each between the P-loop of one monomer and the signature sequence of the other (Manolaridis et al., 2018). In the IFS, the two NBD monomers are partially separated, yet some contacts are maintained at the cytosolic tip of the transporter. The degree of NBD separation varies between the available experimental structures, from fully-inward open (e.g. nucleotide-free estrone 3-sulfate (E_1S) transporter (Manolaridis et al., 2018)) to more, but not completely closed states (e.g. E_1S or topotecan bound transporter in the presence of ATP (Yu et al., 2021)). The TMD pair forms the slit-like hydrophobic cavity 1, where the physiological E_1S substrate and various inhibitors have been proven to bind (Orlando and Liao, 2020, Yu et al., 2021, Jackson et al., 2018, Kowal et al., 2021, Manolaridis et al., 2018). In contrast, in the OFS, cavity 1 is completely collapsed, and the NBDs form a tightly closed interface (FIGURE C.1B).

We chose to model the unresolved flexible intracellular loop regions in the NBDs and include them in our simulations since they are likely to affect the substrate entry and may possess a similar gating function to analogous regions in bacterial transporters (Bi et al., 2018, Caffalette et al., 2019, Chen et al., 2020) (e.g. the loop region between the first and second NBD β-strands, residues 49-57). Similarly, we included the model of the linker segment, connecting individual NBDs to TMDs, as residues in this region have been shown to play a unique role in coupling ATP hydrolysis to substrate efflux, and the related conformational changes of the transporter (Macalou et al., 2016). Multiple systems were constructed from the experimental IFS and OFS structures: an apo IFS, a substrate-bound IFS, and a substrate and

ATP-Mg^{2+}-bound IFS transporter; and an ATP-Mg^{2+}-bound OFS, an ADP-bound OFS, and an OFS transporter with no nucleotide bound (see Materials and Methods and SI Table S1 for details).

The structures were inserted into a lipid bilayer composed of dimiristoyl-phosphatidylcholine (DMPC) with 20% cholesterol (CHOL); the latter has been suggested to play a role in the transport regulation of ABCG2 (Ferreira et al., 2017, Telbisz et al., 2007, Nagy et al., 2021). We performed classical MD simulations and Normal Mode Analysis (NMA) on all the above systems.

Moreover, we developed an innovative method, kinetically excited targeted MD (ketMD) that traces possible pathways between two terminal structures, that we applied here to simulate conformational transitions between the IFS and the OFS. Our concept relies on the method developed by Costa et al., Molecular Dynamics with excited Normal Modes (MDeNM) (Costa et al., 2015) designed to enhance protein conformational exploration. In MDeNM, collective motions of the protein described by different combinations of low frequency normal modes are kinetically activated during MD simulations. This enables the coupling of fast and slow degrees of freedom. Recently, we have successfully employed MDeNM to study large functional movements in several biological systems (Dudas et al., 2020, Dudas et al., 2021a) including the gating mechanism of substrate recognition in the sulfotransferase SULT1A1 (Dudas et al., 2021b).

As in the case of ABCG2 the target conformation is specified, the excitation vector was chosen to point towards the target structure instead of being a combination of normal modes, similar to targeted MD (tMD) simulations. However, unlike tMD, where the potential energy function is biased and the protein is guided by steering forces at each simulation step, ketMD relies on kinetic excitations.

At the first step of each excitation cycle, the velocity components pointing from the instantaneous conformation to the target structure are increased, allowing the crossing of larger energy barriers. This excitation step is followed by a relaxation period where no external perturbation is applied, the system can evolve, and the injected kinetic energy dissipates. After each excitation cycle, the excitation direction vector is updated to point to the target structure from the current conformation. In total, 40 consecutive excitation cycles were performed per system (see Materials and Methods for a detailed description).

2.2. Conformational transitions during the ABCG2 transport cycle

The transport cycle of ABCG2 includes two large conformational transitions. Firstly, transition 1, when the IFS transforms into the OFS while the substrate passes from cavity 1 to cavity 2 (from where it is then released to the extracellular space). Secondly, transition 2, when the OFS returns to the initial IFS. The timescale of a complete transport cycle of ABCG2 falls in a range of fraction of seconds or beyond (the initial transport rate of a substrate is 0.1 molecules per ABCG2 dimer per second in the study of Yu et al.) (Szollosi et al., 2018, Yu et al., 2021), a timeframe that currently cannot be simulated by classical MD. With the help of ketMD, here we present all-atom simulations of transitions 1 and 2 of the membrane-embedded ABCG2.

Transition 1 was simulated starting from the IFS with bound E_1S substrate (an endogenic steroid) and ATP-Mg^{2+}, transition 2 from the OFS without bound substrates and in the presence of ATP-Mg^{2+}, ADP or in the absence of bound nucleotides.

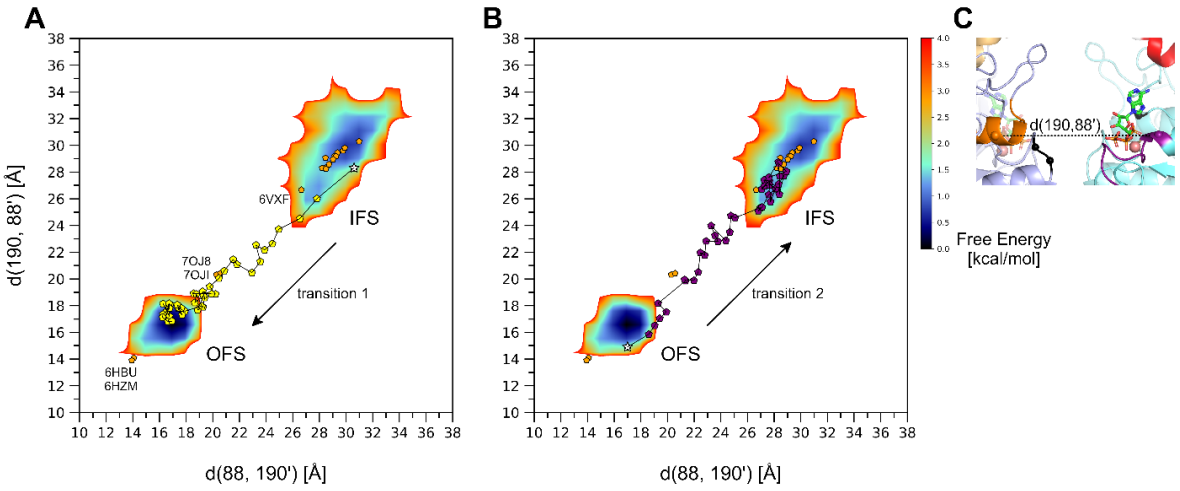

*Figure C.2 Evolution of the openness at the catalytic ATP-binding site upon (**A**) transition 1 (yellow pentagons) and (**B**) transition 2 (purple pentagons) of the ketMD simulations, represented by the distance between the Cα atoms of residues S88 of one monomer and E190 of the other. Free Energy Landscapes (FELs) of the MD-generated conformations starting from the E_1S- and ATP-Mg^{2+}bound IFS and the nucleotide-free OFS are included as references. The initial conformations are indicated as stars, available experimental structures are marked with orange pentagons. (**C**) The catalytic ATP-binding site and the monitored distance shown in the IFS state. The following experimental structures, which fall in the IFS region, are shown but not labelled in panels A and B : PDB 5NJ3, 6ETI, 6FEQ, 6FFC, 6HCO, 6HIJ, 6VXH, 6VXI, 6VXJ, 7NEQ, 7NEZ, 7NFD, 7OJH.*

2.3. Role of the NBDs

Upon the transition from the IFS to the OFS (transition 1), the two NBDs form a tightly packed dimer. The resulting interface establishes the two catalytic ATP-binding sites between the P-loop (residues 80-88) of one monomer and the signature sequence ('VSGGERKR', residues 186-193) of the other (Manolaridis et al., 2018). The formation of the two catalytic ATP-binding sites upon the transition of the NBDs can be monitored by the evolution of the distances between the residues at the edges of the P-loop of one monomer and the signature sequence of the other, symmetrically two distances, each corresponding to one of the two ATP-binding sites, namely d(88CA,190'CA) and d(190CA,88'CA) (FIGURE C.2C).

These distances gradually decrease during the ketMD simulation from the initial 30.6 Å and 28.3 Å to less than 17 Å (FIGURE C.2A). For reference, the distance is around 14 Å in the E211Q mutant ATP-Mg^{2+}-bound OFS target structure (PDB ID: 6HBU). This distance averages 15.4 Å across the three 100-ns-long MD runs starting from the wild-type, ATP-Mg^{2+}-bound OFS, with values greater than 19 Å present in the trajectories (FIGURE C.2A,B, Free Energy Landscape (FEL) of the MD generated conformations calculated based on Equation C.1 in Materials and Methods) suggesting that during the ketMD simulation of transition 1, the catalytic ATP-binding

sites were successfully formed between the P-loop and the signature sequence similarly to what can be observed in the reference OFS MD simulations.

The backbone RMSD (root mean square deviation) of the NBD dimer with respect to the target experimental OFS structure was also monitored during the ketMD simulation to follow the closure of the whole NBD region (SI Figure S1A). The initial RMSD of 7.2 Å gradually decreased to 2 Å during the 40 excitation cycles. As reference, the same RMSD among the classical MD generated OFS conformations is on average 1.8 Å with a standard deviation of 0.23 Å (SI Figure S1C). Based on these results and visual inspection of the generated conformations (FIGURE C.3A,B, FIGURE C.4A,B; SI Video S1 and Figure S1A,C), we conclude that a full NBDs transition was successfully achieved together with the catalytic ATP-binding site formation during the ketMD simulation of transition 1.

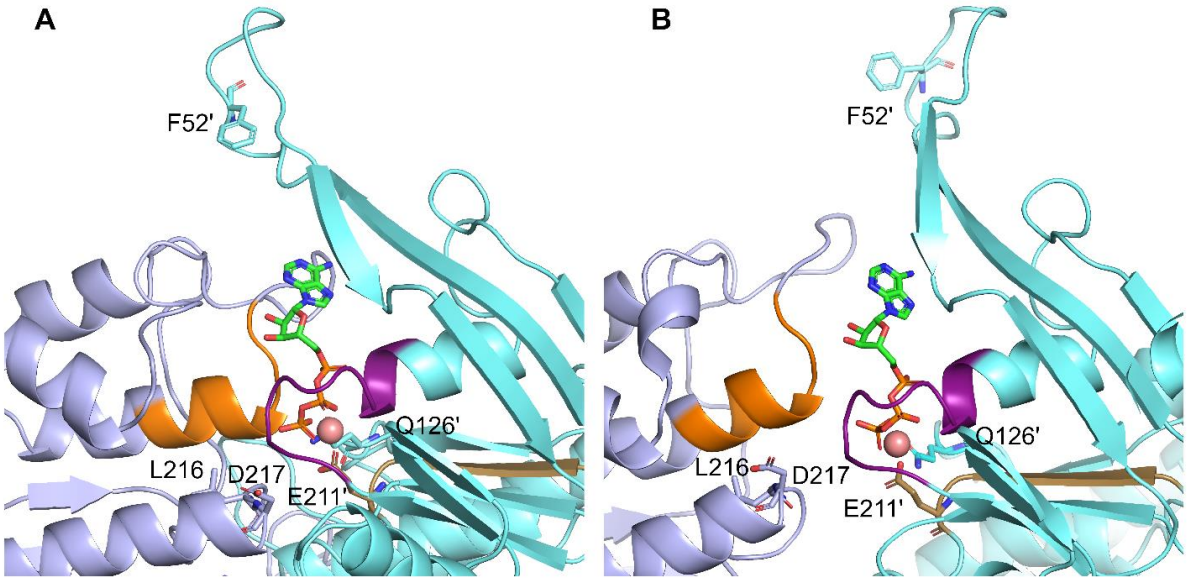

*Figure C.3: The nucleotide binding site in (**A**) the OFS cryo-EM structure (PDB 6HBU) and (**B**) at the end of the ketMD simulation of transition 1. P-loop is highlighted in purple, the signature sequence in orange, and the Walker B motif in tan. The ATP is in licorice, the Mg^{2+} ion in sphere representation.*

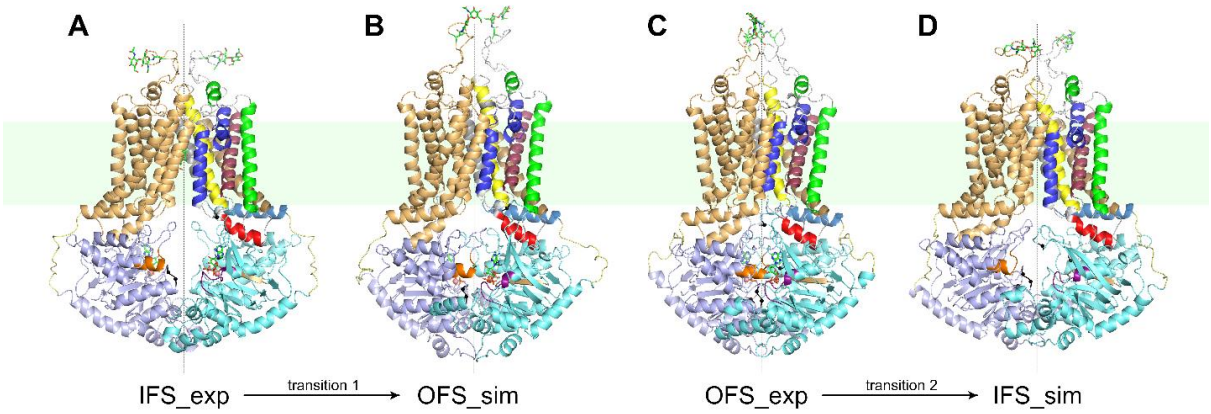

Figure C.4: Conformational transitions during the ketMD simulations. ***(A)*** *The experimental structure (PDB 6HCO) with the modelled missing loops and the added two ATP-Mg^{2+} that was used (after equilibration) as starting structure for the ketMD simulation of transition 1,* ***(B)*** *the final simulated conformation of transition 1,* ***(C)*** *the experimental structure (PDB 6HBU) with the modelled missing loops that was used (after equilibration) as starting structure for the ketMD simulations of transition 2 (either with bound ATP-Mg^{2+}, ADP, or no bound nucleotides),* ***(D)*** *the final simulated conformation of transition 1 (in the absence of bound nucleotides). The ATPs are in licorice, the Mg^{2+} ions in sphere representation. The rotational symmetry axis of the homodimer is indicated by dashed lines.*

In the opposite direction, upon the transition from the OFS to the IFS (transition 2), the strong interactions stabilizing the NBD dimer must be broken to obtain the partially separated NBDs, characteristic of the IFS. Some of the strongest interactions exist between the P-loop and D-loop (P81/T82-D217), the P-loop and the signature sequence (T82-R193), and the Q-loop and the signature sequence (D127-R191). The interaction energy between the 2 NBDs is approximately -320 kcal/mol for the MD equilibrated OFS (in the absence of the nucleotides) and -150 kcal/mol for the IFS conformation. Upon the partial separation of the NBDs during the ketMD simulations, we observe a continuous weakening of the interactions (less negative interaction energy), reaching the reference of -150 kcal/mol after the 25th excitation cycle.

Simultaneously, the distances d(88CA,190'CA) and d(190CA,88'CA) gradually increase from the initial 17 Å and 14.9 Å to over 28 Å (FIGURE C.2B), which demonstrates the dissociation of the catalytic ATP-binding sites. The backbone RMSD of the NBD dimer with respect to the target experimental IFS structure (PDB 6HCO) gradually decreased from the initial 7.2 Å to 1.3 Å during the ketMD simulation (SI Figure S1B). The same RMSD has a mean of 1.7 Å with a standard deviation of 0.27 Å among the classical MD generated IFS conformations (SI Figure S1D). Visual inspection of the ketMD generated conformations together with the analyses above clearly confirmed that the NBDs got partially separated and a complete NBDs transition occurred (FIGURE C.4C,D, SI Video S1 and Figure S1B,D).

To analyze the effect of the presence of ATP, ADP, or the absence of nucleotides on the dissociation of the catalytic ATP-binding sites, we also performed ketMD simulations of transition 2 in the presence of ATP or ADP. During equilibration, the distance between the P-loop and the signature sequence was preserved in the presence of ATP, while it increased

slightly in the presence of ADP and without nucleotides. Detaching the NBDs at the ATP-binding site was more easily achieved in the absence of nucleotides, and most difficult in the presence of ATP (SI Figure S2).

2.4. Collapse and recovery of the substrate binding cavities

Experimental data suggest that during the transition from the IFS to the OFS, cavity 1 completely collapses while the previously occluded cavity 2 opens (Manolaridis et al., 2018, Goebel et al., 2021, Taylor et al., 2017, Jackson et al., 2018). Once the transporter is reset to its IFS, cavity 1 becomes accessible again. The collapse of cavity 1 occurs as a result of the 2 "coupling helices" (CpH, residues 451-462, corresponding to the C-terminal part of TM2, highlighted in red in FIGURE C.1) approaching the 2-fold symmetry axis (Manolaridis et al., 2018).

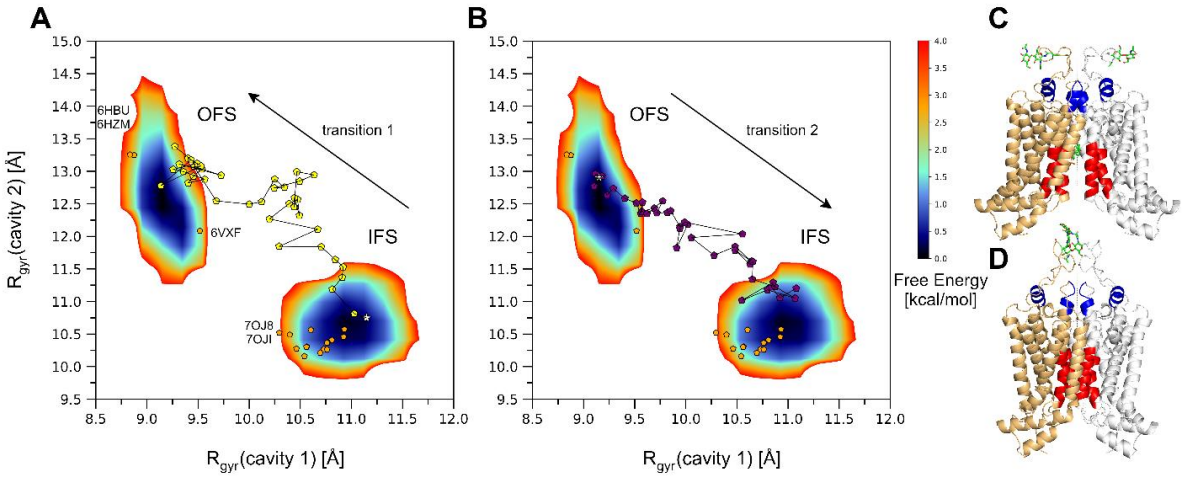

Figure C.5: Changes in the substrate-binding cavities represented by the radius of gyration (R_{gyr}) of the helical structures bordering the cavities. ***(A)*** *The collapse of cavity 1 and the opening of cavity 2 during the ketMD simulation of transition 1 denoted by yellow pentagons and* ***(B)*** *the recovery of cavity 1 and the deflation of cavity 2 during the ketMD simulation of transition 2 denoted by purple pentagons. The initial conformations are indicated as stars. Free Energy Landscapes (FELs) of the classical MD generated conformations starting from the E_1S- and ATP-Mg^{2+}-bound IFS and the nucleotide-free OFS are included as reference in panels A and B, available experimental structures are marked with orange pentagons for reference. The regions determining the R_{gyr} of cavity 1 (highlighted in red, corresponding to the x-axis of panels A and B) and cavity 2 (highlighted in blue, corresponding to the y-axis of panels A and B) are shown* ***(C)*** *in the IFS experimental structure (PDB 6HCO, open cavity 1 and deflated cavity 2) and* ***(D)*** *in the OFS experimental structure (PDB 6HBU, collapsed cavity 1 and widely open cavity 2). The following experimental structures, which fall in the IFS region, are shown but not labelled in panels A and B: PDB 5NJ3, 6ETI, 6FEQ, 6FFC, 6HCO, 6HIJ, 6VXF, 6VXI, 6VXJ, 7NEQ, 7NEZ, 7NFD, 7OJH.*

To assess the changes of the substrate binding cavities during the conformational transitions, the radius of gyration (R_{gyr}) of the helical segments bordering the cavities was calculated. Parts of TM3, TM3', TM5, and TM5' (residues 436-446, 436'-446', 536-547, and 536'-547') for cavity 1, and the upper part of TM3 and TM3' (residues 420-425 and 420'-425') together with the short helical structure within the long loop region connecting TM5 and TM6 (residues 610-617 and 610'-617') for cavity 2 were included for the R_{gyr} calculations (FIGURE C.5). The R_{gyr}

corresponding to cavity 1 is equal to 10.8 Å in the IFS experimental structure (PDB 6HCO), versus 8.9 Å in the OFS structure (PDB 6HBU) where cavity 1 is completely collapsed. When simulating transition 1 with ketMD starting from the IFS, the R_{gyr} corresponding to cavity 1 was reduced to values under 9.3 Å (FIGURE C.5A). For reference, the same R_{gyr} has an average of 9.17 Å with a standard deviation of 0.13 Å among the classical MD generated OFS conformations (the heat-maps in FIGURE C.5A,B correspond to the classical MD simulations). By the end of the ketMD simulation of transition 1, cavity 1 is collapsed and the phenyl rings of residues F439 and F439', initially stacked against the ring system of E_1S (Manolaridis et al., 2018), have moved as close as 3.3 Å from each other, leaving no space for substrates. In the opposite direction starting from the OFS, the R_{gyr} corresponding to cavity 1 increased to as great as 11 Å (FIGURE C.5B), while cavity 1 became exposed and accessible from the cytosol. The average of the R_{gyr} among the reference classical MD generated IFS conformations is 10.9 Å, the standard deviation is 0.2 Å.

The R_{gyr} corresponding to cavity 2 is equal to 10.3 Å in the IFS, while 13.2 Å in the OFS reference experimental structure (PDB 6HCO and 6HBU respectively). Starting from the IFS, during the ketMD simulation of transition 1 as cavity 2 became more exposed to the extracellular space, it also became more voluminous with R_{gyr} values reaching 13.3 Å (FIGURE C.5A); while during the simulation of transition 2 starting from the OFS, cavity 2 approached a more deflated state with R_{gyr} values decreasing to around 11 Å. The variations in cavity 2 volume predominantly originate from the rearrangements of the loop regions connecting TM5 and TM6 and the inflating-deflating motions of the cavity are coupled to the conformational transitions between the IFS and the OFS. An additional binding site was proposed by an *in silico* docking study, delimited by TM1, TM2, TM3, and TM4 and formed primarily by residues Q398, S440, S443, R482, and L539' (Laszlo et al., 2016). Such a site was preserved throughout our ketMD and subsequent classical MD simulations, although E_1S did not approach it during its translocation in our simulations as it is located more peripherally than cavities 1 and 2, and the leucine gate. That site encompasses residue P480 as well as R482, which was suggested to play an important role in substrate transport (Ozvegy et al., 2002) and ATP hydrolysis but not in substrate binding (Ejendal et al., 2006).

2.5. Substrate translocation

In addition to other substrates and inhibitors (SI Table S2), there are currently two E_1S-bound experimental ABCG2 structures (PDB ID 6HCO and 7OJ8). In both cases, the substrate is bound to cavity 1. Experimental structures with a substrate bound to cavity 2, or transient structures along the translocation pathway have not been resolved. With the ketMD run starting from the IFS, it was possible to simulate the translocation of E_1S from cavity 1 to the extracellular space through the leucine gate and cavity 2. In addition to the excitation applied to the transporter, the substrate motion was also kinetically promoted during our ketMD simulation. The velocity components of its atoms pointing towards the extracellular space, perpendicular to the membrane surface, were repeatedly increased, each time followed by a 5 ps relaxation. Subsequently, we performed 10-ns classical MD simulations starting from the ketMD generated

transient conformations along the translocation pathway, to gain insight into the substrate-transporter interactions.

Initially, E_1S was bound to cavity 1, stabilized mainly by the 'sandwich-like' stacking interactions of F439 and F439'. In our ketMD starting conformation, a hydrogen bond formed between N436 and the sulfate group of E_1S further stabilizes the substrate in cavity 1, although this interaction is non-existent in 7OJ8. The substrate remained bound to cavity 1 until the 7th excitation cycle. Key binding residues may be substrate-dependent, except for F439 which is essential for engaging in the transport, as demonstrated by Gose et al. (Gose et al., 2020). The efflux of small molecules investigated in their study was affected by mutation at position F439, but not at N436. However, the latter mutation has been reported to abolish the transport of E_1S, a bulky compound (Manolaridis et al., 2018).

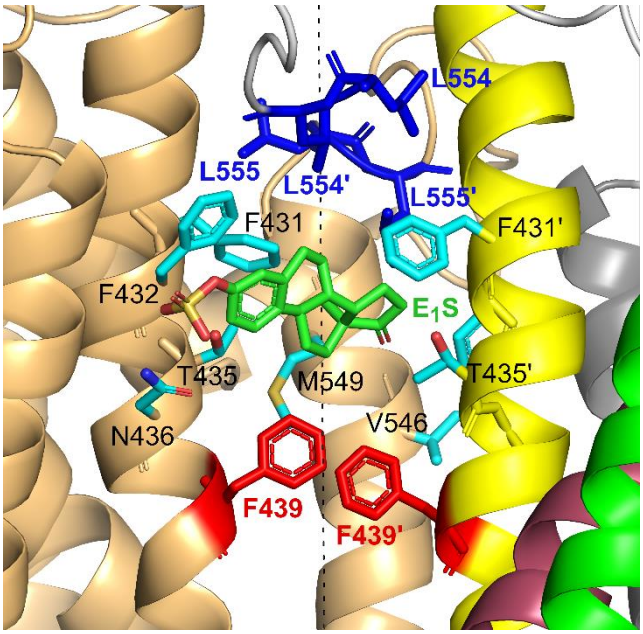

Figure C.6: The pocket-like formation observed during the substrate translocation after leaving cavity 1 but before reaching cavity 2, located between the F439 valve (in red) and the leucine gate (in blue). Residues forming strong interactions with the substrate are labelled and are shown in cyan licorice representation.

During the 8th excitation cycle, E_1S escaped from the 'sandwich-like' trap of the two F439 residues and moved towards cavity 2. As soon as the substrate left, F439 and F439' came into close contact, creating a valve-like construction, similar to what is observed in the OFS cryo-EM structure (PDB 6HBU). Any kind of return movement towards the cytosol is prevented with this valve closed. In our ketMD simulation this was followed by a relatively stable period during which the substrate was trapped between cavities 1 and 2, with movements to cavity 2 still blocked by the closed leucine gate. The stabilizing interactions on the side of this pocket-like formation, located between cavities 1 and 2, involve residues F431, F432, T435, N436, V546, and M549 of the two monomers (FIGURE C.6 and SI Figure S3). Interestingly, Krapf et al. have also proposed F431, F432, and T435 to interact with quinazolines inhibiting ABCG2 (Krapf et al., 2018). Our substrate did not move further until the 18th excitation cycle even though the conformational transition continued and the 'coupling helices' moved closer together. This

demonstrates that passing through the leucine gate that necessitates the separation of the leucine residues of the two monomers requires energy. We argue that the conformational transition from the IFS to the OFS alone cannot induce leucine gate opening and substrate passage, as previously suggested for the E211Q mutant by Manolaridis et al. (Manolaridis et al., 2018). Our findings are consistent with the observations of Nagy et al., who determined a free energy barrier associated with the substrate passing the leucine gate between 7-13 kcal/mol for uric acid, investigated by metadynamics simulations (Nagy et al., 2021).

Once the leucine residues were separated, the substrate was able to slip between them. Here, we identified extensive interactions between E_1S and the leucine gate, especially L554 and L554'. In addition, strong interactions were formed with Q424, Q424', F431', S552', and F578'. The substrate must first escape the grip of the leucine residues and their surroundings to reach cavity 2. In our simulation, we observed this during the 23rd excitation cycle. In the L554A mutant transporter, possibly reduced attractive interactions may explain its (two-fold) higher transport activity than wild-type ABCG2, as reported in the study of Manolaridis et al. (Manolaridis et al., 2018). The sulfate group of E_1S was the last to leave the gate region in our ketMD simulation.

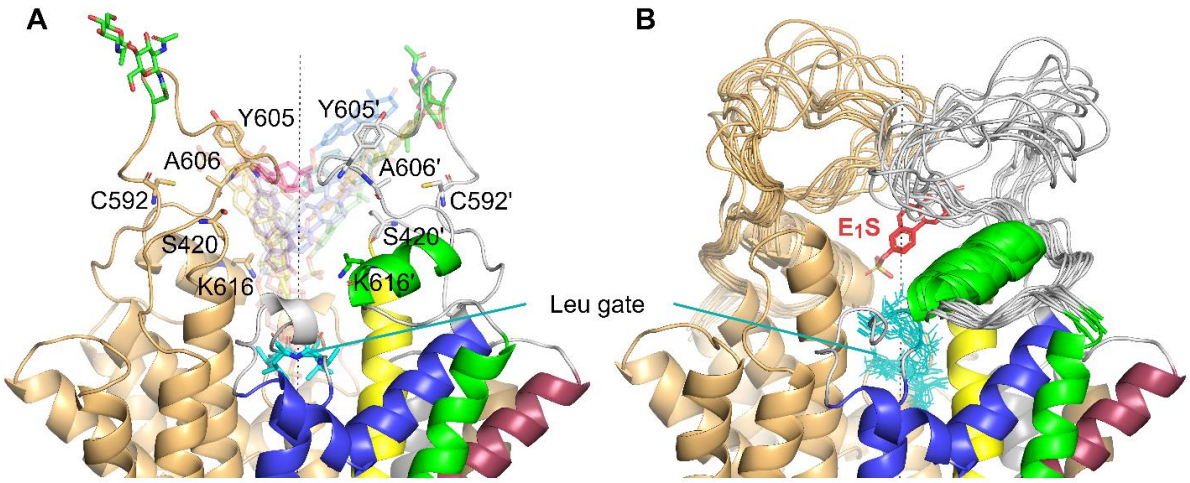

*Figure C.7: Substrate behavior in cavity 2. (**A**) Different substrate positions in cavity 2 observed during the classical MD simulations starting from the ketMD-generated transient conformations, from the crossing of the leucine gate to the leaving of the cavity. Residues forming strong interactions with the substrate are labelled and are shown in licorice representation. (**B**) The fluctuation of the external loop regions corresponding to the substrate positions in panel A.*

The substrate behavior in cavity 2 is very different from what can be observed either in cavity 1 or between cavities 1 and 2. Before arriving at cavity 2, the substrate was tightly bound and closely surrounded by transporter residues. In contrast, the substrate was loosely bound here as it explored the cavity volume, making close contacts with residues at its boundary. These contacts involved S420, C592, Y605, and A606 of both monomers and K616 of one of the monomers (FIGURE C.7 and SI Figure S4). The substrate's further kinetic excitation resulted in its complete detachment from the transporter into the extracellular space.

Recently multiple IFS structures have been resolved with bound nucleotides, thus we also performed ketMD simulation of transition 1 starting from the structure PDB 7OJ8 (Yu et al., 2021) (ATP-bound ABCG2 in the presence of E_1S in cavity 1). The comparison of the P-loop regions (SI Figure S5) showed almost no difference between the ATP binding between the PDB 7OJ8 structure and the model we constructed using the IFS structure (PDB 6HCO) with the nucleotide taken from the OFS structure (PDB 6HBU). The NBDs feature a semi-closed dimer in the starting structure (PDB 7OJ8). After some opening during the equilibration, during the ketMD simulation a tightly packed NBD dimer was reached, the catalytic ATP-binding sites were formed (the backbone RMSD of the NBD dimer with respect to the target OFS structure (PDB 6HBU) was decreased from 3.2 Å to 1.7 Å). E_1S left the grip of the residues F439 and F439' sooner (after the 3rd excitation cycle) and also its crossing through the leucine gate occurred earlier in the ketMD simulations (after the 12th excitation cycle), compared to the ketMD simulation starting from the more open IFS structure (PDB 6HCO), while cavity 1 collapsed and cavity 2 became more exposed to the extracellular space (the backbone RMSD of the TMD dimer with respect to the target OFS structure was reduced from 3.4 Å to 1.7 Å). The ketMD simulation starting from PDB 7OJ8 further supports the existence of a stable pocket-like formation between cavities 1 and 2 (SI Figure S6) where E_1S was trapped for 9 consecutive excitation cycles.

2.6. Effect of substrate and nucleotide binding

We analyzed the different stages of the transport cycle by performing classical MD simulations and NMA. We built IFS systems in their apo-form, with bound E_1S, and bound E_1S and ATP-Mg^{2+} together; the OFS systems were constructed with bound ATP-Mg^{2+}, ADP, and without nucleotides (for their construction see Materials and Methods and SI Table S1). For all of these systems after equilibration, we first performed a 100-ns classical MD simulation. FIGURE C.8 shows the MD frames in the subspace of NBDs and TMDs difference vectors. Conformations were first overlapped to the mean-conformation of the IFS and the OFS experimental structures (PDB 6HCO and 6HBU respectively) and were then projected to the NBD and TMD difference vectors, pointing from the OFS to the IFS structure. After overlapping the OFS and the IFS structures, the difference vector points for each Cα atom from its 3D coordinates in the OFS to its position in the IFS structure. The so obtained difference vector of 3N elements (N is the number of Cα atoms of the system, each having xyz coordinates) was then used to project conformational differences from the mean experimental structure.

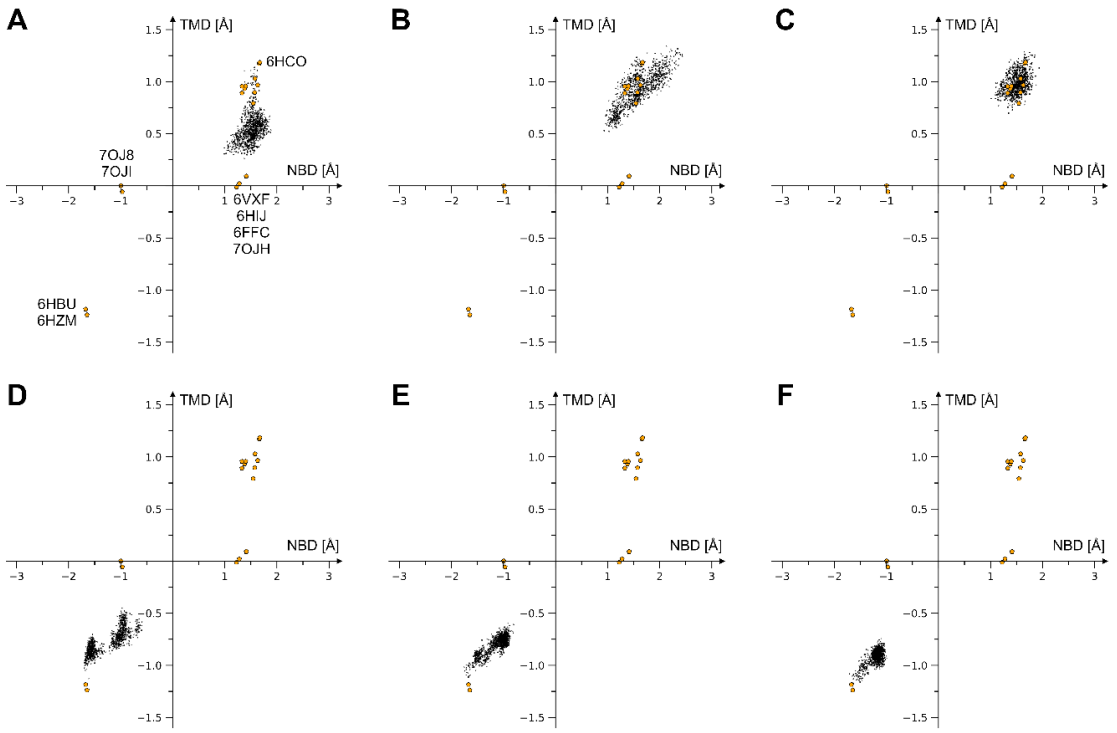

*Figure C.8: The classical MD generated conformations projected in the subspace of the NBDs' and TMDs' difference vectors. The difference vector, after overlapping the IFS (PDB 6HCO) and the OFS (PDB 6HBU) experimental structures, points for each Cα atom from its 3D coordinates in the OFS to its position in the IFS structure. The NBDs and TMDs difference vectors were used for the projection of the conformational differences from the mean experimental structure (of PDB 6HCO and 6HBU) in the case of (**A**) the apo IFS, (**B**) the substrate-bound IFS, and (**C**) the substrate- and ATP-Mg^{2+}-bound IFS transporter, (**D**) the ATP-Mg^{2+}-bound OFS, (**E**) the ADP-bound OFS, and (**F**) the nucleotide-free OFS ABCG2 BCRP. Available experimental structures are marked with orange pentagons. The following experimental structures, which fall in the IFS region, are shown but not labelled: PDB 5NJ3, 6ETI, 6FEQ, 6VXI, 6VXJ, 7NEQ, 7NEZ, 7NFD.*

We found that the presence of the substrate in cavity 1 is essential to couple the movements between the NBDs and the TMDs. In the absence of a bound substrate and nucleotides (apo-form), the TMDs approach a neutral configuration while the NBDs stay far apart (FIGURE C.8A). Further analyses revealed that cavity 2 opens while cavity 1 starts collapsing in the absence of a substrate in cavity 1, approaching the state of the apo-closed experimental structure (PDB 6VXF, nucleotide-free apo state), where the arrangement of the TM helices more closely resembles that seen in the outward facing ATP bound state, whereas the lack of NBD dimerization more closely resembles that of the inward facing state (Orlando and Liao, 2020) (SI Figure S7A). In contrast, the substrate-bound IFS showed coupled motions between the NBDs and TMDs (FIGURE C.8B). In the presence of the substrate in cavity 1, we did not observe larger changes in the state of cavities 1 and 2 (SI Figure S7B,C), the addition of the nucleotides did not induce the onset of a clear transition to the OFS on the simulated time scale (FIGURE C.8C). We hypothesize that the binding order of substrate and nucleotides related to their physiological concentrations could play a role during the transport cycle, although the time-

scale and the character of our simulations do not allow us to draw further conclusions in this regard.

In all the MD simulations starting from the OFS, no conformations moved further away in the direction opposite to the IFS (FIGURE C.8D,E,F). Moreover, the conformations rapidly drifted in the direction of the IFS, independent of the bound nucleotide. We interpret this as a consequence of removing the E211Q mutation which is present in the only OFS cryo-EM structures available (6HBU and 6HZM). This mutation can generate a more tightly packed NBD dimer-interface than what might exist in the case of the wild-type transporter (see position of E211 in FIGURE C.3). As a result, it may also generate or stabilize a TMD configuration that is more extremely open to the extracellular space. In all our OFS MD simulations, d(88,190') exhibited a stable state at around 17 Å (this distance is 13.9 Å in the E211Q mutant OFS cryo-EM structure, 6HBU, SI Figure S8D,E,F). The R_{gyr} corresponding to cavity 2 slightly decreased (closing of cavity 2) while the R_{gyr} corresponding to cavity 1 slightly increased (opening of cavity 1) or remained unchanged during the nucleotide-bound OFS MD simulations (SI Figure S7A,B, for cavity 2 the average of ATP-bound OFS is 12.28 Å, ADP-bound OFS is 12.2 Å versus the experimental OFS of 13.2 Å , and for cavity 1 the average of ATP-bound OFS is 9.0 Å, ADP-bound OFS is 9.18 Å versus the experimental OFS of 8.9 Å). This suggests that the most stable states during the nucleotide-bound OFS classical MD simulations were somewhat less extreme than the experimental OFS structure (PDB 6HBU).

Interestingly, in the case of the ATP-Mg^{2+}-bound transporter, a steady state was present where one of the ATP-binding site distance d(88,190') was around 20 Å (SI Figure S8D). Similar distances exist in the ATP-bound IFS structures in the presence of E_1S and topotecan (an exogenous substrate), which are 20.3 Å and 20.5 Å, respectively (PDB ID 7OJI and 7OJ8) (Yu et al., 2021). It is unclear whether physiologically the ATP-bound OFS is a state with high probability (as the ATPs might be hydrolyzed upon the translocation of the substrate). Our results and the available experimental structures suggest that the sole presence of the ATPs may determine the openness of the NBD dimer.

Furthermore, we argue that contrary to cavity 1, cavity 2 is never fully collapsed, at any stages of the transport cycle even though experiments suggest that it can get occluded (Manolaridis et al., 2018, Goebel et al., 2021, Taylor et al., 2017, Jackson et al., 2018). Its volume shows inflating-deflating variations between the IFS and OFS states, predominantly due to the rearrangements of the loop regions connecting TM5 and TM6. However, cavity 2 is more voluminous than cavity 1 even in its deflated state. The restraining region in cavity 2 is the passage between the upper tips of TM3 and TM3' (residues 420-425). However, this opening shows a high overlap of a large fluctuation between the IFS and the OFS (SI Figure S9, the average is 7.3 Å and 6.2 Å, the standard deviation is 1.41 Å and 1.39 Å for the IFS and OFS free MD simulations respectively). Moreover, our ketMD and subsequent free MD simulations showed that in the presence of E_1S at this region, the minimum distance between TM3 and TM3' can decrease to below 7 Å. This demonstrates that E_1S, which is a bulky compound could pass through this passage and be present in cavity 2, even in the OFS. It also shows a sufficiently

large space for the substrate in cavity 2 throughout the entire transport cycle. This may allow simultaneous substrate binding in cavities 1 and 2, resulting in an accelerated export mechanism.

Prior to our ketMD simulations, we also performed Normal Mode Analysis (NMA) of all the IFS and OFS systems and calculated the fluctuations of the previously discussed distances at 300 K, according to Equation C.2 (see Materials and Methods), which we derived from the classical formula of harmonic approximation to the amplitudes of atomic vibrations (Levy and Karplus, 1979, Levy et al., 1982). We found that the fluctuations of d(88,190') and d(190,88') are both significantly damped in the OFS systems compared to the IFS (SI Figure S10). Visual inspection of the corresponding NMs also revealed that such fluctuations correspond to global transition-like motions in the IFS systems whereas they are local motions of higher frequencies in the OFS systems. Based on the harmonic approximation of NMs, we conclude that it is energetically costly to start transition 2 to return to the initial IFS, which may require the energy released upon ATP hydrolysis, supporting the suggestion for the mechanism by Manolaridis et al. (Manolaridis et al., 2018).

3. Materials and Methods

3.1. Transporter structure preparation

All simulations were performed using the human homodimer ABCG2. Cryo-EM structures from the Protein Data Bank were taken as starting coordinates, entry 6HCO (Manolaridis et al., 2018) (IFS, E_1S-bound) and 7OJ8 (IFS, ATP- and E_1S-bound) for the IFSs, and 6HBU (Manolaridis et al., 2018) (OFS, ATP-Mg^{2+}-bound) for the OFSs. The different structural elements of ABCG2 are presented in SI Table S3. The structures PDB 6HCO and 6HBU were solved with the E211Q mutation, which in this study was reverted to the wild type using CHARMM-GUI (Jo et al., 2008). The human-specific 5D3 antibody (Fab) molecules were removed from the structure. The missing loop regions in the NBDs (residues 47-60, 302-327, 355-371) were modelled using the DaReUS-Loop web server (Karami et al., 2019), and the missing C-terminal S655 was built from the internal coordinate table of CHARMM (Brooks et al., 2009). The first 34 missing N-terminal residues were neglected in all our simulations. Disulfide bridges were set between C592 and C608 in each subunit, and between C603 residues linking the two subunits. The PPM web server (Lomize et al., 2012) was used to determine the orientation of the transporter within the membrane. The pKa values of the protein titratable groups were calculated with PROPKA (Sondergaard et al., 2011), and protonation states were assigned at pH 7.0 outside, and pH 4.0 inside the membrane. The parameters of the substrate E_1S molecule were determined using the CHARMM General Force Field (CGenFF) 2.5 (Vanommeslaeghe et al., 2010).

Multiple systems were constructed starting from the experimental IFS and OFS structures. Using PDB 6HCO an apo IFS (by removing E_1S from cavity 1), a single substrate-bound IFS, and a substrate- and ATP-Mg^{2+}-bound IFS transporter (by taking the ATP-Mg^{2+} positions after overlapping the backbone residues 80-94 of 6HCO on 6HBU (RMSD of 0.6 Å), SI Figure S11). An additional ATP-Mg^{2+}-bound IFS transporter was constructed using PDB 7OJ8. Furthermore, using the structure PDB 6HBU an ATP-Mg^{2+}-bound OFS, an ADP-bound OFS (by

cleaving away the γ-phosphates of the ATPs *in silico* and removing the Mg^{2+} ions), and an OFS transporter with no nucleotides bound (by removing both the ATPs and the Mg^{2+} ions, see SI Table S1 for details).

The PPM-oriented structures were inserted into a lipid bilayer composed of dimiristoylphosphatidylcholine (DMPC) with 20% cholesterol (CHOL) following the work of Ferreira et al. (Ferreira et al., 2017) and the TIP3-solvated systems were generated by CHARMM-GUI. The NaCl concentration was set to 0.15 M.

Each system was energy minimized by alternating 250 steps Steepest Descent (SD) and 250 steps of Adopted Basis Newton-Raphson (ABNR) minimization, 10 times each. This was followed by 10000 Conjugate Gradient (CONJ) steps. The minimization steps were performed with CHARMM (Brooks et al., 2009) using the all-atom additive CHARMM C36m (Huang et al., 2017) force field (FF), with harmonic constraints applied to the backbone (10 kcal/mol/$Å^2$) and the side chain (5 kcal/mol/$Å^2$) heavy atoms.

The systems were then equilibrated at 300 K with progressively decreasing harmonic restraining force constants (every 100 ps) by adopting the values 10, 5, 2.5, 1, 0.5, 0.1 kcal/mol/$Å^2$ for the backbone heavy atoms and 5, 2.5, 1.25, 0.5, 0.25, 0.05 kcal/mol/$Å^2$ for the side chain heavy atoms in an NVT ensemble. The pressure was set to 1 atm and the integration time step to 1 fs. Finally, a 5 ns NPT equilibration run was performed at 300 K, 1 atm, with an integration time step of 2 fs. Equilibration runs were performed with NAMD (Phillips et al., 2020) using the C36m FF. Langevin dynamics was used for constant temperature control with a damping coefficient of 1 ps^{-1}. Constant pressure was achieved using the Nose-Hoover method with a piston oscillation period of 50 fs and a piston oscillation decay time of 25 fs. For energy calculations, the dielectric constant was set to 1. The Particle Mesh Ewald (PME) method was used to calculate electrostatic interactions with a grid spacing of 1 Å or less having the order of 6. The real-space summation was truncated at 12.0 Å, and the width of Gaussian distribution was set to 0.34 $Å^{-1}$. Van der Waals interactions were reduced to zero by 'switch' truncation operating between 10.0 and 12.0 Å.

3.2. Molecular Dynamics simulations

Three, 100-ns-long classical Molecular Dynamics (MD) simulations with different initial velocity distributions were carried out on the systems that were also used in our ketMD simulations (the substrate- and ATP-Mg^{2+}-bound IFS based on PDB 6HCO and the OFS with no bound nucleotide based on 6HBU), using the same initial conformations as ketMD in order to compare the conformational space exploration of MD to the ketMD simulations. A single 100-ns-long MD simulation was carried out for the other four IFS and OFS systems listed previously. NAMD was used for all of these runs with the C36m FF. The integration time step was 2 fs, and the coordinates were saved every 10 ps. The same parameters were used as for the 5 ns NPT equilibration runs described above. Further 10 ns classical MD simulations were carried out starting from the transient conformations along transitions 1 and 2, generated by the ketMD simulations to identify transporter-substrate interactions along the translocation pathway. The systems with the transient conformations were first de-excited by releasing the excess kinetic

energy introduced along the excitation target direction for 10 ps before starting the classical MD simulations. This was achieved by applying a harmonic restraint potential along the target direction, allowing the fast dissipation of the excitation energy. The applied harmonic force constant was 1000 kcal/mol/Å^2.

3.3. Normal Mode Analysis

Normal Mode Analysis (NMA) of each system was performed using the same C36m FF, starting from the equilibrated conformations. The lipid and solvent molecules were first removed. The potential energy of the transporter with the bound ligands was then energy minimized using the SD method with decreasing harmonic restraining potentials applied to the heavy atoms. The harmonic restraining force constants were decreased every 500 steps adopting the values 10, 1, 0.1, and 0 kcal/mol/Å^2. ABNR minimization followed until an RMS energy gradient of 10^{-6} kcal/mol/Å was reached. The normal modes of the energy minimized structures were calculated using the iterative Mixed-Basis Diagonalization (DIMB) routine (Mouawad and Perahia, 1993, Perahia and Mouawad, 1995) of the VIBRAN module in CHARMM.

3.4. Kinetically excited targeted MD

We implemented a method, kinetically excited targeted MD (ketMD), to simulate the conformational transitions between the IFS and the OFS. Our concept relies on the MDeNM method (Costa et al., 2015), designed to enhance the conformational exploration of proteins. Similar to MDeNM, ketMD is based on kinetic excitations. In each excitation cycle, the velocity components pointing from the instantaneous conformation to the target are increased at the first step of the MD simulation. Then, the injected kinetic energy dissipates during a relaxation period while no external perturbation is introduced, and the system progresses. The kinetic excitation corresponded to an overall temperature rise of 2 K in the systems (as was suggested by Kaynak et al. (Kaynak et al., 2022) for MDeNM). As the excitation kinetic energy dissipates rapidly (in less than 1 ps (Costa et al., 2015, Floquet et al., 2015)), 40 consecutive excitation cycles were performed, each containing a 5 ps relaxation MD simulation. Thus, the total ketMD simulation time was 40 x 5 ps = 200 ps per system. We performed ketMD simulations with excitation applied also to the substrate, starting from the substrate-bound IFS. The velocity components of the substrate, perpendicular to the membrane surface pointing to the extracellular space, were also increased at the first step of the MD simulations in each excitation cycle, corresponding to an additional 0.5 K temperature rise of the given system.

3.5. Free Energy Landscape (FEL) calculations

FELs of the MD-generated conformations were calculated within the subspace of d(88,190') vs. d(190,88') and the R_{gyr} corresponding to cavities 1 and 2. The most populated state was used as a reference for calculating free energy differences. The free energy difference (ΔG_α) of a given state α was determined by considering the probability of the occurrence of the states $P(q_\alpha)$ and $P_{max}(q)$ given by the equation:

$$\Delta G_\alpha = -k_B T ln\left[\frac{P(q_\alpha)}{P_{max}(q)}\right] \qquad (C.1)$$

where k_B is the Boltzmann constant, T is the temperature of the simulation, $P(q_\alpha)$ is an estimate of the probability density function obtained from the bi-dimensional histogram of the conformational distribution, and $P_{max}(q)$ is the probability of the most populated state. The free energy differences should be considered here as entropic quantities reflecting the populations in terms of energy.

3.6. Distance RMSF in NMs

The harmonic approximation to the amplitudes of inter-atomic distance vibration contributions by the different NMs at a given temperature was calculated by evaluating the equation:

$$\langle \Delta \boldsymbol{d}^2_{p,q,i} \rangle = \frac{k_B T}{\omega_i^2} \left\| \frac{\boldsymbol{e}_{q,i}}{\sqrt{m_q}} - \frac{\boldsymbol{e}_{p,i}}{\sqrt{m_p}} \right\|^2 \qquad (C.2)$$

where $d_{(p,q,i)}$ is the instantaneous distance vector between atom p and q in the i^{th} NM, k_B the Boltzmann constant, T the absolute temperature of the system, ω_i the frequency of the i^{th} NM, $e_{(p,i)}$ and $e_{(q,i)}$ the mass-weighted displacement vectors of atom p and q in in the i^{th} NM, and m_p and m_q the mass of atoms p and q, respectively.

3.7. Interaction Energies

The interaction energy (E_{int}) between two groups of atoms was calculated as a sum of pairwise non-bonded electrostatic and van der Waals energies. For the energy calculations CHARMM was used with a distance dielectric constant of 2. The interactions were calculated by considering the atoms of the substrate and a given transporter residue. The energy values reported are statistical averages of the given E_{int} calculated among the conformations retrieved from the free MD simulations.

4. Conclusions

In this study, we developed and employed an innovative enhanced MD simulation approach, termed ketMD (kinetically excited targeted Molecular Dynamics), which uses kinetic excitation to promote protein movements corresponding to large conformational changes towards a specified target structure, without biasing the potential energy function. With the help of ketMD, we successfully simulated the conformational transitions of the ABCG2 transport cycle, and revealed the complex molecular mechanism of the physiological E_1S substrate translocation. We observed a valve-like function of residues that initially engage in stacking interactions against the substrate in cavity 1 (F439 and F439'). We found that they prevent backwards movements of the substrate towards the cytosol once it escapes their grasp and moves towards the leucine gate. We also identified a pocket-like construction between this valve and the leucine gate, where the substrate is stabilized before it moves to cavity 2.

Furthermore, using MD simulations and NMA of the different transport stages, we have shown that the presence of the substrate in cavity 1 is essential to couple the movements between the NBDs and the TMDs. Additionally, we observed that cavity 2 was never completely collapsed, at any stages of the transport cycle. Therefore, we hypothesize that simultaneous substrate binding in cavity 1 and 2 may occur, which could result in an accelerated export mechanism.

Finally, the harmonic approximation of the ABCG2 dynamics by NMA revealed that low frequency global transition-like motions exist in the IFS but were absent in our calculations for the OFS transporter, where partial transition-like movements are present but are more localized and of higher frequencies. Accordingly, our results further support previous assumptions that transition 2, starting from the OFS transporter, is energetically costly and ABCG2 requires the energy released upon ATP hydrolysis to return to its initial IFS.

Our observations shed new light on the complex molecular mechanism of the ABCG2 transport, and the results highlighted the utility of including enhanced *in silico* sampling techniques, such as ketMD, in transporter studies. In the future, the provided transition pathways can help to identify novel ABCG2 substrates and inhibitors, and probe new drug candidates for MDR and DDI.

Acknowledgements

The authors thank Université Paris Cité, Inserm Instute, and Ecole Normale Supérieure Paris-Saclay for supporting this research. They also thank Dr J. M. Krieger (Centro Nacional de Biotecnología, Consejo Superior de Investigaciones Científicas, Spain) and Dr E. Balog (Department of Biophysics and Radiation Biology, Semmelweis University, Hungary) for the helpful discussions.

Funding

B.D. has received funding from Université Paris Cité (the Idex project).

Competing Interests

The authors declare no competing interests.

5. References

BI, Y., MANN, E., WHITFIELD, C. & ZIMMER, J. 2018. Architecture of a channel-forming O-antigen polysaccharide ABC transporter. *Nature,* 553, 361-365. doi:10.1038/nature25190

BROOKS, B. R., BROOKS, C. L., 3RD, MACKERELL, A. D., JR., NILSSON, L., PETRELLA, R. J., ROUX, B., WON, Y., ARCHONTIS, G., BARTELS, C., BORESCH, S., CAFLISCH, A., CAVES, L., CUI, Q., DINNER, A. R., FEIG, M., FISCHER, S., GAO, J., HODOSCEK, M., IM, W., KUCZERA, K., LAZARIDIS, T., MA, J., OVCHINNIKOV, V., PACI, E., PASTOR, R. W., POST, C. B., PU, J. Z., SCHAEFER, M., TIDOR, B., VENABLE, R. M., WOODCOCK, H. L., WU, X., YANG, W., YORK, D. M. & KARPLUS, M. 2009. CHARMM: the biomolecular simulation program. *J Comput Chem,* 30, 1545-614. doi:10.1002/jcc.21287

BROZIK, A., HEGEDUS, C., ERDEI, Z., HEGEDUS, T., OZVEGY-LACZKA, C., SZAKACS, G. & SARKADI, B. 2011. Tyrosine kinase inhibitors as modulators of ATP binding cassette multidrug transporters: substrates, chemosensitizers or inducers of acquired multidrug resistance? *Expert Opin Drug Metab Toxicol,* 7, 623-42. doi:10.1517/17425255.2011.562892

CAFFALETTE, C. A., COREY, R. A., SANSOM, M. S. P., STANSFELD, P. J. & ZIMMER, J. 2019. A lipid gating mechanism for the channel-forming O antigen ABC transporter. *Nat Commun,* 10, 824. doi:10.1038/s41467-019-08646-8

CHAPY, H., SAUBAMEA, B., TOURNIER, N., BOURASSET, F., BEHAR-COHEN, F., DECLEVES, X., SCHERRMANN, J. M. & CISTERNINO, S. 2016. Blood-brain and retinal barriers show dissimilar ABC transporter impacts and concealed effect of P-glycoprotein on a novel verapamil influx carrier. *Br J Pharmacol,* 173, 497-510. doi:10.1111/bph.13376

CHEN, L., HOU, W. T., FAN, T., LIU, B., PAN, T., LI, Y. H., JIANG, Y. L., WEN, W., CHEN, Z. P., SUN, L., ZHOU, C. Z. & CHEN, Y. 2020. Cryo-electron Microscopy Structure and Transport Mechanism of a Wall Teichoic Acid ABC Transporter. *mBio,* 11 doi:10.1128/mBio.02749-19

CISTERNINO, S., MERCIER, C., BOURASSET, F., ROUX, F. & SCHERRMANN, J. M. 2004. Expression, up-regulation, and transport activity of the multidrug-resistance protein Abcg2 at the mouse blood-brain barrier. *Cancer Res,* 64, 3296-301. doi:10.1158/0008-5472.can-03-2033

COSTA, M. G., BATISTA, P. R., BISCH, P. M. & PERAHIA, D. 2015. Exploring free energy landscapes of large conformational changes: molecular dynamics with excited normal modes. *J Chem Theory Comput,* 11, 2755-67. doi:10.1021/acs.jctc.5b00003

DANILOWICZ, C., HERMANS, L., COLJEE, V., PREVOST, C. & PRENTISS, M. 2017. ATP hydrolysis provides functions that promote rejection of pairings between different copies of long repeated sequences. *Nucleic Acids Res,* 45, 8448-8462. doi:10.1093/nar/gkx582

DIESTRA, J. E., SCHEFFER, G. L., CATALA, I., MALIEPAARD, M., SCHELLENS, J. H., SCHEPER, R. J., GERMA-LLUCH, J. R. & IZQUIERDO, M. A. 2002. Frequent expression of the multi-drug resistance-associated protein BCRP/MXR/ABCP/ABCG2 in human tumours detected by the BXP-21 monoclonal antibody in paraffin-embedded material. *J Pathol,* 198, 213-9. doi:10.1002/path.1203

DUDAS, B., MERZEL, F., JANG, H., NUSSINOV, R., PERAHIA, D. & BALOG, E. 2020. Nucleotide-Specific Autoinhibition of Full-Length K-Ras4B Identified by Extensive Conformational Sampling. *Front Mol Biosci,* 7, 145. doi:10.3389/fmolb.2020.00145

DUDAS, B., PERAHIA, D. & BALOG, E. 2021a. Revealing the activation mechanism of autoinhibited RalF by integrated simulation and experimental approaches. *Sci Rep,* 11, 10059. doi:10.1038/s41598-021-89169-5

DUDAS, B., TOTH, D., PERAHIA, D., NICOT, A. B., BALOG, E. & MITEVA, M. A. 2021b. Insights into the substrate binding mechanism of SULT1A1 through molecular dynamics with excited normal modes simulations. *Sci Rep,* 11, 13129. doi:10.1038/s41598-021-92480-w

EJENDAL, K. F., DIOP, N. K., SCHWEIGER, L. C. & HRYCYNA, C. A. 2006. The nature of amino acid 482 of human ABCG2 affects substrate transport and ATP hydrolysis but not substrate binding. *Protein Sci,* 15, 1597-607. doi:10.1110/ps.051998406

FERREIRA, R. J., BONITO, C. A., CORDEIRO, M., FERREIRA, M. U. & DOS SANTOS, D. 2017. Structure-function relationships in ABCG2: insights from molecular dynamics simulations and molecular docking studies. *Sci Rep,* 7, 15534. doi:10.1038/s41598-017-15452-z

FETSCH, P. A., ABATI, A., LITMAN, T., MORISAKI, K., HONJO, Y., MITTAL, K. & BATES, S. E. 2006. Localization of the ABCG2 mitoxantrone resistance-associated protein in normal tissues. *Cancer Lett,* 235, 84-92. doi:10.1016/j.canlet.2005.04.024

FLOQUET, N., COSTA, M. G., BATISTA, P. R., RENAULT, P., BISCH, P. M., RAUSSIN, F., MARTINEZ, J., MORRIS, M. C. & PERAHIA, D. 2015. Conformational Equilibrium of CDK/Cyclin Complexes by Molecular Dynamics with Excited Normal Modes. *Biophys J,* 109, 1179-89. doi:10.1016/j.bpj.2015.07.003

GHODE, A., GROSS, L. Z. F., TEE, W. V., GUARNERA, E., BEREZOVSKY, I. N., BIONDI, R. M. & ANAND, G. S. 2020. Synergistic Allostery in Multiligand-Protein Interactions. *Biophys J,* 119, 1833-1848. doi:10.1016/j.bpj.2020.09.019

GILLET, J. P. & GOTTESMAN, M. M. 2011. Advances in the molecular detection of ABC transporters involved in multidrug resistance in cancer. *Curr Pharm Biotechnol,* 12, 686-92. doi:10.2174/138920111795163931

GOEBEL, J., CHMIELEWSKI, J. & HRYCYNA, C. A. 2021. The roles of the human ATP-binding cassette transporters P-glycoprotein and ABCG2 in multidrug resistance in cancer and at endogenous sites: future opportunities for structure-based drug design of inhibitors. *Cancer Drug Resist,* 4, 784-804. doi:10.20517/cdr.2021.19

GOLDWASER, E., LAURENT, C., LAGARDE, N., FABREGA, S., NAY, L., VILLOUTREIX, B. O., JELSCH, C., NICOT, A. B., LORIOT, M. A. & MITEVA, M. A. 2022. Machine learning-driven identification of drugs inhibiting cytochrome P450 2C9. *PLoS Comput Biol,* 18, e1009820. doi:10.1371/journal.pcbi.1009820

GOSE, T., SHAFI, T., FUKUDA, Y., DAS, S., WANG, Y., ALLCOCK, A., GAVAN MCHARG, A., LYNCH, J., CHEN, T., TAMAI, I., SHELAT, A., FORD, R. C. & SCHUETZ, J. D. 2020. ABCG2 requires a single aromatic amino acid to "clamp" substrates and inhibitors into the binding pocket. *FASEB J,* 34, 4890-4903. doi:10.1096/fj.201902338RR

GOTTESMAN, M. M., FOJO, T. & BATES, S. E. 2002. Multidrug resistance in cancer: role of ATP-dependent transporters. *Nat Rev Cancer,* 2, 48-58. doi:10.1038/nrc706

HILLGREN, K. M., KEPPLER, D., ZUR, A. A., GIACOMINI, K. M., STIEGER, B., CASS, C. E., ZHANG, L. & INTERNATIONAL TRANSPORTER, C. 2013. Emerging transporters of clinical importance: an update from the International Transporter Consortium. *Clin Pharmacol Ther,* 94, 52-63. doi:10.1038/clpt.2013.74

HUANG, J., RAUSCHER, S., NAWROCKI, G., RAN, T., FEIG, M., DE GROOT, B. L., GRUBMULLER, H. & MACKERELL, A. D., JR. 2017. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat Methods,* 14, 71-73. doi:10.1038/nmeth.4067

IMAI, Y., ASADA, S., TSUKAHARA, S., ISHIKAWA, E., TSURUO, T. & SUGIMOTO, Y. 2003. Breast cancer resistance protein exports sulfated estrogens but not free estrogens. *Mol Pharmacol,* 64, 610-8. doi:10.1124/mol.64.3.610

JACKSON, S. M., MANOLARIDIS, I., KOWAL, J., ZECHNER, M., TAYLOR, N. M. I., BAUSE, M., BAUER, S., BARTHOLOMAEUS, R., BERNHARDT, G., KOENIG, B., BUSCHAUER, A., STAHLBERG, H., ALTMANN, K. H. & LOCHER, K. P. 2018. Structural basis of small-molecule inhibition of human multidrug transporter ABCG2. *Nat Struct Mol Biol,* 25, 333-340. doi:10.1038/s41594-018-0049-1

JO, S., KIM, T., IYER, V. G. & IM, W. 2008. CHARMM-GUI: a web-based graphical user interface for CHARMM. *J Comput Chem,* 29, 1859-65. doi:10.1002/jcc.20945

JONES, P. M. & GEORGE, A. M. 2004. The ABC transporter structure and mechanism: perspectives on recent research. *Cell Mol Life Sci,* 61, 682-99. doi:10.1007/s00018-003-3336-9

KARAMI, Y., REY, J., POSTIC, G., MURAIL, S., TUFFERY, P. & DE VRIES, S. J. 2019. DaReUS-Loop: a web server to model multiple loops in homology models. *Nucleic Acids Res,* 47, W423-W428. doi:10.1093/nar/gkz403

KAYNAK, B. T., KRIEGER, J. M., DUDAS, B., DAHMANI, Z. L., COSTA, M. G. S., BALOG, E., SCOTT, A. L., DORUKER, P., PERAHIA, D. & BAHAR, I. 2022. Sampling of Protein Conformational Space Using Hybrid Simulations: A Critical Assessment of Recent Methods. *Front Mol Biosci,* 9, 832847. doi:10.3389/fmolb.2022.832847

KHUNWEERAPHONG, N. & KUCHLER, K. 2021. Multidrug Resistance in Mammals and Fungi-From MDR to PDR: A Rocky Road from Atomic Structures to Transport Mechanisms. *Int J Mol Sci,* 22 doi:10.3390/ijms22094806

KOWAL, J., NI, D., JACKSON, S. M., MANOLARIDIS, I., STAHLBERG, H. & LOCHER, K. P. 2021. Structural Basis of Drug Recognition by the Multidrug Transporter ABCG2. *J Mol Biol,* 433, 166980. doi:10.1016/j.jmb.2021.166980

KRAPF, M. K., GALLUS, J., NAMASIVAYAM, V. & WIESE, M. 2018. 2,4,6-Substituted Quinazolines with Extraordinary Inhibitory Potency toward ABCG2. *J Med Chem,* 61, 7952-7976. doi:10.1021/acs.jmedchem.8b01011

LASZLO, L., SARKADI, B. & HEGEDUS, T. 2016. Jump into a New Fold-A Homology Based Model for the ABCG2/BCRP Multidrug Transporter. *PLoS One,* 11, e0164426. doi:10.1371/journal.pone.0164426

LEVY, R. M. & KARPLUS, M. 1979. Vibrational Approach to the Dynamics of an α-Helix. *Biopolymers,* 18, 2465-2495. doi:10.1002/bip.1979.360181008

LEVY, R. M., PERAHIA, D. & KARPLUS, M. 1982. Molecular dynamics of an alpha-helical polypeptide: Temperature dependence and deviation from harmonic behavior. *Proc Natl Acad Sci U S A,* 79, 1346-50. doi:10.1073/pnas.79.4.1346

LOMIZE, M. A., POGOZHEVA, I. D., JOO, H., MOSBERG, H. I. & LOMIZE, A. L. 2012. OPM database and PPM web server: resources for positioning of proteins in membranes. *Nucleic Acids Res,* 40, D370-6. doi:10.1093/nar/gkr703

MACALOU, S., ROBEY, R. W., JABOR GOZZI, G., SHUKLA, S., GROSJEAN, I., HEGEDUS, T., AMBUDKAR, S. V., BATES, S. E. & DI PIETRO, A. 2016. The linker region of breast cancer resistance protein ABCG2 is critical for coupling of ATP-dependent drug transport. *Cell Mol Life Sci,* 73, 1927-37. doi:10.1007/s00018-015-2118-5

MALIEPAARD, M., SCHEFFER, G. L., FANEYTE, I. F., VAN GASTELEN, M. A., PIJNENBORG, A. C., SCHINKEL, A. H., VAN DE VIJVER, M. J., SCHEPER, R. J. & SCHELLENS, J. H. 2001. Subcellular localization and distribution of the breast cancer resistance protein transporter in normal human tissues. *Cancer Res,* 61, 3458-64.

MANOLARIDIS, I., JACKSON, S. M., TAYLOR, N. M. I., KOWAL, J., STAHLBERG, H. & LOCHER, K. P. 2018. Cryo-EM structures of a human ABCG2 mutant trapped in ATP-bound and substrate-bound states. *Nature,* 563, 426-430. doi:10.1038/s41586-018-0680-3

MAO, Q. & UNADKAT, J. D. 2015. Role of the breast cancer resistance protein (BCRP/ABCG2) in drug transport--an update. *AAPS J,* 17, 65-82. doi:10.1208/s12248-014-9668-6

MARTINY, V. Y., CARBONELL, P., CHEVILLARD, F., MOROY, G., NICOT, A. B., VAYER, P., VILLOUTREIX, B. O. & MITEVA, M. A. 2015. Integrated structure- and ligand-based in silico approach to predict inhibition of cytochrome P450 2D6. *Bioinformatics,* 31, 3930-7. doi:10.1093/bioinformatics/btv486

MO, W. & ZHANG, J. T. 2012. Human ABCG2: structure, function, and its role in multidrug resistance. *Int J Biochem Mol Biol,* 3, 1-27.

MORTIER, J., RAKERS, C., BERMUDEZ, M., MURGUEITIO, M. S., RINIKER, S. & WOLBER, G. 2015. The impact of molecular dynamics on drug design: applications for the characterization of ligand-macromolecule complexes. *Drug Discov Today,* 20, 686-702. doi:10.1016/j.drudis.2015.01.003

MOUAWAD, L. & PERAHIA, D. 1993. Diagonalization in a Mixed Basis: A Method to Compute low-Frequency Normal Modes for large Macromolecules. *Biopolymers,* 33, 599-611 doi:https://doi.org/10.1002/bip.360330409

NAGY, T., TOTH, A., TELBISZ, A., SARKADI, B., TORDAI, H., TORDAI, A. & HEGEDUS, T. 2021. The transport pathway in the ABCG2 protein and its regulation revealed by molecular dynamics simulations. *Cell Mol Life Sci,* 78, 2329-2339. doi:10.1007/s00018-020-03651-3

OLDHAM, M. L., DAVIDSON, A. L. & CHEN, J. 2008. Structural insights into ABC transporter mechanism. *Curr Opin Struct Biol,* 18, 726-33. doi:10.1016/j.sbi.2008.09.007

ORLANDO, B. J. & LIAO, M. 2020. ABCG2 transports anticancer drugs via a closed-to-open switch. *Nat Commun,* 11, 2264. doi:10.1038/s41467-020-16155-2

OZVEGY, C., VARADI, A. & SARKADI, B. 2002. Characterization of drug transport, ATP hydrolysis, and nucleotide trapping by the human ABCG2 multidrug transporter. Modulation of substrate specificity by a point mutation. *J Biol Chem,* 277, 47980-90. doi:10.1074/jbc.M207857200

PERAHIA, D. & MOUAWAD, L. 1995. Computation of low-frequency normal modes in macromolecules: improvements to the method of diagonalization in a mixed basis and application to hemoglobin. *Comput Chem,* 19, 241-6. doi:10.1016/0097-8485(95)00011-g

PHILLIPS, J. C., HARDY, D. J., MAIA, J. D. C., STONE, J. E., RIBEIRO, J. V., BERNARDI, R. C., BUCH, R., FIORIN, G., HENIN, J., JIANG, W., MCGREEVY, R., MELO, M. C. R., RADAK, B. K., SKEEL, R. D., SINGHAROY, A., WANG, Y., ROUX,

B., AKSIMENTIEV, A., LUTHEY-SCHULTEN, Z., KALE, L. V., SCHULTEN, K., CHIPOT, C. & TAJKHORSHID, E. 2020. Scalable molecular dynamics on CPU and GPU architectures with NAMD. *J Chem Phys,* 153, 044130. doi:10.1063/5.0014475

PRUEKSARITANONT, T., CHU, X., GIBSON, C., CUI, D., YEE, K. L., BALLARD, J., CABALU, T. & HOCHMAN, J. 2013. Drug-drug interaction studies: regulatory guidance and an industry perspective. *AAPS J,* 15, 629-45. doi:10.1208/s12248-013-9470-x

ROBEY, R. W., TO, K. K., POLGAR, O., DOHSE, M., FETSCH, P., DEAN, M. & BATES, S. E. 2009. ABCG2: a perspective. *Adv Drug Deliv Rev,* 61, 3-13. doi:10.1016/j.addr.2008.11.003

SALINAS, M., KESSLER, P., DOUGUET, D., SARRAF, D., TONALI, N., THAI, R., SERVENT, D. & LINGUEGLIA, E. 2021. Mambalgin-1 pain-relieving peptide locks the hinge between alpha4 and alpha5 helices to inhibit rat acid-sensing ion channel 1a. *Neuropharmacology,* 185, 108453. doi:10.1016/j.neuropharm.2021.108453

SONDERGAARD, C. R., OLSSON, M. H., ROSTKOWSKI, M. & JENSEN, J. H. 2011. Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of pKa Values. *J Chem Theory Comput,* 7, 2284-95. doi:10.1021/ct200133y

SUN, H. & SCOTT, D. O. 2010. Structure-based drug metabolism predictions for drug design. *Chem Biol Drug Des,* 75, 3-17. doi:10.1111/j.1747-0285.2009.00899.x

SUZUKI, M., SUZUKI, H., SUGIMOTO, Y. & SUGIYAMA, Y. 2003. ABCG2 transports sulfated conjugates of steroids and xenobiotics. *J Biol Chem,* 278, 22644-9. doi:10.1074/jbc.M212399200

SZOLLOSI, D., ROSE-SPERLING, D., HELLMICH, U. A. & STOCKNER, T. 2018. Comparison of mechanistic transport cycle models of ABC exporters. *Biochim Biophys Acta Biomembr,* 1860, 818-832. doi:10.1016/j.bbamem.2017.10.028

TAYLOR, N. M. I., MANOLARIDIS, I., JACKSON, S. M., KOWAL, J., STAHLBERG, H. & LOCHER, K. P. 2017. Structure of the human multidrug transporter ABCG2. *Nature,* 546, 504-509. doi:10.1038/nature22345

TELBISZ, A., MULLER, M., OZVEGY-LACZKA, C., HOMOLYA, L., SZENTE, L., VARADI, A. & SARKADI, B. 2007. Membrane cholesterol selectively modulates the activity of the human ABCG2 multidrug transporter. *Biochim Biophys Acta,* 1768, 2698-713. doi:10.1016/j.bbamem.2007.06.026

THOMAS, C. & TAMPE, R. 2020. Structural and Mechanistic Principles of ABC Transporters. *Annu Rev Biochem,* 89, 605-636. doi:10.1146/annurev-biochem-011520-105201

TOYODA, Y., TAKADA, T. & SUZUKI, H. 2019. Inhibitors of Human ABCG2: From Technical Background to Recent Updates With Clinical Implications. *Front Pharmacol,* 10, 208. doi:10.3389/fphar.2019.00208

VANOMMESLAEGHE, K., HATCHER, E., ACHARYA, C., KUNDU, S., ZHONG, S., SHIM, J., DARIAN, E., GUVENCH, O., LOPES, P., VOROBYOV, I. & MACKERELL, A. D., JR. 2010. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J Comput Chem,* 31, 671-90. doi:10.1002/jcc.21367

YU, Q., NI, D., KOWAL, J., MANOLARIDIS, I., JACKSON, S. M., STAHLBERG, H. & LOCHER, K. P. 2021. Structures of ABCG2 under turnover conditions reveal a key step in the drug transport mechanism. *Nat Commun,* 12, 4376. doi:10.1038/s41467-021-24651-2

IV. Conclusions and Perspectives

„Ah vége, vége! Vagy ki tudja? Diák marad az ember, amíg él,
Leckéjét a sírig tanulja: Nehezebbet folyvást a réginél!”

“Oh, finally it’s over. Or is it? You’re a student for as long as you live,
Learning lessons till the very end: one harder than the one before.”

Reviczky Gyula

Drug metabolizing enzymes and transporters are important mediators in a variety of biological processes related to the action and safety of drugs. Drug elimination through metabolism and excretion is a complex process as was demonstrated in the previous chapters and is governed by drug metabolizing enzymes and drug transporters. The prediction of the fate of administrated drugs in the human body remains a challenging task that has long relied on expensive animal studies also raising concerns of ethical aspects and reliability. Therefore, an increasing number of computational approaches have been involved in order to efficiently predict the metabolic outcome of drugs and new drug candidates. Several studies have created prediction models using ligand-based or structure-based information, and in 2013, the host laboratory trained the first machine learning classification models that integrated ligand- and structure-based information for the prediction of the inhibition of different SULT isoforms, and later for CYP enzymes. Understanding the collective functional movements of the proteins and their interactions with small ligands is a crucial step in extracting structure-based information for such prediction models.

In my PhD work, I have focused on two essential phase II DMEs, SULT1A1 and UGT1A1, as well as on the drug efflux transporter ABCG2. My research addressed both the dynamics of the different proteins and their interactions with small molecules. This also included methodological developments, such as an enhanced molecular dynamics simulation tool and different prediction tools through the integration of structure-based and machine learning modeling.

Classical MD simulations and MD in combination with Normal Modes (MDeNM) were performed on SULT1A1 in order to elucidate the molecular mechanisms guiding the recognition of diverse substrates and inhibitors by the enzyme. MDeNM, being a multi-replica protocol designed to enhance conformational exploration in a subspace defined by a set of low-frequency NMs, allowed the exploration of an extended conformational space of the cofactor-bound SULT1A1, inaccessible by classical MD on the simulated timescale. Then I performed docking of the substrate molecules, estradiol and fulvestrant, on the generated enzyme conformational ensemble, and the results of the docking and subsequent classical MD simulations on the complexes demonstrated that large conformational changes of SULT1A1 could occur even in the presence of the cofactor, and that more widely open conformations can accommodate a large substrate like fulvestrant with higher affinity. The conformational exploration revealed that the loops L1, L2, and L3 exhibit an extremely high flexibility in the presence of the cofactor and the results of the docking simulations of a set of known substrates and inhibitors suggested that such significant structural adaptations may be sufficient to accommodate large ligands. Previously it was suggested that PAPS binding restricts the access for larger substrates to SULT1A1. However, the results presented in the thesis demonstrated that large conformational changes of SULT1A1 could occur even in the presence of the cofactor and SULT1A1 can accommodate large ligands independently of the cofactor movements. Altogether, my work shed new light on the complex mechanisms of substrate specificity and inhibition of SULT1A1, and has highlighted the utility of including MDeNM in protein-ligand

interactions studies where major rearrangements are expected. The generated conformations can be efficiently used to train *in silico* prediction models with even higher accuracy for larger molecules.

Structure-based modeling and ligand-based information were integrated to build the first machine learning prediction models of UGT1A1 inhibition. In my PhD work, I performed molecular dynamics simulations of the enzyme starting from a human homology model in the presence of the cofactor to explore the structural variability of the catalytic site, and observed significant flexibility, important for the accommodation of the diverse ligands. RMSD-based clustering of the generated conformational ensemble enabled me to extract a set of diverse representative enzyme conformations. They were used in docking simulations of a set of known substrates and inhibitors together with decoys, which helped to identify six conformations that can efficiently differentiate between actives and decoys. The loop regions in the substrate-binding cavity of these enzyme conformations exhibited highly different configurations, nevertheless the catalytically essential residues displayed conserved relative positions and a sub-pocket close to the catalytic site was detected that may be important for the wide substrate recognition. Machine learning models were then created using Random Forest and Support Vector Machine algorithms which integrated a rational selection of ligand-based descriptors together with information on enzyme-ligand interactions. The trained models showed excellent performance and were implemented in the DrugME software. DrugME can be efficiently used to identify new UGT1A1 inhibitors and can be helpful for the prediction of drug-drug interactions of new drug candidates and safety treatments while they also provide structural information on enzyme-ligand interactions.

In the part of my PhD work focusing on ABCG2, I have developed an innovative enhanced MD simulation technique, termed ketMD (kinetically excited targeted Molecular Dynamics) which uses kinetic excitation to promote protein movements corresponding to large conformational changes towards a specified target structure, without biasing the potential energy function. I employed ketMD to simulate the ABCG2 transport cycle together with the translocation of the physiological substrate, estrone 3-sulfate. With the help of ketMD, I successfully simulated the conformational transitions between the two extreme states of ABCG2 and revealed the complex molecular mechanism of the E_1S substrate translocation. I observed a valve-like function of two phenylalanine residues (F439 and F439') in cavity 1 that initially engage in stacking interactions against the substrate. The closure of the valve prevented backwards movements of the substrate towards the cytosol, and I identified crucial interactions along the translocation pathway. Classical MD simulations and Normal Mode Analysis revealed that the presence of the substrate in cavity 1 is necessary to couple the movements of the transmembrane domain to the nucleotide-binding domain. I also observed that cavity 2 was never completely collapsed, at any stages of the transport cycle which might enable simultaneous substrate binding in cavities 1 and 2 and accelerate the export rate of the transporter. The work on ABCG2 shed new light on the complex molecular mechanism of its transport and highlighted the utility of including enhanced *in silico* sampling techniques, such

as ketMD, in transporter studies. In the near future, the generated conformations along the transition pathway will be used to develop new machine learning models for the prediction of ABCG2 inhibitors and substrates in order to probe new drug candidates for multi-drug resistance and predict drug-drug interactions.

During the three years of my PhD work I had the great chance to deepen my knowledge in the biology of drug metabolism, elimination, and drug-drug interactions and also in the application and development of new *in silico* approaches to study the enzymes and transporters that govern these complex processes. The observations and results of the work presented in the thesis contribute to the better understanding of the molecular mechanisms of phase II metabolizing enzymes and ABC transporters, and their interactions with ligands. In particular, the DrugME software that now incorporates the developed predictive models integrating structure-based and machine learning modeling will greatly contribute to the future predictions of drug-drug interactions.

V. References

„Most, mikor ugyanúgy, mint mindig,
legfőbb ideje, hogy.”

“Now, it is the same as ever.
It’s high time that…”

Tandori Dezső

References

1. DiMasi, J.A., H.G. Grabowski, and R.W. Hansen (2016) *Innovation in the pharmaceutical industry: New estimates of R&D costs.* J Health Econ. 47:20-33. https://doi.org/10.1016/j.jhealeco.2016.01.012
2. Zhang, Z. and W. Tang (2018) *Drug metabolism in drug discovery and development.* Acta Pharm Sin B. 8:721-732. https://doi.org/10.1016/j.apsb.2018.04.003
3. Vidal, R.S., et al. (2018) *Metabolic Reprogramming During Multidrug Resistance in Leukemias.* Front Oncol. 8:90. https://doi.org/10.3389/fonc.2018.00090
4. Kaur, G., et al. (2020) *Drug-metabolizing enzymes: role in drug resistance in cancer.* Clin Transl Oncol. 22:1667-1680. https://doi.org/10.1007/s12094-020-02325-7
5. Almazroo, O.A., M.K. Miah, and R. Venkataramanan (2017) *Drug Metabolism in the Liver.* Clin Liver Dis. 21:1-20. https://doi.org/10.1016/j.cld.2016.08.001
6. Meyer, U.A. (1996) *Overview of enzymes of drug metabolism.* J Pharmacokinet Biopharm. 24:449-59. https://doi.org/10.1007/BF02353473
7. Xu, C., C.Y. Li, and A.N. Kong (2005) *Induction of phase I, II and III drug metabolism/transport by xenobiotics.* Arch Pharm Res. 28:249-68. https://doi.org/10.1007/BF02977789
8. Dowty, M.E., et al., *ADME, ADMET for Medicinal Chemists*. 2011, ISBN 9780470484074
9. Shah, P., et al. (2020) *Predicting liver cytosol stability of small molecules.* J Cheminform. 12:21. https://doi.org/10.1186/s13321-020-00426-7
10. de Waziers, I., et al. (1990) *Cytochrome P 450 isoenzymes, epoxide hydrolase and glutathione transferases in rat and human hepatic and extrahepatic tissues.* J Pharmacol Exp Ther. 253:387-94.
11. Cribb, A.E., et al. (2005) *The endoplasmic reticulum in xenobiotic toxicity.* Drug Metab Rev. 37:405-42. https://doi.org/10.1080/03602530500205135
12. Di, L. and E. Kerns, *Drug-Like Properties: Concepts, Structure Design and Methods from ADME to Toxicity Optimization*. 2010, Academic Press, ISBN 9780128010761
13. Sneader, W. (2000) *The discovery of aspirin: a reappraisal.* BMJ. 321:1591-4. https://doi.org/10.1136/bmj.321.7276.1591
14. Neuvonen, P.J., J.T. Backman, and M. Niemi (2008) *Pharmacokinetic comparison of the potential over-the-counter statins simvastatin, lovastatin, fluvastatin and pravastatin.* Clin Pharmacokinet. 47:463-74. https://doi.org/10.2165/00003088-200847070-00003
15. Hacker, M., W.S. Messer, and K.A. Machmann, *Pharmacology: Principles and Practice*. 2009, Academic Press
16. Stoll, F., A.H. Goller, and A. Hillisch (2011) *Utility of protein structures in overcoming ADMET-related issues of drug-like compounds.* Drug Discov Today. 16:530-8. https://doi.org/10.1016/j.drudis.2011.04.008
17. Pirmohamed, M., N.R. Kitteringham, and B.K. Park (1994) *The role of active metabolites in drug toxicity.* Drug Saf. 11:114-44. https://doi.org/10.2165/00002018-199411020-00006
18. Gibson, G.G. and P. Skett, *Introduction to Drug Metabolism*. 1986, Chapman and Hall, London, ISBN 978-0-412-26390-3
19. Ionescu, C. and M.R. Caira, *Drug Metabolism: Current Concepts*. 2005, Springer, Netherlands, ISBN 978-1-4020-4141-9
20. Testa, B., A. Pedretti, and G. Vistoli (2012) *Reactions and enzymes in the metabolism of drugs and other xenobiotics.* Drug Discov Today. 17:549-60. https://doi.org/10.1016/j.drudis.2012.01.017
21. Guengerich, F.P. (2008) *Cytochrome p450 and chemical toxicology.* Chem Res Toxicol. 21:70-83. https://doi.org/10.1021/tx700079z
22. Williams, J.A., et al. (2004) *Drug-drug interactions for UDP-glucuronosyltransferase substrates: a pharmacokinetic explanation for typically observed low exposure (AUCi/AUC) ratios.* Drug Metab Dispos. 32:1201-8. https://doi.org/10.1124/dmd.104.000794
23. Guengerich, F.P. (2006) *A malleable catalyst dominates the metabolism of drugs.* Proc Natl Acad Sci U S A. 103:13565-6. https://doi.org/10.1073/pnas.0606333103
24. Nebert, D.W. and D.W. Russell (2002) *Clinical importance of the cytochromes P450.* Lancet. 360:1155-62. https://doi.org/10.1016/S0140-6736(02)11203-7
25. Zanger, U.M. and M. Schwab (2013) *Cytochrome P450 enzymes in drug metabolism: regulation of gene expression, enzyme activities, and impact of genetic variation.* Pharmacol Ther. 138:103-41. https://doi.org/10.1016/j.pharmthera.2012.12.007
26. Wienkers, L.C. and T.G. Heath (2005) *Predicting in vivo drug interactions from in vitro drug discovery data.* Nat Rev Drug Discov. 4:825-33. https://doi.org/10.1038/nrd1851
27. Isvoran, A., et al. (2017) *Pharmacogenomics of the cytochrome P450 2C family: impacts of amino acid variations on drug metabolism.* Drug Discov Today. 22:366-376. https://doi.org/10.1016/j.drudis.2016.09.015

28. Zanger, U.M., et al. (2001) *Comprehensive analysis of the genetic factors determining expression and function of hepatic CYP2D6.* Pharmacogenetics. 11:573-85. https://doi.org/10.1097/00008571-200110000-00004
29. Martiny, V.Y. and M.A. Miteva (2013) *Advances in molecular modeling of human cytochrome P450 polymorphism.* J Mol Biol. 425:3978-92. https://doi.org/10.1016/j.jmb.2013.07.010
30. Werk, A.N. and I. Cascorbi (2014) *Functional gene variants of CYP3A4.* Clin Pharmacol Ther. 96:340-8. https://doi.org/10.1038/clpt.2014.129
31. Zhou, S.F., et al. (2010) *Structure, function, regulation and polymorphism and the clinical significance of human cytochrome P450 1A2.* Drug Metab Rev. 42:268-354. https://doi.org/10.3109/03602530903286476
32. Louet, M., et al. (2018) *Insights into molecular mechanisms of drug metabolism dysfunction of human CYP2C9*30.* PLoS One. 13:e0197249. https://doi.org/10.1371/journal.pone.0197249
33. Lee, C.R., J.A. Goldstein, and J.A. Pieper (2002) *Cytochrome P450 2C9 polymorphisms: a comprehensive review of the in-vitro and human data.* Pharmacogenetics. 12:251-63. https://doi.org/10.1097/00008571-200204000-00010
34. Rowland, P., et al. (2006) *Crystal structure of human cytochrome P450 2D6.* J Biol Chem. 281:7614-22. https://doi.org/10.1074/jbc.M511232200
35. Ferguson, C.S. and R.F. Tyndale (2011) *Cytochrome P450 enzymes in the brain: emerging evidence of biological significance.* Trends Pharmacol Sci. 32:708-14. https://doi.org/10.1016/j.tips.2011.08.005
36. Martiny, V.Y., et al. (2015) *Integrated structure- and ligand-based in silico approach to predict inhibition of cytochrome P450 2D6.* Bioinformatics. 31:3930-7. https://doi.org/10.1093/bioinformatics/btv486
37. Poulsen, L.L. and D.M. Ziegler (1979) *The liver microsomal FAD-containing monooxygenase. Spectral characterization and kinetic studies.* J Biol Chem. 254:6449-55.
38. Eswaramoorthy, S., et al. (2006) *Mechanism of action of a flavin-containing monooxygenase.* Proc Natl Acad Sci U S A. 103:9832-7. https://doi.org/10.1073/pnas.0602398103
39. Edmondson, D.E. and C. Binda, *Monoamine Oxidases*. Membrane Protein Complexes: Structure and Function. 2018, Springer Singapore, ISBN 978-981-10-7756-2
40. Hille, R. (2005) *Molybdenum-containing hydroxylases.* Arch Biochem Biophys. 433:107-16. https://doi.org/10.1016/j.abb.2004.08.012
41. Sarmiento-Pavia, P.D. and M.E. Sosa-Torres (2021) *Bioinorganic insights of the PQQ-dependent alcohol dehydrogenases.* J Biol Inorg Chem. 26:177-203. https://doi.org/10.1007/s00775-021-01852-0
42. Edenberg, H.J. and J.N. McClintick (2018) *Alcohol Dehydrogenases, Aldehyde Dehydrogenases, and Alcohol Use Disorders: A Critical Review.* Alcohol Clin Exp Res. 42:2281-2297. https://doi.org/10.1111/acer.13904
43. Barski, O.A., S.M. Tipparaju, and A. Bhatnagar (2008) *The aldo-keto reductase superfamily and its role in drug metabolism and detoxification.* Drug Metab Rev. 40:553-624. https://doi.org/10.1080/03602530802431439
44. Pey, A.L., C.F. Megarity, and D.J. Timson (2019) *NAD(P)H quinone oxidoreductase (NQO1): an enzyme which needs just enough mobility, in just the right places.* Biosci Rep. 39 https://doi.org/10.1042/BSR20180459
45. Prabha, M., V. Ravi, and N. Ramachandra Swamy (2013) *Activity of hydrolytic enzymes in various regions of normal human brain tissue.* Indian J Clin Biochem. 28:283-91. https://doi.org/10.1007/s12291-012-0273-0
46. Khojasteh, S.C., H. Wong, and C.E.C.A. Hop, *Drug Metabolism and Pharmacokinetics Quick Guide*. 2011, Springer New York, NY, ISBN 978-1-4419-5628-6
47. Benedetti, M.S., et al. (2009) *Drug metabolism and pharmacokinetics.* Drug Metab Rev. 41:344-90. https://doi.org/10.1080/10837450902891295
48. Shimomura, K., et al. (1971) *Analgesic effect of morphine glucuronides.* Tohoku J Exp Med. 105:45-52. https://doi.org/10.1620/tjem.105.45
49. Bauer, T.M., et al. (1995) *Prolonged sedation due to accumulation of conjugated metabolites of midazolam.* Lancet. 346:145-7. https://doi.org/10.1016/s0140-6736(95)91209-6
50. Olson, J.A., et al. (1992) *Enhancement of biological activity by conjugation reactions.* J Nutr. 122:615-24. https://doi.org/10.1093/jn/122.suppl_3.615
51. Croom, E. (2012) *Metabolism of xenobiotics of human environments.* Prog Mol Biol Transl Sci. 112:31-88. https://doi.org/10.1016/B978-0-12-415813-9.00003-9
52. Kaivosaari, S., M. Finel, and M. Koskinen (2011) *N-glucuronidation of drugs and other xenobiotics by human and animal UDP-glucuronosyltransferases.* Xenobiotica. 41:652-69. https://doi.org/10.3109/00498254.2011.563327

53. Gonzalez, F.J., M. Coughtrie, and R.H. Tukey, *Drug Metabolism*, in *The Pharmacological Basis of Therapeutics*, L.L. Bruton, R. Hilal-Dandan, and B.C. Knollmann, (Eds.), 2017, Goodman & Gilman's, ISBN 978-0071624428
54. Falany, C.N. (1991) *Molecular enzymology of human liver cytosolic sulfotransferases.* Trends Pharmacol Sci. 12:255-9. https://doi.org/10.1016/0165-6147(91)90566-b
55. Mueller, J.W., et al. (2015) *The Regulation of Steroid Action by Sulfation and Desulfation.* Endocr Rev. 36:526-63. https://doi.org/10.1210/er.2015-1036
56. Visser, T.J. (1994) *Role of sulfation in thyroid hormone metabolism.* Chem Biol Interact. 92:293-303. https://doi.org/10.1016/0009-2797(94)90071-x
57. Coughtrie, M.W.H. (2016) *Function and organization of the human cytosolic sulfotransferase (SULT) family.* Chem Biol Interact. 259:2-7. https://doi.org/10.1016/j.cbi.2016.05.005
58. Gamage, N., et al. (2006) *Human sulfotransferases and their role in chemical metabolism.* Toxicol Sci. 90:5-22. https://doi.org/10.1093/toxsci/kfj061
59. Edavana, V.K., et al. (2012) *Sulfation of 4-hydroxy toremifene: individual variability, isoform specificity, and contribution to toremifene pharmacogenomics.* Drug Metab Dispos. 40:1210-5. https://doi.org/10.1124/dmd.111.044040
60. Isvoran, A., et al. (2022) *Pharmacogenetics of human sulfotransferases and impact of amino acid exchange on Phase II drug metabolism.* Drug Discov Today. 27:103349. https://doi.org/10.1016/j.drudis.2022.103349
61. Marto, N., et al. (2017) *Implications of sulfotransferase activity in interindividual variability in drug response: clinical perspective on current knowledge.* Drug Metab Rev. 49:357-371. https://doi.org/10.1080/03602532.2017.1335749
62. Duffel, M.W., et al. (2001) *Enzymatic aspects of the phenol (aryl) sulfotransferases.* Drug Metab Rev. 33:369-95. https://doi.org/10.1081/dmr-120001394
63. Riches, Z., et al. (2009) *Quantitative evaluation of the expression and activity of five major sulfotransferases (SULTs) in human tissues: the SULT "pie".* Drug Metab Dispos. 37:2255-61. https://doi.org/10.1124/dmd.109.028399
64. Hebbring, S.J., et al. (2007) *Human SULT1A1 gene: copy number differences and functional implications.* Hum Mol Genet. 16:463-70. https://doi.org/10.1093/hmg/ddl468
65. Cook, I., et al. (2013) *The gate that governs sulfotransferase selectivity.* Biochemistry. 52:415-24. https://doi.org/10.1021/bi301492j
66. Marto, N., et al. (2021) *A simple method to measure sulfonation in man using paracetamol as probe drug.* Sci Rep. 11:9036. https://doi.org/10.1038/s41598-021-88393-3
67. Hui, Y., et al. (2015) *Sulfation of afimoxifene, endoxifen, raloxifene, and fulvestrant by the human cytosolic sulfotransferases (SULTs): A systematic analysis.* J Pharmacol Sci. 128:144-9. https://doi.org/10.1016/j.jphs.2015.06.004
68. Glatt, H. (2000) *Sulfotransferases in the bioactivation of xenobiotics.* Chem Biol Interact. 129:141-70. https://doi.org/10.1016/s0009-2797(00)00202-7
69. Hildebrandt, M.A., et al. (2004) *Human SULT1A3 pharmacogenetics: gene duplication and functional genomic studies.* Biochem Biophys Res Commun. 321:870-8. https://doi.org/10.1016/j.bbrc.2004.07.038
70. Brix, L.A., et al. (1999) *Analysis of the substrate specificity of human sulfotransferases SULT1A1 and SULT1A3: site-directed mutagenesis and kinetic studies.* Biochemistry. 38:10474-9. https://doi.org/10.1021/bi990795q
71. Richard, K., et al. (2001) *Sulfation of thyroid hormone and dopamine during human development: ontogeny of phenol sulfotransferases and arylsulfatase in liver, lung, and brain.* J Clin Endocrinol Metab. 86:2734-42. https://doi.org/10.1210/jcem.86.6.7569
72. Fujita, K., et al. (1997) *Molecular cloning and characterization of rat ST1B1 and human ST1B2 cDNAs, encoding thyroid hormone sulfotransferases.* J Biochem. 122:1052-61. https://doi.org/10.1093/oxfordjournals.jbchem.a021846
73. Kurogi, K., et al. (2021) *SULT genetic polymorphisms: physiological, pharmacological and clinical implications.* Expert Opin Drug Metab Toxicol. 17:767-784. https://doi.org/10.1080/17425255.2021.1940952
74. Glatt, H., et al. (2001) *Human cytosolic sulphotransferases: genetics, characteristics, toxicological aspects.* Mutat Res. 482:27-40. https://doi.org/10.1016/s0027-5107(01)00207-x
75. Song, W.C. (2001) *Biochemistry and reproductive endocrinology of estrogen sulfotransferase.* Ann N Y Acad Sci. 948:43-50. https://doi.org/10.1111/j.1749-6632.2001.tb03985.x
76. Huang, J., et al. (2010) *Kinetic analysis of bile acid sulfation by stably expressed human sulfotransferase 2A1 (SULT2A1).* Xenobiotica. 40:184-94. https://doi.org/10.3109/00498250903514607

77. Rowland, A., J.O. Miners, and P.I. Mackenzie (2013) *The UDP-glucuronosyltransferases: their role in drug metabolism and detoxification.* Int J Biochem Cell Biol. 45:1121-32. https://doi.org/10.1016/j.biocel.2013.02.019
78. Ohno, S. and S. Nakajin (2009) *Determination of mRNA expression of human UDP-glucuronosyltransferases and application for localization in various human tissues by real-time reverse transcriptase-polymerase chain reaction.* Drug Metab Dispos. 37:32-40. https://doi.org/10.1124/dmd.108.023598
79. Mathews, C.K. and K.E. Van Holde, *Biochemistry, second edition*. 1996, The Benjamin-Cummings Publishing Company, Menlo Park, California ISBN 9780805339314
80. Burchell, B., et al. (1991) *The UDP glucuronosyltransferase gene superfamily: suggested nomenclature based on evolutionary divergence.* DNA Cell Biol. 10:487-94. https://doi.org/10.1089/dna.1991.10.487
81. Meech, R., et al. (2019) *The UDP-Glycosyltransferase (UGT) Superfamily: New Members, New Functions, and Novel Paradigms.* Physiol Rev. 99:1153-1222. https://doi.org/10.1152/physrev.00058.2017
82. Fisher, M.B., et al. (2001) *The role of hepatic and extrahepatic UDP-glucuronosyltransferases in human drug metabolism.* Drug Metab Rev. 33:273-97. https://doi.org/10.1081/dmr-120000653
83. Csala, M., et al. (2007) *Transport and transporters in the endoplasmic reticulum.* Biochim Biophys Acta. 1768:1325-41. https://doi.org/10.1016/j.bbamem.2007.03.009
84. Bosma, P.J., et al. (1994) *Bilirubin UDP-glucuronosyltransferase 1 is the only relevant bilirubin glucuronidating isoform in man.* J Biol Chem. 269:17960-4.
85. Bosma, P.J. (2003) *Inherited disorders of bilirubin metabolism.* J Hepatol. 38:107-17. https://doi.org/10.1016/s0168-8278(02)00359-8
86. Desai, A.A., F. Innocenti, and M.J. Ratain (2003) *UGT pharmacogenomics: implications for cancer risk and cancer therapeutics.* Pharmacogenetics. 13:517-23. https://doi.org/10.1097/01.fpc.0000054116.14659.e5
87. Nagar, S. and R.L. Blanchard (2006) *Pharmacogenetics of uridine diphosphoglucuronosyltransferase (UGT) 1A family members and its role in patient response to irinotecan.* Drug Metab Rev. 38:393-409. https://doi.org/10.1080/03602530600739835
88. Andrade, R.J., et al. (2009) *Pharmacogenomics in drug induced liver injury.* Curr Drug Metab. 10:956-70. https://doi.org/10.2174/138920009790711805
89. Iyer, L., et al. (1999) *Phenotype-genotype correlation of in vitro SN-38 (active metabolite of irinotecan) and bilirubin glucuronidation in human liver tissue with UGT1A1 promoter polymorphism.* Clin Pharmacol Ther. 65:576-82. https://doi.org/10.1016/S0009-9236(99)70078-0
90. Posner, J., et al. (1989) *The pharmacokinetics of lamotrigine (BW430C) in healthy subjects with unconjugated hyperbilirubinaemia (Gilbert's syndrome).* Br J Clin Pharmacol. 28:117-20. https://doi.org/10.1111/j.1365-2125.1989.tb03514.x
91. Marques, S.C. and O.N. Ikediobi (2010) *The clinical application of UGT1A1 pharmacogenetic testing: gene-environment interactions.* Hum Genomics. 4:238-49. https://doi.org/10.1186/1479-7364-4-4-238
92. Radominska-Pandya, A., et al. (1999) *Structural and functional studies of UDP-glucuronosyltransferases.* Drug Metab Rev. 31:817-99. https://doi.org/10.1081/dmr-100101944
93. Ritter, J.K. (2007) *Intestinal UGTs as potential modifiers of pharmacokinetics and biological responses to drugs and xenobiotics.* Expert Opin Drug Metab Toxicol. 3:93-107. https://doi.org/10.1517/17425255.3.1.93
94. Lautala, P., et al. (2000) *The specificity of glucuronidation of entacapone and tolcapone by recombinant human UDP-glucuronosyltransferases.* Drug Metab Dispos. 28:1385-9.
95. Ma, L., et al. (2012) *Glucuronidation of edaravone by human liver and kidney microsomes: biphasic kinetics and identification of UGT1A9 as the major UDP-glucuronosyltransferase isoform.* Drug Metab Dispos. 40:734-41. https://doi.org/10.1124/dmd.111.043356
96. Kuang, Y., et al. (2021) *Glabrone as a specific UGT1A9 probe substrate and its application in discovering the inhibitor glycycoumarin.* Eur J Pharm Sci. 161:105786. https://doi.org/10.1016/j.ejps.2021.105786
97. Coffman, B.L., et al. (1997) *Human UGT2B7 catalyzes morphine glucuronidation.* Drug Metab Dispos. 25:1-4.
98. Sun, D., et al. (2007) *Glucuronidation of active tamoxifen metabolites by the human UDP glucuronosyltransferases.* Drug Metab Dispos. 35:2006-14. https://doi.org/10.1124/dmd.107.017145
99. Menard, V., et al. (2013) *Expression of UGT2B7 is driven by two mutually exclusive promoters and alternative splicing in human tissues: changes from prenatal life to adulthood and in kidney cancer.* Pharmacogenet Genomics. 23:684-96. https://doi.org/10.1097/FPC.0000000000000008
100. Mitchell, S.C. (2020) *N-acetyltransferase: the practical consequences of polymorphic activity in man.* Xenobiotica. 50:77-91. https://doi.org/10.1080/00498254.2019.1618511

101. Knights, K.M., M.J. Sykes, and J.O. Miners (2007) *Amino acid conjugation: contribution to the metabolism and toxicity of xenobiotic carboxylic acids.* Expert Opin Drug Metab Toxicol. 3:159-68. https://doi.org/10.1517/17425255.3.2.159
102. Hayes, J.D., J.U. Flanagan, and I.R. Jowsey (2005) *Glutathione transferases.* Annu Rev Pharmacol Toxicol. 45:51-88. https://doi.org/10.1146/annurev.pharmtox.45.120403.095857
103. Lefevre, P.G., *The Present State of the Carrier Hypothesis*, in *Current Topics in Membranes and Transport*, F. Bronner and A. Kleinzeller, (Eds.), 1975, Academic Press Inc, ISBN 9780121533076
104. Petzinger, E. and J. Geyer (2006) *Drug transporters in pharmacokinetics.* Naunyn Schmiedebergs Arch Pharmacol. 372:465-75. https://doi.org/10.1007/s00210-006-0042-9
105. Agarwal, S., L. Chinn, and L. Zhang (2013) *An overview of transporter information in package inserts of recently approved new molecular entities.* Pharm Res. 30:899-910. https://doi.org/10.1007/s11095-012-0924-0
106. Lee, S.C., et al. (2017) *Evaluation of transporters in drug development: Current status and contemporary issues.* Adv Drug Deliv Rev. 116:100-118. https://doi.org/10.1016/j.addr.2017.07.020
107. International Transporter, C., et al. (2010) *Membrane transporters in drug development.* Nat Rev Drug Discov. 9:215-36. https://doi.org/10.1038/nrd3028
108. Tweedie, D., et al. (2013) *Transporter studies in drug development: experience to date and follow-up on decision trees from the International Transporter Consortium.* Clin Pharmacol Ther. 94:113-25. https://doi.org/10.1038/clpt.2013.77
109. Hillgren, K.M., et al. (2013) *Emerging transporters of clinical importance: an update from the International Transporter Consortium.* Clin Pharmacol Ther. 94:52-63. https://doi.org/10.1038/clpt.2013.74
110. Brouwer, K.L., et al. (2013) *In vitro methods to support transporter evaluation in drug discovery and development.* Clin Pharmacol Ther. 94:95-112. https://doi.org/10.1038/clpt.2013.81
111. Giacomini, K.M., A. Galetin, and S.M. Huang (2018) *The International Transporter Consortium: Summarizing Advances in the Role of Transporters in Drug Development.* Clin Pharmacol Ther. 104:766-771. https://doi.org/10.1002/cpt.1224
112. Zamek-Gliszczynski, M.J., et al. (2018) *Transporters in Drug Development: 2018 ITC Recommendations for Transporters of Emerging Clinical Importance.* Clin Pharmacol Ther. 104:890-899. https://doi.org/10.1002/cpt.1112
113. EMA, Committee for Human Medicinal Products, *Guideline on the investigation of drug interactions*, CPMP/EWP/560/95/Rev. 1 Corr. 2**, 2012, London, UK
114. FDA, Center for Drug Evaluation and Research, *In Vitro Drug Interaction Studies - Cytochrome P450 Enzyme- and Transporter-Mediated Drug Interactions - Guidance for Industry*, 2020, Silver Spring, MD, USA
115. Mao, Q., Y. Lai, and J. Wang (2018) *Drug Transporters in Xenobiotic Disposition and Pharmacokinetic Prediction.* Drug Metab Dispos. 46:561-566. https://doi.org/10.1124/dmd.118.081356
116. *Drug Transporters in Drug Disposition, Effects and Toxicity*. Advances in Experimental Medicine and Biology, ed. X. Liu and G. Pan. 2019, Springer Singapore, ISBN 978-981-13-7646-7
117. Pizzagalli, M.D., A. Bensimon, and G. Superti-Furga (2021) *A guide to plasma membrane solute carrier proteins.* FEBS J. 288:2784-2835. https://doi.org/10.1111/febs.15531
118. Colas, C., P.M. Ung, and A. Schlessinger (2016) *SLC Transporters: Structure, Function, and Drug Discovery.* Medchemcomm. 7:1069-1081. https://doi.org/10.1039/C6MD00005C
119. Stein, W.D. and T. Litman, *Channels, Carriers, and Pumps*. 2015, Academic Press, ISBN 9780124165793
120. Cesar-Razquin, A., et al. (2015) *A Call for Systematic Research on Solute Carriers.* Cell. 162:478-87. https://doi.org/10.1016/j.cell.2015.07.022
121. Rask-Andersen, M., et al. (2013) *Solute carriers as drug targets: current use, clinical trials and prospective.* Mol Aspects Med. 34:702-10. https://doi.org/10.1016/j.mam.2012.07.015
122. Salphati, L. (2009) *Transport-metabolism interplay.* Mol Pharm. 6:1629-30. https://doi.org/10.1021/mp900266r
123. Benet, L.Z. (2009) *The drug transporter-metabolism alliance: uncovering and defining the interplay.* Mol Pharm. 6:1631-43. https://doi.org/10.1021/mp900253n
124. Zhou, F., et al. (2017) *Recent advance in the pharmacogenomics of human Solute Carrier Transporters (SLCs) in drug disposition.* Adv Drug Deliv Rev. 116:21-36. https://doi.org/10.1016/j.addr.2016.06.004
125. Shu, Y., et al. (2007) *Effect of genetic variation in the organic cation transporter 1 (OCT1) on metformin action.* J Clin Invest. 117:1422-31. https://doi.org/10.1172/JCI30558
126. Koepsell, H. (2020) *Organic Cation Transporters in Health and Disease.* Pharmacol Rev. 72:253-319. https://doi.org/10.1124/pr.118.015578
127. Burckhardt, G. (2012) *Drug transport by Organic Anion Transporters (OATs).* Pharmacol Ther. 136:106-30. https://doi.org/10.1016/j.pharmthera.2012.07.010

128. Nigam, S.K., et al. (2015) *The organic anion transporter (OAT) family: a systems biology perspective.* Physiol Rev. 95:83-123. https://doi.org/10.1152/physrev.00025.2013
129. Ganapathy, V. and F.H. Leibach (1986) *Carrier-mediated reabsorption of small peptides in renal proximal tubule.* Am J Physiol. 251:F945-53. https://doi.org/10.1152/ajprenal.1986.251.6.F945
130. Ganapathy and F.H. Leibach (1985) *Is intestinal peptide transport energized by a proton gradient?* Am J Physiol. 249:G153-60. https://doi.org/10.1152/ajpgi.1985.249.2.G153
131. Terada, T. and K. Inui (2012) *Recent advances in structural biology of peptide transporters.* Curr Top Membr. 70:257-74. https://doi.org/10.1016/B978-0-12-394316-3.00008-9
132. Smith, D.E., B. Clemencon, and M.A. Hediger (2013) *Proton-coupled oligopeptide transporter family SLC15: physiological, pharmacological and pathological implications.* Mol Aspects Med. 34:323-36. https://doi.org/10.1016/j.mam.2012.11.003
133. Otsuka, M., et al. (2005) *A human transporter protein that mediates the final excretion step for toxic organic cations.* Proc Natl Acad Sci U S A. 102:17923-8. https://doi.org/10.1073/pnas.0506483102
134. American Diabetes, A. (2022) *Introduction: Standards of Medical Care in Diabetes-2022.* Diabetes Care. 45:S1-S2. https://doi.org/10.2337/dc22-Sint
135. Damme, K., et al. (2011) *Mammalian MATE (SLC47A) transport proteins: impact on efflux of endogenous substrates and xenobiotics.* Drug Metab Rev. 43:499-523. https://doi.org/10.3109/03602532.2011.602687
136. Dean, M., A. Rzhetsky, and R. Allikmets (2001) *The human ATP-binding cassette (ABC) transporter superfamily.* Genome Res. 11:1156-66. https://doi.org/10.1101/gr.184901
137. Wilkens, S. (2015) *Structure and mechanism of ABC transporters.* F1000Prime Rep. 7:14. https://doi.org/10.12703/P7-14
138. Stieger, B. and B. Gao (2015) *Drug transporters in the central nervous system.* Clin Pharmacokinet. 54:225-42. https://doi.org/10.1007/s40262-015-0241-y
139. Kaminski, W.E., A. Piehler, and J.J. Wenzel (2006) *ABC A-subfamily transporters: structure, function and disease.* Biochim Biophys Acta. 1762:510-24. https://doi.org/10.1016/j.bbadis.2006.01.011
140. Bossaerts, L., R. Cacace, and C. Van Broeckhoven (2022) *The role of ATP-binding cassette subfamily A in the etiology of Alzheimer's disease.* Mol Neurodegener. 17:31. https://doi.org/10.1186/s13024-022-00536-w
141. Slot, A.J., S.V. Molinski, and S.P. Cole (2011) *Mammalian multidrug-resistance proteins (MRPs).* Essays Biochem. 50:179-207. https://doi.org/10.1042/bse0500179
142. Favre, G., et al. (2017) *The ABCC6 Transporter: A New Player in Biomineralization.* Int J Mol Sci. 18 https://doi.org/10.3390/ijms18091941
143. Jemnitz, K., et al. (2010) *ABCC2/Abcc2: a multispecific transporter with dominant excretory functions.* Drug Metab Rev. 42:402-36. https://doi.org/10.3109/03602530903491741
144. Ritter, C.A., et al. (2005) *Cellular export of drugs and signaling molecules by the ATP-binding cassette transporters MRP4 (ABCC4) and MRP5 (ABCC5).* Drug Metab Rev. 37:253-78. https://doi.org/10.1081/dmr-200047984
145. Terry, S.F. (2020) *The Human Face of ABCC6.* FEBS Lett. 594:4151-4157. https://doi.org/10.1002/1873-3468.14002
146. Bisaccia, F., et al. (2021) *Structural and Functional Characterization of the ABCC6 Transporter in Hepatic Cells: Role on PXE, Cancer Therapy and Drug Resistance.* Int J Mol Sci. 22 https://doi.org/10.3390/ijms22062858
147. Hunt, J.F., C. Wang, and R.C. Ford (2013) *Cystic fibrosis transmembrane conductance regulator (ABCC7) structure.* Cold Spring Harb Perspect Med. 3:a009514. https://doi.org/10.1101/cshperspect.a009514
148. Kruh, G.D., et al. (2007) *ABCC10, ABCC11, and ABCC12.* Pflugers Arch. 453:675-84. https://doi.org/10.1007/s00424-006-0114-1
149. Mao, Q. and J.D. Unadkat (2015) *Role of the breast cancer resistance protein (BCRP/ABCG2) in drug transport--an update.* AAPS J. 17:65-82. https://doi.org/10.1208/s12248-014-9668-6
150. Doyle, L. and D.D. Ross (2003) *Multidrug resistance mediated by the breast cancer resistance protein BCRP (ABCG2).* Oncogene. 22:7340-58. https://doi.org/10.1038/sj.onc.1206938
151. Xu, J., H. Peng, and J.T. Zhang (2007) *Human multidrug transporter ABCG2, a target for sensitizing drug resistance in cancer chemotherapy.* Curr Med Chem. 14:689-701. https://doi.org/10.2174/092986707780059580
152. Pena-Solorzano, D., et al. (2017) *ABCG2/BCRP: Specific and Nonspecific Modulators.* Med Res Rev. 37:987-1050. https://doi.org/10.1002/med.21428
153. Mao, Q. and J.D. Unadkat (2005) *Role of the breast cancer resistance protein (ABCG2) in drug transport.* AAPS J. 7:E118-33. https://doi.org/10.1208/aapsj070112

154. Leveque, D., et al. (2010) *[Mechanisms of pharmacokinetic drug-drug interactions].* Rev Med Interne. 31:170-9. https://doi.org/10.1016/j.revmed.2009.07.009
155. Manikandan, P. and S. Nagini (2018) *Cytochrome P450 Structure, Function and Clinical Significance: A Review*. Curr Drug Targets. 19:38-54. https://doi.org/10.2174/1389450118666170125144557
156. Lin, J.H. (2006) *CYP induction-mediated drug interactions: in vitro assessment and clinical implications.* Pharm Res. 23:1089-116. https://doi.org/10.1007/s11095-006-0277-7
157. Labrecque, M.P., G.G. Prefontaine, and T.V. Beischlag (2013) *The aryl hydrocarbon receptor nuclear translocator (ARNT) family of proteins: transcriptional modifiers with multi-functional protein interfaces.* Curr Mol Med. 13:1047-65. https://doi.org/10.2174/15665240113139990042
158. Kalra, B.S. (2007) *Cytochrome P450 enzyme isoforms and their therapeutic implications: an update.* Indian J Med Sci. 61:102-16.
159. Kumar, S., R. Sharma, and A. Roychowdhury (2012) *Modulation of cytochrome-P450 inhibition (CYP) in drug discovery: a medicinal chemistry perspective.* Curr Med Chem. 19:3605-21. https://doi.org/10.2174/092986712801323180
160. Hakkola, J., et al. (2020) *Inhibition and induction of CYP enzymes in humans: an update.* Arch Toxicol. 94:3671-3722. https://doi.org/10.1007/s00204-020-02936-7
161. Boxenbaum, H. (1999) *Cytochrome P450 3A4 in vivo ketoconazole competitive inhibition: determination of Ki and dangers associated with high clearance drugs in general.* J Pharm Pharm Sci. 2:47-52.
162. James, M.O. and S. Ambadapadi (2013) *Interactions of cytosolic sulfotransferases with xenobiotics.* Drug Metab Rev. 45:401-14. https://doi.org/10.3109/03602532.2013.835613
163. Miners, J.O. and P.I. Mackenzie (1991) *Drug glucuronidation in humans.* Pharmacol Ther. 51:347-69. https://doi.org/10.1016/0163-7258(91)90065-t
164. Samara, E.E., et al. (1997) *Effect of valproate on the pharmacokinetics and pharmacodynamics of lorazepam.* J Clin Pharmacol. 37:442-50. https://doi.org/10.1002/j.1552-4604.1997.tb04322.x
165. Muller, F. and M.F. Fromm (2011) *Transporter-mediated drug-drug interactions.* Pharmacogenomics. 12:1017-37. https://doi.org/10.2217/pgs.11.44
166. Fenner, K.S., et al. (2009) *Drug-drug interactions mediated through P-glycoprotein: clinical relevance and in vitro-in vivo correlation using digoxin as a probe drug.* Clin Pharmacol Ther. 85:173-81. https://doi.org/10.1038/clpt.2008.195
167. Greiner, B., et al. (1999) *The role of intestinal P-glycoprotein in the interaction of digoxin and rifampin.* J Clin Invest. 104:147-53. https://doi.org/10.1172/JCI6663
168. Kruijtzer, C.M., et al. (2002) *Increased oral bioavailability of topotecan in combination with the breast cancer resistance protein and P-glycoprotein inhibitor GF120918.* J Clin Oncol. 20:2943-50. https://doi.org/10.1200/JCO.2002.12.116
169. Kyrklund, C., et al. (2003) *Gemfibrozil increases plasma pravastatin concentrations and reduces pravastatin renal clearance.* Clin Pharmacol Ther. 73:538-44. https://doi.org/10.1016/S0009-9236(03)00052-3
170. Schneck, D.W., et al. (2004) *The effect of gemfibrozil on the pharmacokinetics of rosuvastatin.* Clin Pharmacol Ther. 75:455-63. https://doi.org/10.1016/j.clpt.2003.12.014
171. Ferlay, J., et al. *Global Cancer Observatory: Cancer Today. Lyon, France: International Agency for Research on Cancer. Available from: https://gco.iarc.fr/today, accessed 2021 February*. 2020.
172. Bukowski, K., M. Kciuk, and R. Kontek (2020) *Mechanisms of Multidrug Resistance in Cancer Chemotherapy.* Int J Mol Sci. 21 https://doi.org/10.3390/ijms21093233
173. Pathania, S., et al. (2018) *Drug metabolizing enzymes and their inhibitors' role in cancer resistance.* Biomed Pharmacother. 105:53-65. https://doi.org/10.1016/j.biopha.2018.05.117
174. Li, Y., et al. (2017) *Tumoral expression of drug and xenobiotic metabolizing enzymes in breast cancer patients of different ethnicities with implications to personalized medicine.* Sci Rep. 7:4747. https://doi.org/10.1038/s41598-017-04250-2
175. Osborne, M.J., et al. (2019) *Overcoming Drug Resistance through the Development of Selective Inhibitors of UDP-Glucuronosyltransferase Enzymes.* J Mol Biol. 431:258-272. https://doi.org/10.1016/j.jmb.2018.11.007
176. Pirmohamed, M. (2001) *Pharmacogenetics and pharmacogenomics.* Br J Clin Pharmacol. 52:345-7. https://doi.org/10.1046/j.0306-5251.2001.01498.x
177. Yan, L. and R.A. Beckman (2005) *Pharmacogenetics and pharmacogenomics in oncology therapeutic antibody development.* Biotechniques. 39:565-8.
178. McGraw, J. and D. Waller (2012) *Cytochrome P450 variations in different ethnic populations.* Expert Opin Drug Metab Toxicol. 8:371-82. https://doi.org/10.1517/17425255.2012.657626

179. King, B.P., et al. (2004) *Upstream and coding region CYP2C9 polymorphisms: correlation with warfarin dose and metabolism.* Pharmacogenetics. 14:813-22. https://doi.org/10.1097/00008571-200412000-00004
180. Allabi, A.C., J.L. Gala, and Y. Horsmans (2005) *CYP2C9, CYP2C19, ABCB1 (MDR1) genetic polymorphisms and phenytoin metabolism in a Black Beninese population.* Pharmacogenet Genomics. 15:779-86. https://doi.org/10.1097/01.fpc.0000174787.92861.91
181. Zhou, S.F. (2009) *Polymorphism of human cytochrome P450 2D6 and its clinical significance: Part I.* Clin Pharmacokinet. 48:689-723. https://doi.org/10.2165/11318030-000000000-00000
182. Chen, B.H., et al. (2015) *Mechanism of sulfotransferase pharmacogenetics in altered xenobiotic metabolism.* Expert Opin Drug Metab Toxicol. 11:1053-71. https://doi.org/10.1517/17425255.2015.1045486
183. Lazarus, P., et al. (2009) *Potential role of UGT pharmacogenetics in cancer treatment and prevention: focus on tamoxifen.* Ann N Y Acad Sci. 1155:99-111. https://doi.org/10.1111/j.1749-6632.2009.04114.x
184. Guillemette, C. (2003) *Pharmacogenomics of human UDP-glucuronosyltransferase enzymes.* Pharmacogenomics J. 3:136-58. https://doi.org/10.1038/sj.tpj.6500171
185. Hu, D.G., et al. (2016) *Genetic polymorphisms of human UDP-glucuronosyltransferase (UGT) genes and cancer risk.* Drug Metab Rev. 48:47-69. https://doi.org/10.3109/03602532.2015.1131292
186. Carulli, N., et al. (1976) *Alteration of drug metabolism in Gilbert's syndrome.* Gut. 17:581-7. https://doi.org/10.1136/gut.17.8.581
187. Bruckmueller, H. and I. Cascorbi (2021) *ABCB1, ABCG2, ABCC1, ABCC2, and ABCC3 drug transporter polymorphisms and their impact on drug bioavailability: what is our current understanding?* Expert Opin Drug Metab Toxicol. 17:369-396. https://doi.org/10.1080/17425255.2021.1876661
188. Schaefer, M., I. Roots, and T. Gerloff (2006) *In-vitro transport characteristics discriminate wild-type ABCB1 (MDR1) from ALA893SER and ALA893THR polymorphisms.* Pharmacogenet Genomics. 16:855-61. https://doi.org/10.1097/01.fpc.0000230113.03710.34
189. Whirl-Carrillo, M., et al. (2012) *Pharmacogenomics knowledge for personalized medicine.* Clin Pharmacol Ther. 92:414-7. https://doi.org/10.1038/clpt.2012.96
190. Imai, Y., et al. (2002) *C421A polymorphism in the human breast cancer resistance protein gene is associated with low expression of Q141K protein and low-level drug resistance.* Mol Cancer Ther. 1:611-6.
191. Morisaki, K., et al. (2005) *Single nucleotide polymorphisms modify the transporter activity of ABCG2.* Cancer Chemother Pharmacol. 56:161-72. https://doi.org/10.1007/s00280-004-0931-x
192. Mizuarai, S., N. Aozasa, and H. Kotani (2004) *Single nucleotide polymorphisms result in impaired membrane localization and reduced atpase activity in multidrug transporter ABCG2.* Int J Cancer. 109:238-46. https://doi.org/10.1002/ijc.11669
193. Toyoda, Y., et al. (2019) *Functional Characterization of Clinically-Relevant Rare Variants in ABCG2 Identified in a Gout and Hyperuricemia Cohort.* Cells. 8 https://doi.org/10.3390/cells8040363
194. Sarkadi, B., L. Homolya, and T. Hegedus (2020) *The ABCG2/BCRP transporter and its variants - from structure to pathology.* FEBS Lett. 594:4012-4034. https://doi.org/10.1002/1873-3468.13947
195. Ekroos, M. and T. Sjogren (2006) *Structural basis for ligand promiscuity in cytochrome P450 3A4.* Proc Natl Acad Sci U S A. 103:13682-7. https://doi.org/10.1073/pnas.0603236103
196. Nair, P.C., R.A. McKinnon, and J.O. Miners (2016) *Cytochrome P450 structure-function: insights from molecular dynamics simulations.* Drug Metab Rev. 48:434-52. https://doi.org/10.1080/03602532.2016.1178771
197. Karplus, M. and J.A. McCammon (2002) *Molecular dynamics simulations of biomolecules.* Nat Struct Biol. 9:646-52. https://doi.org/10.1038/nsb0902-646
198. Sridhar, J., et al. (2017) *Review of Ligand Specificity Factors for CYP1A Subfamily Enzymes from Molecular Modeling Studies Reported to-Date.* Molecules. 22 https://doi.org/10.3390/molecules22071143
199. Banu, H., N. Renuka, and G. Vasanthakumar (2011) *Reduced catalytic activity of human CYP2C9 natural alleles for gliclazide: molecular dynamics simulation and docking studies.* Biochimie. 93:1028-36. https://doi.org/10.1016/j.biochi.2011.02.008
200. Sano, E., et al. (2010) *Mechanism of the decrease in catalytic activity of human cytochrome P450 2C9 polymorphic variants investigated by computational analysis.* J Comput Chem. 31:2746-58. https://doi.org/10.1002/jcc.21568
201. Lertkiatmongkol, P., et al. (2013) *Distal effect of amino acid substitutions in CYP2C9 polymorphic variants causes differences in interatomic interactions against (S)-warfarin.* PLoS One. 8:e74053. https://doi.org/10.1371/journal.pone.0074053
202. Zhang, T., et al. (2011) *Long-range effects of a peripheral mutation on the enzymatic activity of cytochrome P450 1A2.* J Chem Inf Model. 51:1336-46. https://doi.org/10.1021/ci200112b

203. Wilderman, P.R., et al. (2012) *Investigation by site-directed mutagenesis of the role of cytochrome P450 2B4 non-active-site residues in protein-ligand interactions based on crystal structures of the ligand-bound enzyme.* FEBS J. 279:1607-20. https://doi.org/10.1111/j.1742-4658.2011.08411.x
204. Zhou, Y.H., et al. (2006) *On the human CYP2C9*13 variant activity reduction: a molecular dynamics simulation and docking study.* Biochimie. 88:1457-65. https://doi.org/10.1016/j.biochi.2006.05.001
205. Cui, Y.L. and R.L. Wu (2017) *Molecular dynamics investigations of membrane-bound CYP2C19 polymorphisms reveal distinct mechanisms for peripheral variants by long-range effects on the enzymatic activity.* Mol Biosyst. 13:1070-1079. https://doi.org/10.1039/c6mb00827e
206. Fukuyoshi, S., et al. (2016) *Molecular Dynamics Simulations to Investigate the Influences of Amino Acid Mutations on Protein Three-Dimensional Structures of Cytochrome P450 2D6.1, 2, 10, 14A, 51, and 62.* PLoS One. 11:e0152946. https://doi.org/10.1371/journal.pone.0152946
207. Yadav, A., et al. (2022) *Mining of molecular insights of CYP2A6 and its variants complex with coumarin (CYP2A6*-coumarin) using molecular dynamics simulation.* J Biomol Struct Dyn:1-12. https://doi.org/10.1080/07391102.2022.2062785
208. Mustafa, G., et al. (2019) *Influence of Transmembrane Helix Mutations on Cytochrome P450-Membrane Interactions and Function.* Biophys J. 116:419-432. https://doi.org/10.1016/j.bpj.2018.12.014
209. Mancini, G. and C. Zazza (2015) *F429 Regulation of Tunnels in Cytochrome P450 2B4: A Top Down Study of Multiple Molecular Dynamics Simulations.* PLoS One. 10:e0137075. https://doi.org/10.1371/journal.pone.0137075
210. Kobayashi, K., et al. (2014) *Evaluation of influence of single nucleotide polymorphisms in cytochrome P450 2B6 on substrate recognition using computational docking and molecular dynamics simulation.* PLoS One. 9:e96789. https://doi.org/10.1371/journal.pone.0096789
211. Kato, K., et al. (2021) *Deciphering Structural Alterations Associated with Activity Reductions of Genetic Polymorphisms in Cytochrome P450 2A6 Using Molecular Dynamics Simulations.* Int J Mol Sci. 22 https://doi.org/10.3390/ijms221810119
212. Cui, Y.L., et al. (2013) *Molecular dynamic investigations of the mutational effects on structural characteristics and tunnel geometry in CYP17A1.* J Chem Inf Model. 53:3308-17. https://doi.org/10.1021/ci400553w
213. Skopalik, J., P. Anzenbacher, and M. Otyepka (2008) *Flexibility of human cytochromes P450: molecular dynamics reveals differences between CYPs 3A4, 2C9, and 2A6, which correlate with their substrate preferences.* J Phys Chem B. 112:8165-73. https://doi.org/10.1021/jp800311c
214. Mustafa, G., et al. (2019) *Differing Membrane Interactions of Two Highly Similar Drug-Metabolizing Cytochrome P450 Isoforms: CYP 2C9 and CYP 2C19.* Int J Mol Sci. 20 https://doi.org/10.3390/ijms20184328
215. Cui, Y.L., F. Xu, and R. Wu (2016) *Molecular dynamics investigations of regioselectivity of anionic/aromatic substrates by a family of enzymes: a case study of diclofenac binding in CYP2C isoforms.* Phys Chem Chem Phys. 18:17428-39. https://doi.org/10.1039/c6cp01128d
216. Ma, G., et al. (2020) *Molecular Basis for Metabolic Regioselectivity and Mechanism of Cytochrome P450s toward Carcinogenic 4-(Methylnitrosamino)-(3-pyridyl)-1-butanone.* Chem Res Toxicol. 33:436-447. https://doi.org/10.1021/acs.chemrestox.9b00353
217. Wright, W.C., et al. (2020) *Clobetasol Propionate Is a Heme-Mediated Selective Inhibitor of Human Cytochrome P450 3A5.* J Med Chem. 63:1415-1433. https://doi.org/10.1021/acs.jmedchem.9b02067
218. Fischer, A. and M. Smiesko (2021) *A Conserved Allosteric Site on Drug-Metabolizing CYPs: A Systematic Computational Assessment.* Int J Mol Sci. 22 https://doi.org/10.3390/ijms222413215
219. Cook, I., T. Wang, and T.S. Leyh (2015) *Sulfotransferase 1A1 Substrate Selectivity: A Molecular Clamp Mechanism.* Biochemistry. 54:6114-22. https://doi.org/10.1021/acs.biochem.5b00406
220. Wang, T., I. Cook, and T.S. Leyh (2017) *The NSAID allosteric site of human cytosolic sulfotransferases.* J Biol Chem. 292:20305-20312. https://doi.org/10.1074/jbc.M117.817387
221. Isvoran, A., et al. (2022) *Pharmacogenetics of human sulfotransferases and impact of amino acid exchange on Phase II drug metabolism.* Drug Discov Today:103349. https://doi.org/10.1016/j.drudis.2022.103349
222. Rakers, C., et al. (2016) *In Silico Prediction of Human Sulfotransferase 1E1 Activity Guided by Pharmacophores from Molecular Dynamics Simulations.* J Biol Chem. 291:58-71. https://doi.org/10.1074/jbc.M115.685610
223. Laakkonen, L. and M. Finel (2010) *A molecular model of the human UDP-glucuronosyltransferase 1A1, its membrane orientation, and the interactions between different parts of the enzyme.* Mol Pharmacol. 77:931-9. https://doi.org/10.1124/mol.109.063289
224. Li, C. and Q. Wu (2007) *Adaptive evolution of multiple-variable exons and structural diversity of drug-metabolizing enzymes.* BMC Evol Biol. 7:69. https://doi.org/10.1186/1471-2148-7-69

225. Locuson, C.W. and T.S. Tracy (2007) *Comparative modelling of the human UDP-glucuronosyltransferases: insights into structure and mechanism.* Xenobiotica. 37:155-68. https://doi.org/10.1080/00498250601129109
226. Nair, P.C., et al. (2020) *Arginine-259 of UGT2B7 Confers UDP-Sugar Selectivity.* Mol Pharmacol. 98:710-718. https://doi.org/10.1124/molpharm.120.000104
227. Fujiwara, R., et al. (2009) *In silico and in vitro approaches to elucidate the thermal stability of human UDP-glucuronosyltransferase (UGT) 1A9.* Drug Metab Pharmacokinet. 24:235-44. https://doi.org/10.2133/dmpk.24.235
228. Subedi, P., et al. (2022) *Insight into glucocorticoids glucosylation by glucosyltransferase: A combined experimental and in-silico approach.* Biophysical Chemistry. 289 https://doi.org/10.1016/j.bpc.2022.106875
229. Zhang, X., et al. (2012) *Twelve transmembrane helices form the functional core of mammalian MATE1 (multidrug and toxin extruder 1) protein.* J Biol Chem. 287:27971-82. https://doi.org/10.1074/jbc.M112.386979
230. Tsigelny, I.F., et al. (2008) *Modeling of glycerol-3-phosphate transporter suggests a potential 'tilt' mechanism involved in its function.* J Bioinform Comput Biol. 6:885-904. https://doi.org/10.1142/s0219720008003801
231. Tsigelny, I.F., et al. (2011) *Conformational changes of the multispecific transporter organic anion transporter 1 (OAT1/SLC22A6) suggests a molecular mechanism for initial stages of drug and metabolite transport.* Cell Biochem Biophys. 61:251-9. https://doi.org/10.1007/s12013-011-9191-7
232. Adla, S.K., et al. (2021) *Neurosteroids: Structure-Uptake Relationships and Computational Modeling of Organic Anion Transporting Polypeptides (OATP)1A2.* Molecules. 26 https://doi.org/10.3390/molecules26185662
233. Gebauer, L., et al. (2022) *Molecular basis for stereoselective transport of fenoterol by the organic cation transporters 1 and 2.* Biochem Pharmacol. 197:114871. https://doi.org/10.1016/j.bcp.2021.114871
234. Shahraki, O., et al. (2018) *Molecular dynamics simulation and molecular docking studies of 1,4-Dihydropyridines as P-glycoprotein's allosteric inhibitors.* J Biomol Struct Dyn. 36:112-125. https://doi.org/10.1080/07391102.2016.1268976
235. Prajapati, R., et al. (2013) *In silico model for P-glycoprotein substrate prediction: insights from molecular dynamics and in vitro studies.* J Comput Aided Mol Des. 27:347-63. https://doi.org/10.1007/s10822-013-9650-x
236. Domicevica, L. and P.C. Biggin (2015) *Homology modelling of human P-glycoprotein.* Biochem Soc Trans. 43:952-8. https://doi.org/10.1042/BST20150125
237. Gadhe, C.G., G. Kothandan, and S.J. Cho (2013) *In silico study of desmosdumotin as an anticancer agent: homology modeling, docking and molecular dynamics simulation approach.* Anticancer Agents Med Chem. 13:1636-44. https://doi.org/10.2174/18715206113139990302
238. Hosseini Balef, S.S., et al. (2019) *In vitro and in silico evaluation of P-glycoprotein inhibition through (99m) Tc-methoxyisobutylisonitrile uptake.* Chem Biol Drug Des. 93:283-289. https://doi.org/10.1111/cbdd.13411
239. Laszlo, L., B. Sarkadi, and T. Hegedus (2016) *Jump into a New Fold-A Homology Based Model for the ABCG2/BCRP Multidrug Transporter.* PLoS One. 11:e0164426. https://doi.org/10.1371/journal.pone.0164426
240. Mora Lagares, L., et al. (2021) *Structure-Function Relationships in the Human P-Glycoprotein (ABCB1): Insights from Molecular Dynamics Simulations.* Int J Mol Sci. 23 https://doi.org/10.3390/ijms23010362
241. Xing, J., et al. (2020) *Computational Insights into Allosteric Conformational Modulation of P-Glycoprotein by Substrate and Inhibitor Binding.* Molecules. 25 https://doi.org/10.3390/molecules25246006
242. Zhang, B., et al. (2021) *Simultaneous binding mechanism of multiple substrates for multidrug resistance transporter P-glycoprotein.* Phys Chem Chem Phys. 23:4530-4543. https://doi.org/10.1039/d0cp05910b
243. Wang, L., et al. (2019) *Molecular Energetics of Doxorubicin Pumping by Human P-Glycoprotein.* J Chem Inf Model. 59:3889-3898. https://doi.org/10.1021/acs.jcim.9b00429
244. Wang, L. and Y. Sun (2020) *Efflux mechanism and pathway of verapamil pumping by human P-glycoprotein.* Arch Biochem Biophys. 696:108675. https://doi.org/10.1016/j.abb.2020.108675
245. Zhang, Y., et al. (2019) *Exploring movement and energy in human P-glycoprotein conformational rearrangement.* J Biomol Struct Dyn. 37:1104-1119. https://doi.org/10.1080/07391102.2018.1461133
246. Zhang, J., et al. (2015) *Interaction of P-glycoprotein with anti-tumor drugs: the site, gate and pathway.* Soft Matter. 11:6633-41. https://doi.org/10.1039/c5sm01028d
247. Domicevica, L., H. Koldso, and P.C. Biggin (2018) *Multiscale molecular dynamics simulations of lipid interactions with P-glycoprotein in a complex membrane.* J Mol Graph Model. 80:147-156. https://doi.org/10.1016/j.jmgm.2017.12.022

248. Barreto-Ojeda, E., et al. (2018) *Coarse-grained molecular dynamics simulations reveal lipid access pathways in P-glycoprotein.* J Gen Physiol. 150:417-429. https://doi.org/10.1085/jgp.201711907
249. Behmard, E., et al. (2022) *Efflux dynamics of the antiseizure drug, levetiracetam, through the P-glycoprotein channel revealed by advanced comparative molecular simulations.* Sci Rep. 12:13674. https://doi.org/10.1038/s41598-022-17994-3
250. Vesga, L.C., et al. (2021) *Tetrahydroquinoline/4,5-Dihydroisoxazole Molecular Hybrids as Inhibitors of Breast Cancer Resistance Protein (BCRP/ABCG2).* ChemMedChem. 16:2686-2694. https://doi.org/10.1002/cmdc.202100188
251. Zhang, Y.K., et al. (2019) *Regorafenib antagonizes BCRP-mediated multidrug resistance in colon cancer.* Cancer Lett. 442:104-112. https://doi.org/10.1016/j.canlet.2018.10.032
252. Zattoni, I.F., et al. (2022) *A new porphyrin as selective substrate-based inhibitor of breast cancer resistance protein (BCRP/ABCG2).* Chem Biol Interact. 351:109718. https://doi.org/10.1016/j.cbi.2021.109718
253. Ibrahim, M.A.A., et al. (2022) *In Silico Targeting Human Multidrug Transporter ABCG2 in Breast Cancer: Database Screening, Molecular Docking, and Molecular Dynamics Study.* Mol Inform. 41:e2060039. https://doi.org/10.1002/minf.202060039
254. Ibrahim, M.A.A., et al. (2021) *Prospective Drug Candidates as Human Multidrug Transporter ABCG2 Inhibitors: an In Silico Drug Discovery Study.* Cell Biochem Biophys. 79:189-200. https://doi.org/10.1007/s12013-021-00985-y
255. Ibrahim, M.A.A., et al. (2022) *Naturally occurring plant-based anticancerous candidates as prospective ABCG2 inhibitors: an in silico drug discovery study.* Mol Divers https://doi.org/10.1007/s11030-022-10389-6
256. Tadayon, M. and Z. Garkani-Nejad (2019) *In silico study combining QSAR, docking and molecular dynamics simulation on 2,4-disubstituted pyridopyrimidine derivatives.* J Recept Signal Transduct Res. 39:167-174. https://doi.org/10.1080/10799893.2019.1641821
257. Wang, J.Q., et al. (2020) *Reversal of Cancer Multidrug Resistance (MDR) Mediated by ATP-Binding Cassette Transporter G2 (ABCG2) by AZ-628, a RAF Kinase Inhibitor.* Front Cell Dev Biol. 8:601400. https://doi.org/10.3389/fcell.2020.601400
258. Nagy, T., et al. (2021) *The transport pathway in the ABCG2 protein and its regulation revealed by molecular dynamics simulations.* Cell Mol Life Sci. 78:2329-2339. https://doi.org/10.1007/s00018-020-03651-3
259. Islam, M.A., et al. (2022) *Identification of Potential Cytochrome P450 3A5 Inhibitors: An Extensive Virtual Screening through Molecular Docking, Negative Image-Based Screening, Machine Learning and Molecular Dynamics Simulation Studies.* International Journal of Molecular Sciences. 23 https://doi.org/10.3390/ijms23169374
260. Luirink, R.A., et al. (2018) *A combined computational and experimental study on selective flucloxacillin hydroxylation by cytochrome P450 BM3 variants.* J Inorg Biochem. 184:115-122. https://doi.org/10.1016/j.jinorgbio.2018.04.013
261. Chelli, S.M., P. Gupta, and S.K. Belliraj (2019) *An in silico design of bioavailability for kinase inhibitors evaluating the mechanistic rationale in the CYP metabolism of erlotinib.* J Mol Model. 25:65. https://doi.org/10.1007/s00894-018-3917-z
262. Goldwaser, E., et al. (2022) *Machine learning-driven identification of drugs inhibiting cytochrome P450 2C9.* PLoS Comput Biol. 18:e1009820. https://doi.org/10.1371/journal.pcbi.1009820
263. Magalhães, R.P., H.S. Fernandes, and S.F. Sousa (2020) *Modelling Enzymatic Mechanisms with QM/MM Approaches: Current Status and Future Challenges.* Israel Journal of Chemistry. 60:655-666. https://doi.org/10.1002/ijch.202000014
264. Tyzack, J.D. and J. Kirchmair (2019) *Computational methods and tools to predict cytochrome P450 metabolism for drug discovery.* Chem Biol Drug Des. 93:377-386. https://doi.org/10.1111/cbdd.13445
265. Shaik, S., et al. (2010) *P450 enzymes: their structure, reactivity, and selectivity-modeled by QM/MM calculations.* Chem Rev. 110:949-1017. https://doi.org/10.1021/cr900121s
266. Friesner, R.A. and V. Guallar (2005) *Ab initio quantum chemical and mixed quantum mechanics/molecular mechanics (QM/MM) methods for studying enzymatic catalysis.* Annu Rev Phys Chem. 56:389-427. https://doi.org/10.1146/annurev.physchem.55.091602.094410
267. Shaik, S., et al., *QM/MM Studies of Structure and Reactivity of Cytochrome P450 Enzymes: Methodology and Selected Applications*, in *Drug Metabolism Prediction*, 2014, ISBN 9783527335664978352733566497835273356649
268. Borowski, T., M. Quesne, and M. Szaleniec (2015) *QM and QM/MM Methods Compared: Case Studies on Reaction Mechanisms of Metalloenzymes.* Adv Protein Chem Struct Biol. 100:187-224. https://doi.org/10.1016/bs.apcsb.2015.06.005
269. Lonsdale, R. and A.J. Mulholland (2014) *QM/MM modelling of drug-metabolizing enzymes.* Curr Top Med Chem. 14:1339-47. https://doi.org/10.2174/1568026614666140506114859

270. Sgrignani, J., et al. (2016) *Enzymatic and Inhibition Mechanism of Human Aromatase (CYP19A1) Enzyme. A Computational Perspective from QM/MM and Classical Molecular Dynamics Simulations.* Mini Rev Med Chem. 16:1112-24. https://doi.org/10.2174/1389557516666160623101129
271. Rydberg, P., F.S. Jorgensen, and L. Olsen (2014) *Use of density functional theory in drug metabolism studies.* Expert Opin Drug Metab Toxicol. 10:215-27. https://doi.org/10.1517/17425255.2014.864278
272. Balding, P.R., et al. (2008) *How do azoles inhibit cytochrome P450 enzymes? A density functional study.* J Phys Chem A. 112:12911-8. https://doi.org/10.1021/jp802087w
273. Leach, A.G. and N.J. Kidley (2011) *Quantitatively interpreted enhanced inhibition of cytochrome P450s by heteroaromatic rings containing nitrogen.* J Chem Inf Model. 51:1048-63. https://doi.org/10.1021/ci2000506
274. Lee, J.Y., N.S. Kang, and Y.K. Kang (2012) *Binding free energies of inhibitors to iron porphyrin complex as a model for Cytochrome P450.* Biopolymers. 97:219-28. https://doi.org/10.1002/bip.22009
275. Ma, G., et al. (2021) *Investigating the molecular mechanism of hydroxylated bromdiphenyl ethers to inhibit the thyroid hormone sulfotransferase SULT1A1.* Chemosphere. 263:128353. https://doi.org/10.1016/j.chemosphere.2020.128353
276. Hong, H., et al. (2008) *Mold(2), molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics.* J Chem Inf Model. 48:1337-44. https://doi.org/10.1021/ci800038f
277. Chandrasekaran, B., et al., *Computer-Aided Prediction of Pharmacokinetic (ADMET) Properties*, in *Dosage Form Design Parameters*, R.K. Tekade, Editor, 2018, Academic Press, ISBN 9780128144213
278. Seo, M., et al. (2020) *Development of Natural Compound Molecular Fingerprint (NC-MFP) with the Dictionary of Natural Products (DNP) for natural product-based drug development.* J Cheminform. 12:6. https://doi.org/10.1186/s13321-020-0410-3
279. Roy, K., S. Kar, and R.N. Das, *Statistical Methods in QSAR/QSPR*, in *A Primer on QSAR/QSPR Modeling*, 2015, Springer, Cham, ISBN 978-3-319-17280-4
280. Kato, H. (2020) *Computational prediction of cytochrome P450 inhibition and induction.* Drug Metab Pharmacokinet. 35:30-44. https://doi.org/10.1016/j.dmpk.2019.11.006
281. Roy, K. and P.P. Roy (2008) *Comparative QSAR studies of CYP1A2 inhibitor flavonoids using 2D and 3D descriptors.* Chem Biol Drug Des. 72:370-82. https://doi.org/10.1111/j.1747-0285.2008.00717.x
282. Chohan, K.K., et al. (2005) *A rapid computational filter for cytochrome P450 1A2 inhibition potential of compound libraries.* J Med Chem. 48:5154-61. https://doi.org/10.1021/jm048959a
283. Lozano, R., A. Frutos, and A. Martinez (2021) *In Silico Model for Predicting CYP2D6-Mediated Drug-Drug Interactions.* Curr Rev Clin Exp Pharmacol. 16:124-127. https://doi.org/10.2174/1574884715666200507130824
284. Lather, V. and M.X. Fernandes (2011) *Comparative QSAR analyses of competitive CYP2C9 inhibitors using three-dimensional molecular descriptors.* Chem Biol Drug Des. 78:112-23. https://doi.org/10.1111/j.1747-0285.2011.01106.x
285. Roy, K. and P. Pratim Roy (2009) *Comparative chemometric modeling of cytochrome 3A4 inhibitory activity of structurally diverse compounds using stepwise MLR, FA-MLR, PLS, GFA, G/PLS and ANN techniques.* Eur J Med Chem. 44:2913-22. https://doi.org/10.1016/j.ejmech.2008.12.004
286. AlQudah, D.A., M.A. Zihlif, and M.O. Taha (2016) *Ligand-based modeling of diverse aryalkylamines yields new potent P-glycoprotein inhibitors.* Eur J Med Chem. 110:204-23. https://doi.org/10.1016/j.ejmech.2016.01.034
287. Vazquez, R.N., et al. (2014) *Molecular factors influencing the affinity of flavonoid compounds on P-glycoprotein efflux transporter.* Curr Comput Aided Drug Des. 10:250-8. https://doi.org/10.2174/1573409910031503022311140
288. Shen, J., et al. (2014) *A genetic algorithm- back propagation artificial neural network model to quantify the affinity of flavonoids toward P-glycoprotein.* Comb Chem High Throughput Screen. 17:162-72. https://doi.org/10.2174/1386207311301010002
289. Lee, Y., et al. (2013) *Computational analysis and predictive modeling of polymorph descriptors.* Chem Cent J. 7:23. https://doi.org/10.1186/1752-153X-7-23
290. Byvatov, E., et al. (2007) *A Virtual Screening Filter for Identification of Cytochrome P450 2C9 (CYP2C9) Inhibitors.* QSAR & Combinatorial Science. 26:618-628. https://doi.org/10.1002/qsar.200630143
291. Ekins, S., et al. (1999) *Three- and four-dimensional quantitative structure activity relationship analyses of cytochrome P-450 3A4 inhibitors.* J Pharmacol Exp Ther. 290:429-38.
292. Wanchana, S., F. Yamashita, and M. Hashida (2003) *QSAR analysis of the inhibition of recombinant CYP 3A4 activity by structurally diverse compounds using a genetic algorithm-combined partial least squares method.* Pharm Res. 20:1401-8. https://doi.org/10.1023/a:1025702009611

293. Taskinen, J., et al. (2003) *Conjugation of catechols by recombinant human sulfotransferases, UDP-glucuronosyltransferases, and soluble catechol O-methyltransferase: structure-conjugation relationships and predictive models.* Drug Metab Dispos. 31:1187-97. https://doi.org/10.1124/dmd.31.9.1187
294. Adenot, M. and R. Lahana (2004) *Blood-brain barrier permeation models: discriminating between potential CNS and non-CNS drugs including P-glycoprotein substrates.* J Chem Inf Comput Sci. 44:239-48. https://doi.org/10.1021/ci034205d
295. Zhou, X.F., et al. (2005) *Quantitative structure-activity relationship and quantitative structure-pharmacokinetics relationship of 1,4-dihydropyridines and pyridines as multidrug resistance modulators.* Pharm Res. 22:1989-96. https://doi.org/10.1007/s11095-005-8112-0
296. Huang, S., et al. (2022) *Development of Simple and Accurate in Silico Ligand-Based Models for Predicting ABCG2 Inhibition.* Front Chem. 10:863146. https://doi.org/10.3389/fchem.2022.863146
297. Le, M.T., et al. (2021) *Prediction model of human ABCC2/MRP2 efflux pump inhibitors: a QSAR study.* Mol Divers. 25:741-751. https://doi.org/10.1007/s11030-020-10047-9
298. Plonka, W., et al. (2021) *CYPlebrity: Machine learning models for the prediction of inhibitors of cytochrome P450 enzymes.* Bioorg Med Chem. 46:116388. https://doi.org/10.1016/j.bmc.2021.116388
299. Zheng, M., et al. (2009) *Site of metabolism prediction for six biotransformations mediated by cytochromes P450.* Bioinformatics. 25:1251-8. https://doi.org/10.1093/bioinformatics/btp140
300. Sicho, M., et al. (2017) *FAME 2: Simple and Effective Machine Learning Model of Cytochrome P450 Regioselectivity.* J Chem Inf Model. 57:1832-1846. https://doi.org/10.1021/acs.jcim.7b00250
301. Mazzolari, A., et al. (2019) *Prediction of UGT-mediated Metabolism Using the Manually Curated MetaQSAR Database.* ACS Med Chem Lett. 10:633-638. https://doi.org/10.1021/acsmedchemlett.8b00603
302. Labute, P. (1999) *Binary QSAR: a new method for the determination of quantitative structure activity relationships.* Pac Symp Biocomput:444-55. https://doi.org/10.1142/9789814447300_0044
303. Schwaha, R. and G.F. Ecker (2011) *Use of shape similarities for the classification of P-glycoprotein substrates and nonsubstrates.* Future Med Chem. 3:1117-28. https://doi.org/10.4155/fmc.11.58
304. Poongavanam, V., N. Haider, and G.F. Ecker (2012) *Fingerprint-based in silico models for the prediction of P-glycoprotein substrates and inhibitors.* Bioorg Med Chem. 20:5388-95. https://doi.org/10.1016/j.bmc.2012.03.045
305. Ohashi, R., et al. (2019) *Development of Simplified in Vitro P-Glycoprotein Substrate Assay and in Silico Prediction Models To Evaluate Transport Potential of P-Glycoprotein.* Mol Pharm. 16:1851-1863. https://doi.org/10.1021/acs.molpharmaceut.8b01143
306. Ghosh, K., et al. (2020) *Identification of structural fingerprints for ABCG2 inhibition by using Monte Carlo optimization, Bayesian classification, and structural and physicochemical interpretation (SPCI) analysis.* SAR QSAR Environ Res. 31:439-455. https://doi.org/10.1080/1062936X.2020.1771769
307. Hu, B., et al. (2020) *Structure-Property Relationships and Machine Learning Models for Addressing CYP3A4-Mediated Victim Drug-Drug Interaction Risk in Drug Discovery.* Mol Pharm. 17:3600-3608. https://doi.org/10.1021/acs.molpharmaceut.0c00637
308. Mishra, N.K., S. Agarwal, and G.P. Raghava (2010) *Prediction of cytochrome P450 isoform responsible for metabolizing a drug molecule.* BMC Pharmacol. 10:8. https://doi.org/10.1186/1471-2210-10-8
309. Kriegl, J.M., et al. (2005) *A support vector machine approach to classify human cytochrome P450 3A4 inhibitors.* J Comput Aided Mol Des. 19:189-201. https://doi.org/10.1007/s10822-005-3785-3
310. Sun, H., et al. (2012) *Prediction of Cytochrome P450 Profiles of Environmental Chemicals with QSAR Models Built from Drug-like Molecules.* Mol Inform. 31:783-792. https://doi.org/10.1002/minf.201200065
311. Terfloth, L., B. Bienfait, and J. Gasteiger (2007) *Ligand-based models for the isoform specificity of cytochrome P450 3A4, 2D6, and 2C9 substrates.* J Chem Inf Model. 47:1688-701. https://doi.org/10.1021/ci700010t
312. Xie, Z., et al. (2010) *The computational model to predict accurately inhibitory activity for inhibitors towards CYP3A4.* Comput Biol Med. 40:845-52. https://doi.org/10.1016/j.compbiomed.2010.09.004
313. Pang, X., et al. (2018) *Screening of cytochrome P450 3A4 inhibitors via in silico and in vitro approaches.* RSC Adv. 8:34783-34792. https://doi.org/10.1039/c8ra06311g
314. Sorich, M.J., et al. (2003) *Comparison of linear and nonlinear classification algorithms for the prediction of drug and chemical metabolism by human UDP-glucuronosyltransferase isoforms.* J Chem Inf Comput Sci. 43:2019-24. https://doi.org/10.1021/ci034108k
315. Urtti, A., et al. (2012) *Prediction of Promiscuous P-Glycoprotein Inhibition Using a Novel Machine Learning Scheme.* PLoS ONE. 7 https://doi.org/10.1371/journal.pone.0033829

316. Eric, S., et al. (2014) *Computational classification models for predicting the interaction of drugs with P-glycoprotein and breast cancer resistance protein.* SAR QSAR Environ Res. 25:939-66. https://doi.org/10.1080/1062936X.2014.976265
317. Montanari, F., et al. (2016) *Selectivity profiling of BCRP versus P-gp inhibition: from automated collection of polypharmacology data to multi-label learning.* J Cheminform. 8:7. https://doi.org/10.1186/s13321-016-0121-y
318. Zhong, L., et al. (2011) *A prediction model of substrates and non-substrates of breast cancer resistance protein (BCRP) developed by GA-CG-SVM method.* Comput Biol Med. 41:1006-13. https://doi.org/10.1016/j.compbiomed.2011.08.009
319. Hazai, E., et al. (2013) *Predicting substrates of the human breast cancer resistance protein using a support vector machine method.* BMC Bioinformatics. 14:130. https://doi.org/10.1186/1471-2105-14-130
320. Jiang, D., et al. (2020) *ADMET evaluation in drug discovery. 20. Prediction of breast cancer resistance protein inhibition through machine learning.* J Cheminform. 12:16. https://doi.org/10.1186/s13321-020-00421-y
321. Ding, Y.L., et al. (2014) *In silico prediction of inhibition of promiscuous breast cancer resistance protein (BCRP/ABCG2).* PLoS One. 9:e90689. https://doi.org/10.1371/journal.pone.0090689
322. Belekar, V., K. Lingineni, and P. Garg (2015) *Classification of Breast Cancer Resistant Protein (BCRP) Inhibitors and Non-Inhibitors Using Machine Learning Approaches.* Comb Chem High Throughput Screen. 18:476-85. https://doi.org/10.2174/1386207318666150525094503
323. Martiny, V.Y., et al. (2013) *In silico mechanistic profiling to probe small molecule binding to sulfotransferases.* PLoS One. 8:e73587. https://doi.org/10.1371/journal.pone.0073587
324. Cook, I., et al. (2013) *High accuracy in silico sulfotransferase models.* J Biol Chem. 288:34494-501. https://doi.org/10.1074/jbc.M113.510974
325. Huang, T.W., et al. (2013) *DR-predictor: incorporating flexible docking with specialized electronic reactivity and machine learning techniques to predict CYP-mediated sites of metabolism.* J Chem Inf Model. 53:3352-66. https://doi.org/10.1021/ci4004688
326. Esposito, C., et al. (2020) *Combining Machine Learning and Molecular Dynamics to Predict P-Glycoprotein Substrates.* J Chem Inf Model. 60:4730-4749. https://doi.org/10.1021/acs.jcim.0c00525
327. Mahmud, S., et al. (2021) *Designing potent inhibitors against the multidrug resistance P-glycoprotein.* J Biomol Struct Dyn:1-13. https://doi.org/10.1080/07391102.2021.1930159
328. Demel, M.A., et al. (2009) *Predicting ligand interactions with ABC transporters in ADME.* Chem Biodivers. 6:1960-9. https://doi.org/10.1002/cbdv.200900138
329. Montanari, F. and G.F. Ecker (2015) *Prediction of drug-ABC-transporter interaction--Recent advances and future challenges.* Adv Drug Deliv Rev. 86:17-26. https://doi.org/10.1016/j.addr.2015.03.001
330. Bahar, I., R.L. Jernigan, and K.A. Dill, *Protein Actions : Principles & Modeling*. 2017, Garland Science, New York, NY, ISBN 978-0-8153-4177-2
331. Braxenthaler, M., et al. (1997) *Chaos in protein dynamics.* Proteins. 29:417-25.
332. Yadava, U. (2018) *Search algorithms and scoring methods in protein-ligand docking.* Endocrinology&Metabolism International Journal. 6 https://doi.org/10.15406/emij.2018.06.00212
333. Moustakas, D.T., et al. (2006) *Development and validation of a modular, extensible docking program: DOCK 5.* J Comput Aided Mol Des. 20:601-19. https://doi.org/10.1007/s10822-006-9060-4
334. Rarey, M., et al. (1996) *A fast flexible docking method using an incremental construction algorithm.* J Mol Biol. 261:470-89. https://doi.org/10.1006/jmbi.1996.0477
335. Ferreira, L.G., et al. (2015) *Molecular docking and structure-based drug design strategies.* Molecules. 20:13384-421. https://doi.org/10.3390/molecules200713384
336. Prieto-Martínez, F.D., M. Arciniega, and J.L. Medina-Franco (2018) *Acoplamiento Molecular: Avances Recientes y Retos.* TIP Revista Especializada en Ciencias Químico-Biológicas. 21 https://doi.org/10.22201/fesz.23958723e.2018.0.143
337. Liu, M. and S. Wang (1999) *MCDOCK: a Monte Carlo simulation approach to the molecular docking problem.* J Comput Aided Mol Des. 13:435-51. https://doi.org/10.1023/a:1008005918983
338. Trott, O. and A.J. Olson (2010) *AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading.* J Comput Chem. 31:455-61. https://doi.org/10.1002/jcc.21334
339. Totrov, M. and R. Abagyan (1997) *Flexible protein-ligand docking by global energy optimization in internal coordinates.* Proteins. Suppl 1:215-20. https://doi.org/10.1002/(sici)1097-0134(1997)1+<215::aid-prot29>3.3.co;2-i
340. Morris, G.M., et al. (2009) *AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility.* J Comput Chem. 30:2785-91. https://doi.org/10.1002/jcc.21256

341. Jones, G., et al. (1997) *Development and validation of a genetic algorithm for flexible docking.* J Mol Biol. 267:727-48. https://doi.org/10.1006/jmbi.1996.0897
342. Gohlke, H., M. Hendlich, and G. Klebe (2000) *Knowledge-based scoring function to predict protein-ligand interactions.* J Mol Biol. 295:337-56. https://doi.org/10.1006/jmbi.1999.3371
343. Empereur-Mot, C., et al. (2015) *Predictiveness curves in virtual screening.* J Cheminform. 7:52. https://doi.org/10.1186/s13321-015-0100-8

VI. Appendix

„Az utolsó csepp nem tehet róla, hogy éppen ő az utolsó csepp,
hiszen amúgy semmiben sem különbözik az előző cseppektől.”

“The last drop cannot be blamed for being the last,
it differs not from the rest.”

Mérő László

A. Publications encompassed within the thesis

Dudas B[#], Bagdad Y[#], Picard M, Perahia D, Miteva MA: *Machine learning and structure-based modeling for the prediction of UDP-glucuronosyltransferase inhibition*, iScience 2022 (Accepted for publication)

Dudas B, Decleves X, Cisternino S, Perahia D, Miteva MA: *ABCG2/BCRP transport mechanism revealed through kinetically excited targeted molecular dynamics simulations,* Comput Struct Biotechnol J. 2022, doi: 10.1016/j.csbj.2022.07.035

Dudas B[#], Toth D[#], Perahia D, Nicot AB, Balog E, Miteva MA. *Insights into the substrate binding mechanism of SULT1A1 through molecular dynamics with excited normal modes simulations.* Sci Rep. 2021, doi: 10.1038/s41598-021-92480-w

B. Other publications

Kaynak BT[#], Krieger JM[#], **Dudas B**[#], Dahmani ZL, Costa MGS, Balog E, Scott AL, Doruker P, Perahia D, Bahar I. *Sampling of Protein Conformational Space Using Hybrid Simulations: A Critical Assessment of Recent Methods.* Front Mol Biosci. 2022, doi: 10.3389/fmolb.2022.832847

Dudas B, Perahia D, Balog E. *Revealing the activation mechanism of autoinhibited RalF by integrated simulation and experimental approaches.* Sci Rep. 2021, doi: 10.1038/s41598-021-89169-5

Dudas B, Merzel F, Jang H, Nussinov R, Perahia D, Balog E. *Nucleotide-Specific Autoinhibition of Full-Length K-Ras4B Identified by Extensive Conformational Sampling.* Front Mol Biosci. 2020, doi: 10.3389/fmolb.2020.00145

[#]1st co-authors

C. Intellectual Property

DrugME software: No: 001.010030.000.S.C.2021.000.31230 (5 %)
For the prediction of phase II drug metabolism inhibition

D. Prizes & Awards

Young Researcher Grand First Prize	Hungarian Biophysical Society (2021)
EBSA Travel Award	European Biophysical Society (2021)
Initiative d'Excellence (IdEx) Doctoral Fellowship	Université Paris Cité (2019)

E. Teaching Assignments

Physical Chemistry (Université Paris Cité), 2020-2022

Molecular modelling and simulation: theoretical principles and applications in biology (Ecole Normale Supérieure Paris-Saclay), 2020-2022

F. Oral presentations

Gordon Research Seminar & Conference, Castelldefels, 2022
ABCG2 transport mechanism revealed through kinetically excited targeted molecular dynamics simulations

Atelier de Modélisation des Molécules d'Intérêt Biologique (AMMIB) Conference, Palaiseau, 2022
Insights into the transport mechanism of BCRP through an enhanced MD simulation tool, kinetically excited targeted Molecular Dynamics

Group of Graphism and Molecular Modeling (GGMM) and French Society of Chemoinformatics (SFCi) Conference, Lille, 2021
Insights into the transport mechanism of BCRP through Molecular Dynamics with excited Normal Modes simulations

Scientific Days of Médicament, Toxicologie, Chimie, Imageries Doctoral School, Le Mée-sur-Seine, 2021
Extended structural dynamics reveal SULT1A1 accommodation to large ligands

Meeting of the Hungarian Biophysical Society, Budapest, 2021
In silico prediction of drug transport inhibition

G. Conference Posters

Dudas B, Decleves X, Cisternino S, Perahia D, Miteva MA: ABCG2 transport mechanism revealed through kinetically excited targeted molecular dynamics simulations, Gordon Research Seminar & Conference, Castelldefels, 2022

Dudas B, Decleves X, Cisternino S, Perahia D, Miteva MA: The transport mechanism of ABCG2 investigated by an enhanced MD simulation tool, kinetically excited targeted Molecular Dynamics, Chemical Computing Group (CCG) Conference, Amsterdam, 2022

Dudas B, D. Tóth, Perahia D, Nicot A, Balog E, Miteva MA: Extended structural dynamics reveal SULT1A1 accommodation to large ligands, European Biophysics Journal V.50, Vienna, 2021

Dudas B, D. Tóth, Perahia D, Nicot A, Balog E, Miteva MA: Mechanistic insights into the effect of structural dynamics of SULT1A1 on substrate binding and selectivity, 28th Young Research Fellows Meeting, Paris, 2021

Dudas B, D. Tóth, Nicot A, Perahia D, Balog E, Miteva MA: Mechanistic insights into the effect of structural dynamics of SULT1A1 on substrate binding and selectivity, OpenTox Congress, online, 2020

H. Substantial summary in English

Title: *Molecular mechanisms of phase II metabolizing enzymes and ABC transporters, and their interactions with small molecules modeled through structure-based and machine learning methods*

Key words: *drug metabolizing enzymes, SULT, UGT, ABC transporters, BCRP, ABCG2, efflux mechanism, conformational exploration, molecular dynamics, normal modes, docking, machine learning*

1. Introduction

Drug discovery and development are expensive and slow processes. A major challenge associated with the identification of promising drug candidates is to find a good balance between the required efficacy, selectivity, and affinity against their intended therapeutic target while at the same time showing appropriate absorption, distribution, metabolism, excretion, and low toxicity (ADME-Tox) properties. The complex process of drug elimination is governed by drug metabolizing enzymes (DMEs) and drug transporters. Xenobiotics and endogenous compounds that should be eliminated from the human body can undergo phase I and/or phase II metabolism and then be excreted by efflux transporters, DMEs and efflux transporters modulate the intracellular bioavailability and pharmacokinetics of drugs and other xenobiotics. Furthermore, the inhibition of DMEs and drug transporters, possibly involved in drug-drug interactions, can directly increase intracellular toxicity while the formation of reactive or toxic metabolites are also a safety liability.

1.1. Phase I drug metabolism

Phase I metabolism (functionalization) primarily includes oxidation. Phase I DMEs aim to unmask a polar functional group on their substrates. In case the phase I metabolite is hydrophilic enough it may get directly eliminated. Otherwise, a consecutive conjugation reaction step can be catalyzed by phase II DMEs. The most important phase I DMEs belong to the cytochrome P450 superfamily that is capable of catalyzing the oxidative biotransformation of most drugs and other lipophilic xenobiotics.

1.2. Phase II drug metabolism

Phase II metabolism (conjugation) links a relatively large endogenous polar group to diverse types of compounds, generally creating water-soluble products with increased molecular weight which can be excreted in bile or urine. Sulfation and glucuronidation are two major phase II reactions. The resulting metabolites are typically of reduced membrane permeability, consequently their excretion is coupled to active drug transporters. In most cases conjugation reactions terminate the biological activity of drugs. However, reactive conjugated metabolites have also been reported. Phase II metabolism can follow the unmasking of a polar functional group by phase I DMEs, however, numerous compounds can be directly conjugated without a preceding phase I reaction.

Xenobiotic sulfation reactions are catalyzed by the cytosolic sulfotransferase enzymes (SULTs) that transfer the sulfonate group from the cofactor 3′-Phosphoadenosine 5′-Phosphosulfate (PAPS) to a hydroxyl or amino group of their substrates. The major enzyme responsible for xenobiotic sulfonation is the widely expressed SULT1A1.

The primary sugar conjugation route in humans is glucuronidation which is catalyzed by uridine 5′-diphosphate-glucuronosyltransferases (UGTs) that transfer the glucuronic acid from the cofactor uridine-diphosphate glucuronic acid to nucleophilic substrates. The isoenzyme UGT1A1 is of particular importance, given its exclusive role in the glucuronidation, and therefore the detoxification of the endogenous heme breakdown byproduct bilirubin together with the glucuronidation of a number of xenobiotics and drugs of clinical interest.

1.3. Drug transporters

Transporters present embedded in the plasma membrane of cells mediate the uptake of endobiotic and xenobiotic molecules from the extracellular space and the elimination of toxic waste from the cytosol. Drug transporters have been identified to influence drug disposition and be involved in drug-drug interactions (DDI) of a large number of drugs and drug candidates as well as to contribute to the multidrug-resistance phenotype of tumor cells. Mechanistically, most drug transporters can be classified as either solute carrier (SLC) transporter or ATP-binding cassette (ABC) transporter. Most SLC transporters are influx transporters, among their substrates there are sugars, amino acids, vitamins, nucleotides, metals, inorganic ions, organic anions, oligopeptides, and drugs. SLC transporters are either passive facilitative transporters or secondary active transporters. As opposed to that, ABC transporters bind ATP and harvest the energy of ATP hydrolysis in order to selectively translocate a variety of substrates across membranes. The ABC transporter superfamily represents the largest family of transmembrane proteins and in eukaryotes they primarily function as exporters. ABCB1, ABCG2, and members of the ABCC subfamily have been proven to be involved in the development of multidrug resistance and drug-drug interactions. In particular, ABCG2 (BCRP) is a key player in preventing the absorption of toxic compounds from the gut, and it also plays an essential protective role at different tissue barriers like the maternal-fetus barrier, the blood-brain barrier (BBB), and the blood-testis barrier. ABCG2 transports a wide variety of drugs and also many phase II metabolites such as sulfate or glucuronide conjugates. The exact mechanisms of the ABC transporter-mediated substrate translocation are not fully understood.

2. Methods

Computational modeling of proteins and their interactions with ligands is of increasing importance in drug discovery and development. Both ligand-based methods that are based on already known ligands, as well as structure-based methods in case the structure of the target protein (or a similar one) is known, can be used to identify substrates or inhibitors. *In silico* modeling of the structures and dynamics of proteins requires a model of interatomic interactions. Different force fields are used to define the potential energy as a function of atomic coordinates. Most commonly used semi-empirical forcefields include bonded terms for bond stretching, bond bending, and bond torsion, and nonbonded terms for van der Waals and

electrostatic interactions. Energy minimization aims to reach conformations of low potential energy by identifying the local minimum closest to the starting structure.

However, to provide a proper description of the conformational population, free energies need to be estimated so that a combination of low energy and high entropy can be ensured. Molecular dynamics simulations solve Newton's equations of motion iteratively thereby creating a time-dependent conformational trajectory. As opposed to energy minimization, molecular dynamics is capable of crossing energy barriers and reaching states with lower free energy. Several simulation tools aim to overcome the time limitations of classical all-atomic molecular dynamics, such as metadynamics, conformational flooding, accelerated MD, or targeted MD. Another way is to use coarse-grained models rather than including full atomistic details.

Normal mode analysis can be efficiently used for identifying and describing the slowest intrinsic motions of macromolecules, which in nature, generally correspond well to collective functional movements. Normal mode analysis relies on the harmonic approximation of the potential energy function around a given local minimum.

In the modeling of protein-ligand interactions, molecular docking simulations can be efficiently used to identify binding sites and poses for ligands, and predict binding affinities. Docking simulations approximate true physical energies with simplified energetics and solvation, they generate and rank protein-ligand configurations by using different search algorithms and scoring functions.

Given a training data set of compounds with known activity, ligand- and structure-based features can be used to train machine learning prediction models for classification (or regression). Random forest and support-vector machine are widely used examples of supervised machine learning algorithms, they can be efficiently used to distinguish between active and inactive compounds of a given target protein.

3. Objectives

One of the major reasons for drug candidate failure is due to problems related to pharmacokinetics/pharmacodynamics during clinical trials. *In silico* prediction of interactions with DMEs and drug transporters that govern the pharmacokinetics of xenobiotics can help reduce the rate of drug candidate failure at an early drug development stage, also reducing associated costs and can help decrease the number of animal tests. In my PhD work I have focused on two phase II DMEs, SULT1A1 and UGT1A1, as well as on an essential drug efflux transporter, ABCG2.

The objective of my PhD work has been to decipher the functional movements encoded in the above listed proteins with the help of different simulation approaches, to identify the effects of ligand binding on their dynamics and functional collective movements, and ultimately to create *in silico* predictive machine learning models capable of distinguishing between inhibitors and inactive compounds with respect to the target proteins.

4. Results

4.1. Substrate binding mechanism of SULT1A1

It has been previously suggested that a considerable shift of SULT structure caused by PAPS binding could control the capability of SULT to bind large substrates. In order to elucidate molecular mechanisms guiding the recognition of diverse substrates and inhibitors by SULT1A1, molecular dynamics (MD) simulations and the recently developed approach of MD with excited normal modes (MDeNM) were performed in the presence of the cofactor (PAPS). MDeNM simulations enable an extended sampling of the conformational space by running multiple short MD simulations during which motions described by a subset of low-frequency Normal Modes are kinetically excited. Root mean square deviation analysis calculated on the binding pocket and the whole protein revealed that MDeNM performed a more exhaustive conformational sampling of the SULT1A1 binding pocket than the classical MD simulations while maintaining the protein's overall structure closer to the initial crystal structure. Root mean square fluctuation analysis of the Cα atoms revealed that MDeNM particularly magnified motions related to the loops L1 and L3, and moderately related to L2, the three loops forming the gate to the binding pocket. Fluctuations at the tips of the loops L1 and L3 are double in the case of MDeNM compared to classical MD simulations, indicating that MDeNM explores the gating motions to a greater extent. L1 exhibited a larger fluctuation than L3 by both MD and MDeNM, implying its involvement in the gating mechanism as was earlier proposed. Analysis of the gate opening revealed that MDeNM reached considerably more widely open conformations whereas MD mapped densely populated tightly closed states.

To gain insight into the mechanism of SULT1A1-ligand interactions, the docking of 132 known substrates or inhibitors was performed into the binding pocket of the centroid conformations collected by MD and MDeNM after clustering. Many ligands expressed similar docking behavior into the MD and MDeNM set of conformations in terms of interaction energy (IE), for some ligands, however, considerable differences were observed. Most of these compounds showed a more favorable IE when docked to the MDeNM set of conformations, demonstrating the benefit of including the MDeNM simulations in addition to MD. In particular, the assessment of ligands for which there was a significant difference between MD and MDeNM revealed that most of the compounds for which MDeNM performed better were of big size, occupying a large volume in the binding pocket, and their poses corresponding to the best IE were accommodated within widely open enzyme conformations that were poorly populated or even not accessible by classical MD simulations.

Two substrates of different sizes were used in additional docking simulations to further investigate the gating mechanism and substrate recognition of SULT1A1. The substrate 17β-estradiol (E2) is a smaller, medium-sized substrate while fulvestrant, an estrogen analogue, is a larger substrate of SULT1A1, with an additional 15-atom long functional sidechain, both substrates were docked into 6000 conformations reached by MD and 6000 other conformations generated by MDeNM. MD and MDeNM conformations were capable of accommodating E2, regardless of their openness which agrees with previous kinetic and

binding studies showing that E2 can bind to open and closed conformations of the enzyme. Fulvestrant showed an obvious preference toward open conformations. The most favorable docking positions of E2 and fulvestrant were further analyzed and were found to be stable by consecutive classical MD simulations starting from the enzyme-cofactor-substrate complexes. MD simulations in the presence of estradiol and fulvestrant showed the induced further opening of the loops in the presence of the bound substrate. Both MD and MDeNM results confirmed that open conformations are available for big ligands (such as fulvestrant, 4-hydroxytamoxifen, or raloxifene) to bind even in the presence of the bound cofactor. Thus, our results proposed a new mechanism for large substrate recognition without implicating a movement of the cofactor to trigger large-substrate binding.

4.2. Prediction of UDP-glucuronosyltransferase inhibition

Strong inhibition of UGT1A1 may trigger adverse drug/herb-drug interactions, or result in disorders of the endobiotic metabolism. In order to address the degree of structural flexibility and the conformational adaptation of its binding pocket in light of the structural variety of the diverse active compounds, molecular dynamics simulations were performed starting from a human homology model in the presence of the cofactor (UDPGA). Root mean square deviation analysis demonstrated that all three MD runs quickly diverted from the initial conformation reaching relatively high differences. Larger interdomain opening-closing motions were revealed by the radius of gyration analysis and visual inspection of the trajectories. The substrate-binding pocket volume itself was found to significantly vary during the simulations, in some conformations the volume reached 1.5 to 2 times the size of the starting structure. The large variations in the substrate-binding pocket volume and the opening towards the lumen can facilitate access to the catalytic site and accommodate the diverse substrates.

Known ligands of UGT1A1, substrates and inhibitors, were collected to be used for docking and to build training and external test datasets for machine learning modeling. Ensemble docking of the active and decoy molecules of the training set was performed on the centroid conformations collected by MD after clustering to identify protein conformations best distinguishing between active and inactive compounds. Using the best retained IE for each compound, enrichment curves were calculated for each protein centroid protein conformation, six of which were kept for the machine learning modeling and the corresponding IE docking scores were used as protein-ligand interactions-based descriptors. The identified protein conformations exhibited large flexibility at their substrate-binding pockets, key residues for the catalytic reaction remained less flexible, and kept their orientation within the binding pocket.

Physicochemical molecular descriptors of the active compounds and decoys were calculated, and highly correlated descriptors together with descriptors of near null variance were removed, and the six IE scores retrieved from the docking simulations were included as structure-based descriptors accounting for the protein-ligand interactions. A reduction in the number of descriptors was performed based on their relative importance in order to avoid overfitting after training a number of random forest models on the training dataset and selecting the subset of descriptors with the highest Gini importance. A total of 56 ligand-based

descriptors and the 6 IE were kept for further model building, the most important descriptors were related to polarity, lipophilicity, and charges. Random forest (RF) and support vector machine (SVM) algorithms were used to create optimized prediction models integrating ligand- and structure-based information, optimizing the corresponding model hyper parameters. The RF and SVM models showed excellent predictive powers for the discrimination of the UGT1A1 active molecules with an accuracy and sensitivity of around 90 %. These are the first machine learning models created for the prediction of UGT1A1 inhibition, they can be helpful for the prediction of drug-drug interactions of novel drug candidates while also providing structural information on enzyme-ligand interactions.

4.3. ABCG2 transport mechanism

ABCG2 is of major interest due to its involvement in multidrug resistance (MDR), and understanding its complex efflux mechanism is essential to preventing MDR and drug-drug interactions (DDI). ABCG2 export is characterized by two major conformational transitions between inward- and outward-facing states (IFS and OFS), the structures of which have been resolved. Yet, the entire transport cycle has not been characterized to date. In order to elucidate the transport mechanism of ABCG2 and its behavior in the different transport stages, different simulations were performed starting from available cryo-EM structures in its IFS and OFS. In the IFS, the two nucleotide-binding domain (NBD) monomers are partially separated while in the OFS they form a tightly-packed dimer establishing the two nucleotide-binding sites. The transmembrane domain (TMD) pair forms the slit-like hydrophobic cavity 1, where physiological substrates and various inhibitors have been proven to bind, and cavity 2 which faces the extracellular space. Cavity 1 is collapsed in the OFS, the two cavities are separated by a leucine gate.

First, the missing intracellular loop regions within the NBDs were modeled since they are likely to affect the substrate entry and may possess a gating function. Similarly, the linker segment was also modeled connecting the individual NBDs and TMDs. Multiple systems were constructed from the experimental IFS and OFS structures in the presence or absence of a physiological substrate (estrone 3-sulfate, E_1S) and nucleotides (ATPs or ADPs) to mimic the different transport stages:

- Apo IFS
- E_1S-bound IFS
- E_1S- and ATP-Mg^{2+}-bound IFS
- ATP-Mg^{2+}-bound OFS
- ADP-bound OFS
- Nucleotide-free OFS

Molecular dynamics (MD) simulations embedded in a lipid bilayer and normal mode analysis (NMA) were performed on the above six systems. The timescale of a complete transport cycle of ABCG2 falls in a range of a fraction of seconds or beyond, a timeframe that currently cannot be simulated by classical MD. Therefore, I developed an enhanced MD simulation tool,

kinetically excited targeted MD (ketMD) that traces possible pathways between two terminal structures, and I applied it to ABCG2 to simulate conformational transitions between its IFS and OFS. Similarly to targeted MD (tMD), the excitation vector was chosen to point towards the target structure, however, unlike tMD, where the potential energy function is biased and the protein is guided by steering forces at each simulation step, ketMD relies on kinetic excitations. At the first step of each excitation cycle, the velocity components pointing from the instantaneous conformation to the target structure are increased, allowing the crossing of larger energy barriers. This excitation step is followed by a relaxation period where no external perturbation is applied, the system can evolve, and the injected kinetic energy dissipates. After each excitation cycle, the excitation direction vector is updated to point from the instantaneous conformation to the target structure.

Transition 1 (IFS to OFS) was simulated starting from the E_1S- and ATP-Mg^{2+}-bound IFS, whereas transition 2 (OFS to IFS) from the OFS without bound substrates and in the presence of ATP-Mg^{2+}, ADP, or in the absence of bound nucleotides. Root mean square deviation analysis, the monitoring of a catalytic distance-pair, and visual inspection confirmed that a full NBD dimerization was successfully achieved together with the catalytic ATP-binding site formation during the ketMD simulation of transition 1, whereas in the opposite direction, the NBDs got partially separated and a complete NBDs transition occurred during the ketMD simulations. During transition 2, detaching the NBDs at the ATP-binding site was more easily achieved in the absence of nucleotides, and most difficult in the presence of ATP.

Radius of gyration analysis of the helical segments bordering the two cavities was performed to investigate their state during the simulations. By the end of the ketMD simulation of transition 1, cavity 1 was collapsed (identically to the OFS structure) while in the opposite direction, starting from a collapsed cavity 1, it became exposed again and accessible from the cytosol. Opposed to this, during the ketMD simulation of transition 1, cavity 2 became more exposed to the extracellular space and also more voluminous, while during the simulation of transition 2, starting from the OFS, cavity 2 approached a more deflated state.

Experimental structures with a substrate bound to cavity 2, or transient structures along the translocation pathway have not been resolved. With the ketMD run starting from the IFS, it was possible to simulate the translocation of E_1S from cavity 1 to the extracellular space through the leucine gate and cavity 2. In addition to the excitation applied to the transporter, the substrate motion was also kinetically promoted towards the extracellular space. The substrate was initially bound in cavity 1 stabilized mainly by the ‘sandwich-like’ stacking interactions of F439 and F439′. As soon as the substrate escaped from the ’sandwich-like‘ trap of the two F439 residues and moved towards cavity 2, F439 and F439′ came into close contact, creating a valve-like construction, similar to what is observed in the OFS cryo-EM structure. Any kind of return movement towards the cytosol was prevented with this valve closed and I identified a pocket-like formation between cavities 1 and 2, where the substrate was trapped as movements to cavity 2 still remained blocked by the closed leucine gate. Several residues at the edge of this site stabilized the position of the substrate, it only moved away from the pocket

once the residues of the leucine gate were separated. The substrate behavior in cavity 2 was very different from what was observed either in cavity 1 or between cavities 1 and 2. Before arriving at cavity 2, the substrate was tightly bound and closely surrounded by transporter residues. In contrast, the substrate was loosely bound here as it explored the cavity volume, making close contacts with residues at its boundary. The substrate's further kinetic excitation resulted in its complete detachment from the transporter into the extracellular space.

The classical MD simulations and normal mode analysis of the different transport stages revealed that the presence of the substrate in cavity 1 is essential to couple the movements between the NBDs and the TMDs. In the absence of a bound substrate and nucleotides (apo-form), the TMDs approached a neutral configuration, cavity 2 opened while cavity 1 started collapsing, while the NBDs stayed far apart. Contrary to cavity 1, cavity 2 was never fully collapsed, at any stages of the transport cycle (during the ketMD and classical MD simulations) even though experiments suggest that it can get occluded. Its volume shows inflating-deflating variations between the IFS and OFS states, however, cavity 2 is more voluminous than cavity 1 even in its deflated state. The ketMD and subsequent classical MD simulations demonstrated that E_1S, which is a bulky compound, could be present in cavity 2, even in the OFS and they also showed a sufficiently large space for the substrate in cavity 2 throughout the entire transport cycle. This may allow simultaneous substrate binding in cavities 1 and 2, resulting in an accelerated export mechanism.

The harmonic approximation of normal modes revealed that transition-like motions are present in the IFS corresponding to low frequencies, whereas they are damped and are local motions of higher frequencies in the OFS systems. This suggests that it is energetically costly to start transition 2 to return to the initial IFS, which may require the energy released upon ATP hydrolysis.

5. Conclusions

The prediction of the fate of administrated drugs in the human body remains a challenging task that has long relied on expensive animal studies also raising concerns about ethical aspects and reliability. Therefore, an increasing number of computational approaches have been involved in order to efficiently predict the metabolic outcome of drugs and new drug candidates. Several studies have created prediction models using ligand-based or structure-based information, and in 2013, the host laboratory trained the first machine learning classification models that integrated ligand- and structure-based information for the prediction of the inhibition of different SULT isoforms, and later for CYP enzymes. Understanding the collective functional movements of the proteins and their interactions with small ligands is a crucial step in extracting structure-based information for such prediction models.

In my PhD work, I have focused on two essential phase II DMEs, SULT1A1 and UGT1A1, as well as on the drug efflux transporter ABCG2. My research addressed both the dynamics of the different proteins and their interactions with small molecules. This also included methodological developments, such as an enhanced molecular dynamics simulation tool and

different prediction tools through the integration of structure-based and machine learning modeling.

Previously it was suggested that PAPS binding restricts the access for larger substrates to SULT1A1. Classical MD simulations and MD in combination with NMs (MDeNM) were performed in order to elucidate the molecular mechanisms guiding the recognition of diverse substrates and inhibitors by SULT1A1 in the presence of the cofactor, which was followed by docking simulations of known ligands. The results of the docking and subsequent classical molecular dynamics simulations on the complexes demonstrated that large conformational changes of SULT1A1 could occur even in the presence of the cofactor, and that more widely open conformations can accommodate a large substrate like fulvestrant with higher affinity. In addition to L3, the L1 loop region at the entrance to the catalytic site exhibited an extremely high flexibility which underlined their functional importance in the gating mechanism of SULT1A1. Altogether, my work shed new light on the complex mechanisms of substrate specificity and inhibition of SULT1A1, and has highlighted the utility of including MDeNM in protein-ligand interactions studies where major rearrangements are expected. The generated conformations can be efficiently used to train *in silico* prediction models with even higher accuracy for larger molecules.

Structure-based modeling and ligand-based information were integrated to build the first machine learning prediction models of UGT1A1 inhibition. In my PhD work, I performed molecular dynamics simulations of the enzyme starting from a human homology model in the presence of the cofactor to explore the structural variability of the catalytic site, and observed significant flexibility, important for the accommodation of the diverse ligands. This was followed by conformational clustering. A set of known ligands, substrates and inhibitors, were collected and ensemble docking was performed to identify conformations that can efficiently differentiate between actives and decoys. Ligand-based descriptors were calculated. Finally, a reasonable selection of ligand-based descriptors together with the interaction energies retrieved from the docking were used to train the first machine learning models for the prediction of UGT1A1 inhibition. The trained models showed excellent performance and were implemented in the DrugME software. DrugME can be efficiently used to identify new UGT1A1 inhibitors and can be helpful for the prediction of drug-drug interactions of new drug candidates while they also provide structural information on enzyme-ligand interactions.

I have developed an innovative enhanced MD simulation technique, termed ketMD (kinetically excited targeted Molecular Dynamics) to simulate the ABCG2 transport cycle together with the translocation of the physiological substrate, estrone 3-sulfate. With the help of ketMD, I successfully simulated the conformational transitions between the two extreme states of ABCG2 and revealed the complex molecular mechanism of the E_1S substrate translocation. I observed a valve-like function of two phenylalanine residues (F439 and F439') in cavity 1 that initially engage in stacking interactions against the substrate. Classical MD simulations and Normal Mode Analysis revealed that the presence of the substrate in cavity 1 is necessary to couple the movements of the transmembrane domain to the nucleotide-binding

domain. I also observed that cavity 2 was never completely collapsed, at any stages of the transport cycle which might enable simultaneous substrate binding in cavities 1 and 2 and accelerate the export rate of the transporter. The work on ABCG2 shed new light on the complex molecular mechanism of its transport and highlighted the utility of including enhanced *in silico* sampling techniques, such as ketMD, in transporter studies. In the near future, the generated conformations along the transition pathway will be used to develop new machine learning models for the prediction of ABCG2 inhibitors and substrates in order to probe new drug candidates for multi-drug resistance and predict drug-drug interactions.

During the three years of my PhD work I had the great chance to deepen my knowledge in the biology of drug metabolism, elimination, and drug-drug interactions and also in the application and development of new *in silico* approaches to study the enzymes and transporters that govern these complex processes. The observations and results of the work presented in the thesis contribute to a better understanding of the molecular mechanisms of phase II metabolizing enzymes and ABC transporters, and their interactions with ligands. In particular, the DrugME software that now incorporates the developed predictive models integrating structure-based and machine learning modeling will greatly contribute to the future predictions of drug-drug interactions.

I. Résumé substantiel en français

Titre : *Mécanismes moléculaires des enzymes de métabolisme de phase II et des transporteurs ABC, et leurs interactions avec de petites molécules modélisées par des méthodes structurales et d'apprentissage automatique*

Mots clés : *enzymes métabolisant des médicaments, SULT, UGT, transporteurs ABC, BCRP, ABCG2, mécanisme d'efflux, exploration conformationnelle, dynamique moléculaire, modes normaux, arrimage moléculaire, apprentissage automatique*

1. Introduction

La découverte et le développement de médicaments sont des processus coûteux et lents. Un défi majeur associé à l'identification de candidats médicaments prometteurs est de trouver un bon équilibre entre l'efficacité, la sélectivité et l'affinité requises contre la cible thérapeutique visée, tout en présentant des propriétés appropriées d'absorption, de distribution, de métabolisme, d'excrétion et de toxicité (ADME-Tox). Le processus complexe d'élimination des médicaments est régi par les enzymes de métabolisation des médicaments (DME) et les transporteurs de médicaments. Les xénobiotiques et les composés endogènes qui doivent être éliminés du corps humain peuvent subir un métabolisme de phase I et/ou de phase II, puis être excrétés par des transporteurs d'efflux. Les DME et les transporteurs d'efflux modulent la biodisponibilité intracellulaire et la pharmacocinétique des médicaments et autres xénobiotiques. En outre, l'inhibition des DME et des transporteurs de médicaments, qui peuvent être impliqués dans des interactions médicamenteuses, peut augmenter directement la toxicité intracellulaire, tandis que la formation de métabolites réactifs ou toxiques constitue également un risque pour la sécurité.

1.1. Métabolisme de phase I des médicaments

Le métabolisme de phase I (fonctionnalisation) comprend principalement l'oxydation. Les DME de phase I visent à démasquer un groupe fonctionnel polaire sur leurs substrats. Si le métabolite de phase I est suffisamment hydrophile, il peut être directement éliminé. Sinon, une étape consécutive de réaction de conjugaison peut être catalysée par les DME de phase II. Les DME de phase I les plus importants appartiennent à la superfamille des cytochromes P450 qui sont capables de catalyser la biotransformation oxydative de la plupart des médicaments et autres xénobiotiques lipophiles.

1.2. Métabolisme de phase II des médicaments

Le métabolisme de phase II (conjugaison) lie un groupe polaire endogène relativement grand à divers types de composés, créant généralement des produits hydrosolubles de poids moléculaire accru qui peuvent être excrétés dans la bile ou l'urine. La sulfonation et la glucuronidation sont deux réactions majeures de la phase II. Les métabolites obtenus ont généralement une perméabilité membranaire réduite, et leur excrétion est donc couplée à des transporteurs de médicaments actifs. Dans la plupart des cas, les réactions de conjugaison mettent fin à l'activité biologique des médicaments. Cependant, des métabolites conjugués réactifs ont également été signalés. Le métabolisme de phase II peut suivre le démasquage d'un

groupe fonctionnel polaire par les DME de phase I, cependant, de nombreux composés peuvent être directement conjugués sans réaction de phase I préalable.

Les réactions de sulfatation des xénobiotiques sont catalysées par les enzymes sulfotransférases cytosoliques (SULT) qui transfèrent le groupe sulfonate du cofacteur 3′-Phosphoadénosine 5′-Phosphosulfate (PAPS) à un groupe hydroxyle ou amino de leurs substrats. La principale enzyme responsable de la sulfonation des xénobiotiques est la SULT1A1, largement exprimée dans le corps humain.

La principale voie de conjugaison des sucres chez l'homme est la glucuronidation, catalysée par les uridine 5' diphosphate glucuronosyltransférases (UGT) qui transfèrent l'acide glucuronique du cofacteur acide uridine-diphosphate glucuronique aux substrats nucléophiles. L'isoenzyme UGT1A1 est particulièrement importante, étant donné son rôle exclusif dans la glucuronidation, et donc la détoxification de la bilirubine, sous-produit endogène de la dégradation de l'hème, ainsi que dans la glucuronidation d'un certain nombre de xénobiotiques et de médicaments d'intérêt clinique.

1.3. Transporteurs de médicaments

Les transporteurs présents dans la membrane plasmique des cellules médient l'absorption des molécules endobiotiques et xénobiotiques de l'espace extracellulaire et l'élimination des produits toxiques du cytosol. Les transporteurs de médicaments ont été identifiés comme influençant la disposition des médicaments et étant impliqués dans les interactions médicamenteuses (DDI) d'un grand nombre de médicaments et de candidats médicaments, ainsi que comme contribuant au phénotype de multirésistance des cellules tumorales. D'un point de vue mécanique, la plupart des transporteurs de médicaments peuvent être classés comme transporteurs de soluté (SLC) ou transporteurs ABC (ATP-binding cassette). La plupart des transporteurs SLC sont des transporteurs d'influx, parmi leurs substrats on trouve des sucres, des acides aminés, des vitamines, des nucléotides, des métaux, des ions inorganiques, des anions organiques, des oligopeptides et des médicaments. Les transporteurs SLC sont soit des transporteurs passifs facilitateurs, soit des transporteurs actifs secondaires. En revanche, les transporteurs ABC se lient à l'ATP et récupèrent l'énergie de l'hydrolyse de l'ATP afin de transloquer sélectivement divers substrats à travers les membranes. La superfamille des transporteurs ABC représente la plus grande famille de protéines transmembranaires et, chez les eucaryotes, ils fonctionnent principalement comme exportateurs. Il a été prouvé que ABCB1, ABCG2 et les membres de la sous-famille ABCC sont impliqués dans le développement de la multirésistance et des interactions médicamenteuses. En particulier, ABCG2 (BCRP) est un acteur clé dans la prévention de l'absorption de composés toxiques à partir de l'intestin, et il joue également un rôle protecteur essentiel au niveau de différentes barrières tissulaires comme la barrière mère-fœtus, la barrière hémato-encéphalique (BHE) et la barrière sang-testicule. ABCG2 transporte une grande variété de médicaments ainsi que de nombreux métabolites de phase II tels que les sulfates ou les glucurones.

2. Méthodes

La modélisation informatique des protéines et de leurs interactions avec les ligands revêt une importance croissante dans la découverte et le développement de médicaments. Les méthodes basées sur des ligands déjà connus, ainsi que les méthodes basées sur la structure dans le cas où la structure de la protéine cible (ou une protéine similaire) est connue, peuvent être utilisées pour identifier des substrats ou des inhibiteurs. La modélisation in silico des structures et de la dynamique des protéines nécessite un modèle d'interactions interatomiques. Différents champs de force sont utilisés pour définir l'énergie potentielle en fonction des coordonnées atomiques. Les champs de force semi-empiriques les plus couramment utilisés comprennent des termes liés pour l'étirement, la flexion et la torsion des liaisons, et des termes non liés pour les interactions de van der Waals et électrostatiques. La minimisation de l'énergie vise à atteindre des conformations à faible énergie potentielle en identifiant le minimum local le plus proche de la structure de départ.

Cependant, pour fournir une description correcte de la population conformationnelle, les énergies libres doivent être estimées de manière à garantir une combinaison d'énergie faible et d'entropie élevée. Les simulations de dynamique moléculaire (MD) résolvent les équations du mouvement de Newton de manière itérative, créant ainsi une trajectoire conformationnelle dépendant du temps. Contrairement à la minimisation de l'énergie, la dynamique moléculaire est capable de franchir des barrières énergétiques et d'atteindre des états ayant une énergie libre plus faible. Plusieurs outils de simulation visent à surmonter les limites temporelles de la dynamique moléculaire classique tout-atomique, comme la métadynamique, l'inondation conformationnelle, la MD accélérée ou la MD ciblée. Une autre solution consiste à utiliser des modèles coarse-grained plutôt que d'inclure tous les détails atomistiques.

L'analyse de modes normaux peut être utilisée efficacement pour identifier et décrire les mouvements intrinsèques les plus lents des macromolécules, qui, dans la nature, correspondent généralement bien aux mouvements fonctionnels collectifs. L'analyse de modes normaux repose sur l'approximation harmonique de la fonction d'énergie potentielle autour d'un minimum local.

Dans la modélisation des interactions protéine-ligand, les simulations d'arrimage moléculaire peuvent être utilisées efficacement pour identifier les sites de liaison et les poses des ligands, et prédire les affinités de liaison. Les simulations d'arrimage s'approchent des énergies physiques réelles avec une énergétique et une solvatation simplifiées, elles génèrent et classent les configurations protéine-ligand en utilisant différents algorithmes de recherche et fonctions de notation.

À partir d'un ensemble de données d'entraînement de composés dont l'activité est connue, les caractéristiques basées sur les ligands et la structure peuvent être utilisées pour entraîner des modèles de prédiction par apprentissage automatique pour la classification (ou la régression). La forêt aléatoire (random forest, RF) et la machine à vecteur de support

(support-vector machine, SVM) sont des exemples largement utilisés d'algorithmes d'apprentissage automatique supervisés, et ils peuvent être utilisés efficacement pour distinguer les composés actifs et inactifs d'une protéine cible donnée.

3. Objectifs

L'une des principales raisons de l'échec d'un candidat médicament est due à des problèmes liés à la pharmacocinétique/pharmacodynamique pendant les essais cliniques. La prédiction *in silico* des interactions avec les DME et les transporteurs de médicaments qui régissent la pharmacocinétique des xénobiotiques peut contribuer à réduire le taux d'échec des candidats médicaments à un stade précoce de leur développement, ainsi que les coûts associés et le nombre de tests sur les animaux. Dans mon travail de doctorat, je me suis concentré sur deux DME de phase II, SULT1A1 et UGT1A1, ainsi que sur un transporteur d'efflux de médicaments essentiel, ABCG2.

L'objectif de mon travail de doctorat était de déchiffrer les mouvements fonctionnels codés dans les protéines susmentionnées à l'aide de différentes approches de simulation, d'identifier les effets de la liaison du ligand sur leur dynamique et leurs mouvements collectifs fonctionnels, et finalement de créer des modèles d'apprentissage automatique prédictifs *in silico* capables de distinguer les inhibiteurs des composés inactifs par rapport aux protéines cibles.

4. Résultats

4.1. Mécanisme de liaison des substrats de SULT1A1

Il a été suggéré précédemment qu'un déplacement considérable de la structure de la SULT, causé par la liaison du cofacteur PAPS, pourrait contrôler la capacité de la SULT à lier des substrats de grande taille. Afin d'élucider les mécanismes moléculaires guidant la reconnaissance de divers substrats et inhibiteurs par SULT1A1, des simulations de dynamique moléculaire (MD) et l'approche récemment développée de ‘MD with excited Normal Modes’ (MDeNM) ont été réalisées en présence du cofacteur. Les simulations MDeNM permettent un échantillonnage étendu de l'espace conformationnel en réalisant plusieurs simulations MD courtes au cours desquelles les mouvements décrits par un sous-ensemble de modes normaux de basse fréquence sont cinétiquement excités. L'analyse de ‘root mean square deviation’ (RMSD) calculé sur la poche de liaison et la protéine entière a révélé que MDeNM a effectué un échantillonnage conformationnel plus exhaustif de la poche de liaison de SULT1A1 que les simulations MD classiques tout en maintenant la structure globale de la protéine plus proche de la structure cristalline initiale. L'analyse des ‘root mean square fluctuation’ des atomes Cα a révélé que MDeNM a particulièrement amplifié les mouvements liés aux boucles L1 et L3, et modérément ceux de L2, les trois boucles formant la porte de la poche de liaison. Les fluctuations aux extrémités des boucles L1 et L3 sont deux fois plus grande dans le cas de MDeNM que celles des simulations MD classiques, ce qui indique que MDeNM explore davantage les mouvements de la porte. L1 a présenté une plus grande fluctuation que L3 à la fois par MD et MDeNM, ce qui implique son rôle dans le mécanisme de gating comme cela a

été proposé précédemment. L'analyse de l'ouverture de la porte a révélé que MDeNM a atteint des conformations beaucoup plus largement ouvertes, alors que MD a cartographié des états fermés très peuplés.

Pour mieux comprendre le mécanisme des interactions entre SULT1A1 et les ligands, l'arrimage de 132 substrats ou inhibiteurs connus a été effectué dans la poche de liaison des conformations centroïdes recueillies par MD et MDeNM après regroupement. De nombreux ligands ont exprimé un comportement d'arrimage similaire dans l'ensemble des conformations recueillies par MD et MDeNM en termes d'énergie d'interaction (IE), pour certains ligands, cependant, des différences considérables ont été observées. La plupart de ces composés ont montré une énergie d'interaction plus favorable lorsqu'ils ont été amarrés à l'ensemble des conformations MDeNM, ce qui démontre l'avantage d'inclure les simulations MDeNM en plus des MD. En particulier, l'évaluation des ligands pour lesquels il y avait une différence significative entre MD et MDeNM a révélé que la plupart des composés pour lesquels MDeNM était plus performant étaient de grande taille, occupant un grand volume dans la poche de liaison, et que leurs poses correspondant au meilleur IE étaient logées dans des conformations enzymatiques largement ouvertes qui étaient peu peuplées ou même non accessibles par les simulations MD classiques.

Deux substrats de taille différente ont été utilisés dans des simulations d'arrimage supplémentaires afin d'étudier plus en détail le mécanisme de porte et la reconnaissance du substrat de SULT1A1. Le substrat 17β-estradiol (E2) est un substrat plus petit et de taille moyenne tandis que le fulvestrant, un analogue d'œstrogène, est un substrat plus grand de SULT1A1, avec une chaîne latérale fonctionnelle supplémentaire de 15 atomes de long. Les deux substrats ont été arrimés dans 6000 conformations générées par MD et 6000 autres conformations générées par MDeNM. Les conformations MD et MDeNM étaient capables d'accueillir l'E2, indépendamment de leur ouverture, ce qui concorde avec les études cinétiques et de liaison précédentes montrant que l'E2 peut se lier à des conformations ouvertes et fermées de l'enzyme. Le fulvestrant a montré une préférence évidente pour les conformations ouvertes. Les positions d'arrimage les plus favorables de l'E2 et du fulvestrant ont été analysées plus en détail et se sont révélées stables par des simulations MD classiques consécutives à partir des complexes enzyme-cofacteur-substrat. Les simulations MD en présence d'estradiol et de fulvestrant ont montré l'ouverture supplémentaire induite des boucles en présence du substrat lié. Les résultats des simulations MD et MDeNM ont confirmé que des conformations ouvertes sont disponibles pour les gros ligands (tels que le fulvestrant, le 4-hydroxytamoxifène ou le raloxifène) pour se lier même en présence du cofacteur lié. Ainsi, nos résultats ont proposé un nouveau mécanisme pour la reconnaissance des grands substrats sans impliquer un mouvement du cofacteur pour déclencher la liaison des grands substrats.

4.2. Prévision de l'inhibition de l'UDP-glucuronosyltransférase

Une forte inhibition de l'UGT1A1 peut déclencher des interactions médicamenteuses indésirables ou entraîner des troubles du métabolisme endobiotique. Afin d'étudier le degré de flexibilité structurelle et l'adaptation conformationnelle de sa poche de liaison à la lumière

de la variété structurale des divers composés actifs, des simulations de dynamique moléculaire ont été réalisées à partir d'un modèle d'homologie humaine en présence du cofacteur (UDPGA). L'analyse de 'root mean square deviation' a montré que les trois simulations de dynamique moléculaire ont rapidement dévié de la conformation initiale, atteignant des différences relativement élevées. Des mouvements d'ouverture-fermeture inter domaines plus importants ont été révélés par l'analyse du rayon de giration et l'inspection visuelle des trajectoires. Le volume de la poche de liaison au substrat varie de manière significative au cours des simulations. Dans certaines conformations, le volume atteint 1,5 à 2 fois la taille de la structure de départ. Les grandes variations du volume de la poche de liaison au substrat et l'ouverture vers le lumen peuvent faciliter l'accès au site catalytique et accueillir les divers substrats.

Des ligands connus de l'UGT1A1, substrats et inhibiteurs, ont été rassemblés pour être utilisés pour l'arrimage et construire des ensembles de données d'entraînement et de test externe pour la modélisation par apprentissage automatique. L'arrimage d'ensemble des molécules actives et des molécules leurres de l'ensemble d'entraînement a été réalisé sur les conformations centroïdes recueillies par MD après regroupement pour identifier les conformations de la protéine qui distinguent le mieux les composés actifs et inactifs. En utilisant la meilleure IE retenue pour chaque composé, des courbes d'enrichissement ont été calculées pour chaque conformation protéique centroïde, dont six ont été conservées pour la modélisation par apprentissage automatique et les scores d'accostage IE correspondants ont été utilisés comme descripteurs basés sur les interactions protéine-ligand. Les conformations protéiques identifiées présentaient une grande flexibilité au niveau de leurs poches de liaison au substrat, les résidus clés pour la réaction catalytique restaient moins flexibles, et conservaient leur orientation à l'intérieur de la poche de liaison.

Les descripteurs moléculaires physico-chimiques des composés actifs et des leurres ont été calculés, et les descripteurs fortement corrélés ainsi que les descripteurs dont la variance est proche de zéro ont été supprimés. Les six scores IE obtenus à partir des simulations d'arrimage ont été inclus en tant que descripteurs basés sur la structure représentant les interactions protéine-ligand. Une réduction du nombre de descripteurs a été effectuée sur la base de leur importance relative afin d'éviter un ajustement excessif après l'entraînement d'un certain nombre de modèles de forêt aléatoire sur l'ensemble de données d'entraînement et la sélection du sous-ensemble de descripteurs ayant la plus grande importance de Gini. Un total de 56 descripteurs basés sur le ligand et les 6 IE ont été conservés pour la construction de modèle, les descripteurs les plus importants étant liés à la polarité, à la lipophilie et aux charges. Les algorithmes random forest (RF) et support-vector machine (SVM) ont été utilisés pour créer des modèles de prédiction optimisés intégrant les informations basées sur les ligands et la structure, en optimisant les hyper paramètres correspondants du modèle. Les modèles RF et SVM ont montré d'excellents pouvoirs prédictifs pour la discrimination des molécules actives de l'UGT1A1 avec une précision et une sensibilité d'environ 90 %. Ce sont les premiers modèles d'apprentissage automatique créés pour la prédiction de l'inhibition de l'UGT1A1. Ils peuvent

être utiles pour la prédiction des interactions médicamenteuses de nouveaux candidats médicaments tout en fournissant également des informations structurales sur les interactions enzyme-ligand.

4.3. Mécanisme de transport ABCG2

ABCG2 est d'un intérêt majeur en raison de son implication dans la multirésistance aux médicaments (MDR), et la compréhension de son mécanisme d'efflux complexe est essentielle pour prévenir la MDR et les interactions médicament-médicament (DDI). L'exportation d'ABCG2 est caractérisée par deux transitions conformationnelles majeures entre les états orientés vers l'intérieur et vers l'extérieur (IFS et OFS), dont les structures ont été résolues. Pourtant, l'ensemble du cycle de transport n'a pas été caractérisé à ce jour. Afin d'élucider le mécanisme de transport d'ABCG2 et son comportement dans les différentes étapes du transport, différentes simulations ont été réalisées à partir des structures cryo-EM disponibles correspondant aux deux états, son IFS et son OFS. Dans l'IFS, les deux monomères du domaine de liaison aux nucléotides (NBD) sont partiellement séparés alors que dans l'OFS, ils forment un dimère serré établissant les deux sites de liaison aux nucléotides. La paire de domaines transmembranaires (TMD) forme la cavité hydrophobe 1 en forme de fente, où il a été prouvé que les substrats physiologiques et divers inhibiteurs se lient, et la cavité 2 qui fait face à l'espace extracellulaire. La cavité 1 est repliée dans l'OFS, les deux cavités sont séparées par une porte en leucine.

Tout d'abord, les régions de boucle intracellulaire manquantes dans les NBD ont été modélisées car elles sont susceptibles d'affecter l'entrée de substrat et peuvent posséder une fonction de porte. De même, le segment de liaison a également été modélisé, reliant les NBD individuels et les TMD. Des systèmes multiples ont été construits à partir des structures expérimentales IFS et OFS en présence ou en l'absence d'un substrat physiologique (estrone 3-sulfate, E_1S) et de nucléotides (ATPs ou ADPs) pour imiter les différentes étapes du transport :

- Apo IFS
- IFS lié à E_1S
- IFS lié à E_1S et ATP-Mg^{2+}
- OFS lié à l'ATP-Mg^{2+}
- OFS lié à l'ADP
- OFS sans nucléotide

Des simulations de dynamique moléculaire (MD) intégrées dans une bicouche lipidique et une analyse en mode normal (NMA) ont été réalisées sur les six systèmes ci-dessus. L'échelle de temps d'un cycle de transport complet d'ABCG2 est de l'ordre de la fraction de seconde ou plus, un délai qui ne peut actuellement pas être simulé par la MD classique. J'ai donc développé un outil de simulation MD amélioré, la 'kinetically excited targeted MD' (ketMD), qui trace les chemins possibles entre deux structures terminales ; il a été appliqué à l'ABCG2 pour simuler les transitions conformationnelles entre les états IFS et OFS. De la même manière que pour la MD ciblée (tMD), le vecteur d'excitation a été choisi pour pointer vers la structure cible,

cependant, contrairement à la tMD, où la fonction d'énergie potentielle est biaisée et où la protéine est guidée par des forces directrices à chaque étape de la simulation, la ketMD repose sur des excitations cinétiques. Lors de la première étape de chaque cycle d'excitation, les composantes de vitesse pointant de la conformation instantanée vers la structure cible sont augmentées, permettant le franchissement de barrières énergétiques plus importantes. Cette étape d'excitation est suivie d'une période de relaxation pendant laquelle aucune perturbation externe n'est appliquée, le système peut évoluer et l'énergie cinétique injectée peut se dissiper. Après chaque cycle d'excitation, le vecteur de direction de l'excitation est mis à jour pour pointer de la conformation instantanée vers la structure cible.

La transition 1 (IFS vers OFS) a été simulée à partir de l'IFS lié à E_1S et à ATP-Mg^{2+}, tandis que la transition 2 (OFS vers IFS) à partir de l'OFS sans substrats liés et en présence d'ATP-Mg^{2+}, d'ADP, ou en l'absence de nucléotides liés. L'analyse de ‘root mean square deviation’, le suivi d'une paire de distances catalytiques et l'inspection visuelle ont confirmé qu'une dimérisation complète des NBD a été réalisée avec succès en même temps que la formation du site de liaison ATP catalytique pendant la simulation ketMD de la transition 1, alors que dans la direction opposée, les NBD se sont partiellement séparés et une transition complète des NBD s'est produite pendant les simulations ketMD. Au cours de la transition 2, le détachement des NBDs au niveau du site de liaison à l'ATP était plus facile à réaliser en l'absence de nucléotides, et plus difficile en présence d'ATP.

Une analyse du rayon de giration des segments hélicoïdaux bordant les deux cavités a été réalisée pour étudier leur état au cours des simulations. A la fin de la simulation ketMD de la transition 1, la cavité 1 était effondrée (identique à la structure OFS) alors qu'en sens inverse, à partir d'une cavité 1 effondrée, elle redevenait exposée et accessible depuis le cytosol. À l'opposé, pendant la simulation de la transition 1 par la ketMD, la cavité 2 est devenue plus exposée à l'espace extracellulaire et également plus volumineuse, tandis que pendant la simulation de la transition 2, en partant de l'OFS, la cavité 2 s'est rapprochée d'un état plus dégonflé.

Des structures expérimentales avec un substrat lié à la cavité 2, ou des structures transitoires le long de la voie de translocation n'ont pas été résolues. Avec l'exécution de la ketMD à partir de l'IFS, il a été possible de simuler la translocation de E_1S de la cavité 1 vers l'espace extracellulaire à travers la porte de la leucine et la cavité 2. En plus de l'excitation appliquée au transporteur, le mouvement du substrat a également été cinétiquement promu vers l'espace extracellulaire. Le substrat était initialement lié à la cavité 1, stabilisé principalement par les interactions d'empilement " en sandwich " de F439 et F439’. Dès que le substrat s'est échappé du piège 'sandwich' des deux résidus F439 et s'est déplacé vers la cavité 2, F439 et F439’ sont entrés en contact étroit, créant une construction en forme de valve, similaire à ce qui est observé dans la structure cryo-EM de OFS. Tout type de mouvement de retour vers le cytosol était empêché avec cette valve fermée et j'ai identifié une formation de type poche entre les cavités 1 et 2, où le substrat était piégé car les mouvements vers la cavité 2 restaient encore bloqués par la porte leucine fermée. Plusieurs résidus au bord de ce site ont

stabilisé la position du substrat, il ne s'est éloigné de la poche qu'une fois les résidus de la porte leucine séparée. Le comportement du substrat dans la cavité 2 était très différent de celui observé dans la cavité 1 ou entre les cavités 1 et 2. Avant d'arriver à la cavité 2, le substrat était étroitement lié et entouré de près par les résidus du transporteur. En revanche, le substrat était faiblement lié ici alors qu'il explorait le volume de la cavité, établissant des contacts étroits avec les résidus à sa limite. L'excitation cinétique supplémentaire du substrat a entraîné son détachement complet du transporteur dans l'espace extracellulaire.

Les simulations MD classiques et l'analyse en mode normal des différentes étapes de transport ont révélé que la présence du substrat dans la cavité 1 est essentielle pour coupler les mouvements entre les NBDs et les TMDs. En l'absence d'un substrat et de nucléotides liés (forme apo), les TMDs se rapprochent d'une configuration neutre, la cavité 2 s'ouvre alors que la cavité 1 commence à s'effondrer, tandis que les NBDs restent éloignés les uns des autres. Contrairement à la cavité 1, la cavité 2 ne s'est jamais complètement effondrée, quelle que soit l'étape du cycle de transport (pendant les simulations MD classique et ketMD), même si les expériences suggèrent qu'elle peut s'occlure. Son volume présente des variations de gonflement et de dégonflement entre les états IFS et OFS, mais la cavité 2 est plus volumineuse que la cavité 1, même dans son état dégonflé. Les simulations ketMD et les simulations MD classiques ultérieures ont démontré que E_1S, qui est un composé volumineux, pouvait être présent dans la cavité 2, même dans l'état OFS, et elles ont également montré un espace suffisamment grand pour le substrat dans la cavité 2 pendant tout le cycle de transport. Cela peut permettre la fixation simultanée du substrat dans les cavités 1 et 2, entraînant un mécanisme d'exportation accéléré.

L'approximation harmonique des modes normaux a révélé que des mouvements de type transition sont présents dans le système IFS correspondant à de basses fréquences, alors qu'ils sont amortis et sont des mouvements locaux de fréquences plus élevées dans les systèmes OFS. Cela suggère qu'il est énergétiquement coûteux d'amorcer la transition 2 pour revenir à l'IFS initial, ce qui peut nécessiter l'énergie libérée lors de l'hydrolyse de l'ATP.

5. Conclusions

La prédiction du devenir des médicaments administrés dans le corps humain reste une tâche difficile qui s'est longtemps appuyée sur des études coûteuses sur les animaux, ce qui soulève également des problèmes d'éthique et de fiabilité. C'est pourquoi un nombre croissant d'approches informatiques ont été utilisées afin de prédire efficacement le résultat métabolique des médicaments et des nouveaux candidats médicaments. Plusieurs études ont créé des modèles de prédiction utilisant des informations basées sur le ligand ou la structure. En 2013, le laboratoire hôte a formé les premiers modèles de classification par apprentissage automatique intégrant des informations basées sur le ligand et la structure pour la prédiction de l'inhibition de différentes isoformes de SULT, puis des enzymes CYP. La compréhension des mouvements fonctionnels collectifs des protéines et de leurs interactions avec de petits ligands est une étape cruciale pour extraire des informations basées sur la structure pour de tels modèles de prédiction.

Dans mon travail de doctorat, je me suis concentré sur deux DME essentielles de la phase II, SULT1A1 et UGT1A1, ainsi que sur le transporteur d'efflux de médicaments ABCG2. Mes recherches ont porté à la fois sur la dynamique des différentes protéines et sur leurs interactions avec de petites molécules. Elles comprenaient également des développements méthodologiques, tels qu'un outil de simulation de dynamique moléculaire amélioré et différents outils de prédiction par l'intégration de la modélisation basée sur la structure et l'apprentissage automatique.

Auparavant, il avait été suggéré que la liaison de la PAPS limitait l'accès de substrats plus importants à la SULT1A1. Des simulations MD classiques et 'MD with excited Normal Modes' (MDeNM) ont été réalisées afin d'élucider les mécanismes moléculaires guidant la reconnaissance de divers substrats et inhibiteurs par SULT1A1 en présence du cofacteur, ce qui a été suivi par des simulations d'arrimage de ligands connus. Les résultats d'arrimage et des simulations classiques de dynamique moléculaire qui ont suivi sur les complexes ont démontré que d'importants changements de conformation de la SULT1A1 pouvaient se produire même en présence du cofacteur, et que des conformations plus largement ouvertes peuvent accueillir un grand substrat comme le fulvestrant avec une plus grande affinité. En plus de L3, la région de la boucle L1 à l'entrée du site catalytique présentait une flexibilité extrêmement élevée qui soulignait son importance fonctionnelle dans le mécanisme de déclenchement de la SULT1A1. Dans l'ensemble, mon travail a jeté une nouvelle lumière sur les mécanismes complexes de la spécificité du substrat et de l'inhibition de la SULT1A1, et a souligné l'utilité d'inclure MDeNM dans les études d'interactions protéine-ligand où des réarrangements majeurs sont attendus. Les conformations générées peuvent être utilisées efficacement pour entraîner des modèles de prédiction in silico avec une précision encore plus élevée pour les molécules plus grandes.

La modélisation basée sur la structure et les informations basées sur le ligand ont été intégrées pour construire les premiers modèles de prédiction par apprentissage automatique de l'inhibition de l'UGT1A1. Dans le cadre de mon travail de doctorat, j'ai réalisé des simulations de dynamique moléculaire de l'enzyme à partir d'un modèle d'homologie humain en présence du cofacteur afin d'explorer la variabilité structurale du site catalytique, et j'ai observé une flexibilité significative, importante pour l'accommodation des divers ligands. Cette étude a été suivie d'un regroupement conformationnel. Un ensemble de ligands connus, substrats et inhibiteurs, a été collecté et un arrimage d'ensemble a été réalisé pour identifier les conformations qui peuvent efficacement différencier les actifs des leurres. Des descripteurs basés sur les ligands ont été calculés. Enfin, une sélection raisonnable de descripteurs basés sur les ligands, ainsi que les énergies d'interaction obtenues à partir d'arrimage, ont été utilisées pour entraîner les premiers modèles d'apprentissage automatique pour la prédiction de l'inhibition de l'UGT1A1. Les modèles formés ont montré d'excellentes performances et ont été mis en œuvre dans le logiciel DrugME. DrugME peut être utilisé efficacement pour identifier de nouveaux inhibiteurs de l'UGT1A1 et peut être utile pour la prédiction des interactions médicamenteuses de nouveaux candidats médicaments, tout en fournissant également des informations structurales sur les interactions enzyme-ligand.

J'ai développé une technique de simulation MD améliorée et innovante, appelée ketMD (kinetically excited targeted Molecular Dynamics) pour simuler le cycle de transport de l'ABCG2 ainsi que la translocation du substrat physiologique, le 3-sulfate d'estrone. À l'aide de ketMD, j'ai réussi à simuler les transitions conformationnelles entre les deux états extrêmes de l'ABCG2 et j'ai révélé le mécanisme moléculaire complexe de la translocation du substrat E_1S. J'ai observé une fonction de valve de deux résidus de phénylalanine (F439 et F439') dans la cavité 1 qui s'engagent initialement dans des interactions d'empilement contre le substrat. Les simulations MD classiques et l'analyse des modes normaux ont révélé que la présence du substrat dans la cavité 1 est nécessaire pour coupler les mouvements du domaine transmembranaire au domaine de liaison aux nucléotides. J'ai également observé que la cavité 2 n'était jamais complètement effondrée, à n'importe quel stade du cycle de transport, ce qui pourrait permettre la fixation simultanée du substrat dans les cavités 1 et 2 et accélérer le taux d'exportation du transporteur. Le travail sur ABCG2 a jeté une nouvelle lumière sur le mécanisme moléculaire complexe de son transport et a mis en évidence l'utilité d'inclure des techniques d'échantillonnage *in silico* améliorées, telles que la ketMD, dans les études sur les transporteurs. Dans un avenir proche, les conformations générées le long de la voie de transition seront utilisées pour développer de nouveaux modèles d'apprentissage automatique pour la prédiction des inhibiteurs et des substrats d'ABCG2, afin de sonder de nouveaux candidats médicaments pour la multirésistance et de prédire les interactions médicamenteuses.

Au cours des trois années de mon travail de doctorat, j'ai eu la grande chance d'approfondir mes connaissances en biologie sur le métabolisme des médicaments, de leur élimination et de leurs interactions, ainsi que dans l'application et le développement de nouvelles approches *in silico* pour étudier les enzymes et les transporteurs qui régissent ces processus complexes. Les observations et les résultats des travaux présentés dans la thèse contribuent à une meilleure compréhension des mécanismes moléculaires des enzymes du métabolisme de phase II et des transporteurs ABC, et de leurs interactions avec des ligands. En particulier, le logiciel DrugME qui incorpore maintenant les modèles prédictifs développés intégrant la modélisation basée sur la structure et l'apprentissage automatique contribuera grandement aux futures prédictions des interactions médicamenteuses.

J. Supporting Information – SULT1A1

Insights into the substrate binding mechanism of SULT1A1 through Molecular Dynamics with excited Normal Modes simulations

Balint Dudas[1,2,#], Daniel Toth[3,#], David Perahia[2], Arnaud B. Nicot[4], Erika Balog[3,*], Maria. A. Miteva[1,*]

[1]Inserm U1268 MCTR, CiTCoM UMR 8038 CNRS - University of Paris, Pharmacy Faculty of Paris, France
[2]Laboratoire de Biologie et Pharmacologie Appliquée, Ecole Normale Supérieure Paris-Saclay, UMR 8113, CNRS, Gif-sur-Yvette, France
[3]Department of Biophysics and Radiation Biology, Semmelweis University, Budapest, Hungary
[4]Inserm, Université de Nantes, Centre de Recherche en Transplantation et Immunologie, UMR 1064, ITUN, F-44000 Nantes, France

[#]1st coauthors
[*]corresponding authors: maria.mitev@inserm.fr, balog.erika@med.semmelweis-univ.hu

Published in Scientific Reports on 2021 Jun 23
doi: 10.1038/s41598-021-92480-w

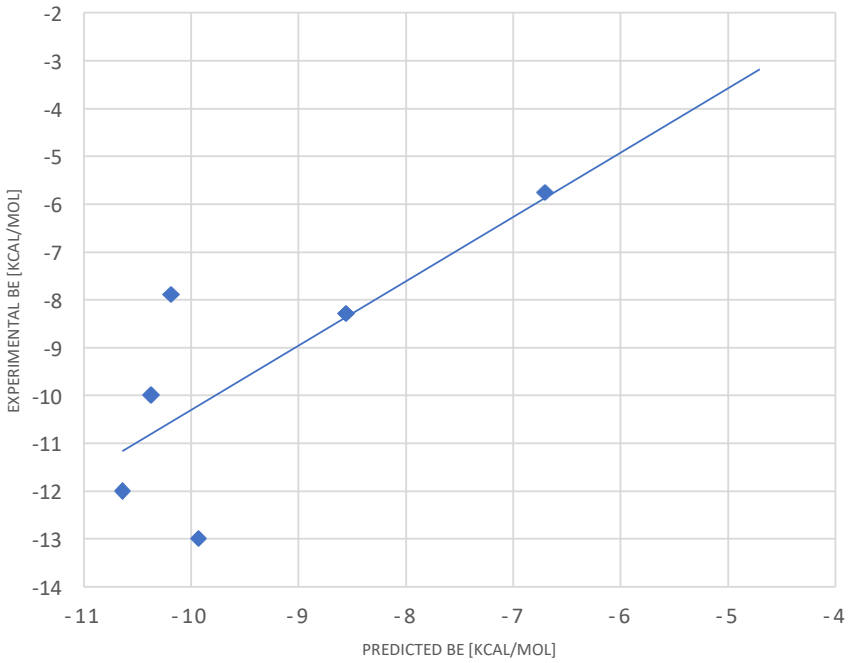

Figure S1. Predicted and experimental binding energy correlation for SULT1A1 ligands. Predicted bindingenergies (BE) were calculated by averaging over the best scored Autodock Vina energies obtained for the best 10 MD conformations and the best 10 MDeNM conformations. The experimental binding energies (BE) were taken or calculated using ligand affinity constants as reported in the literature (see SI Table S1).

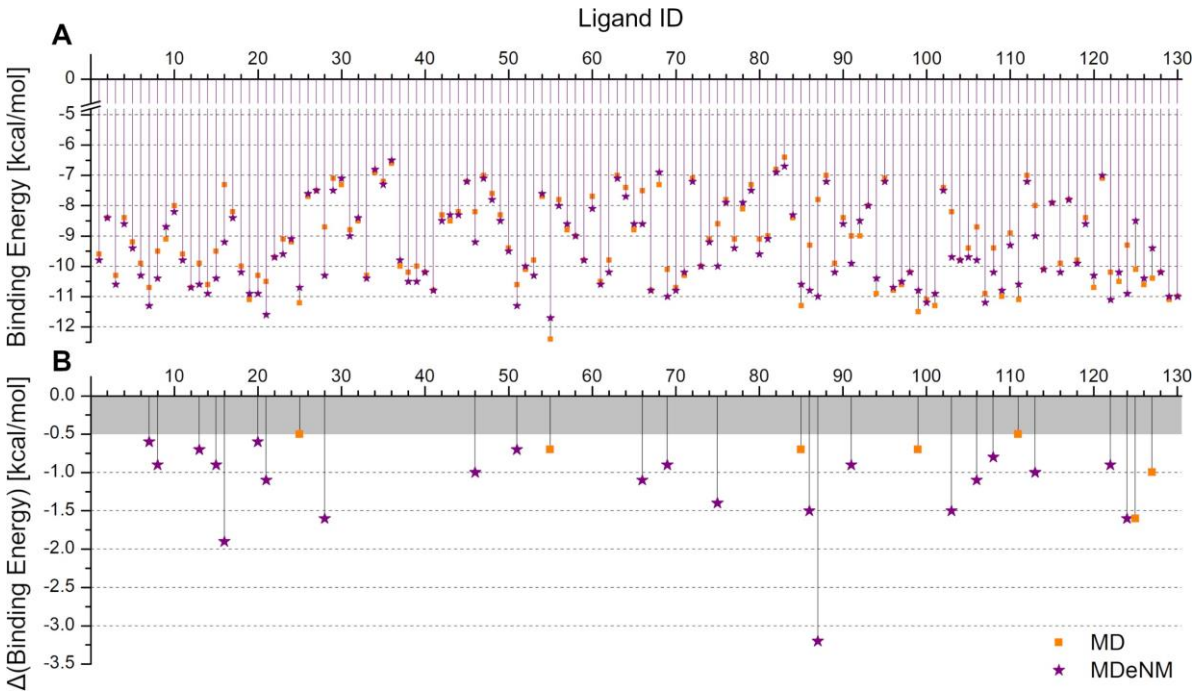

Figure S2. A. The best retained BE for each of the 132 known ligands over the MD (denoted by orange squares) and MDeNM (denoted by purple stars) conformational ensemble and B. the differences between the best BEs retained by MD and MDeNM conformations with the 5 Å distance criterion applied to the substrates. For the better visualization, only differences larger than 0.5 kcal/mol are indicated.

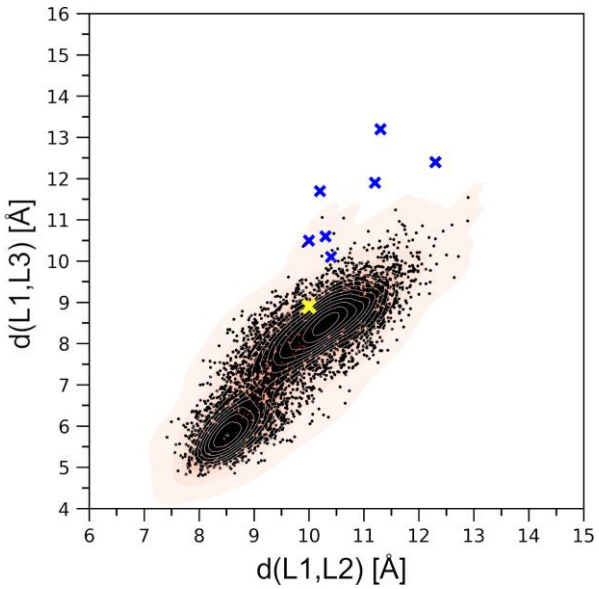

Figure S3. Distribution in the plane defined by d(L1,L2) vs. d(L1,L3) distances of all the MD generated structures (black dots) and the MDeNM structures (blue 'x'-es) that can accommodate competent orientations of bigger ligands with BEs inaccessible for any MD generated conformation. The location of the crystal structure (4GRA.pdb) is shown in yellow 'x'.

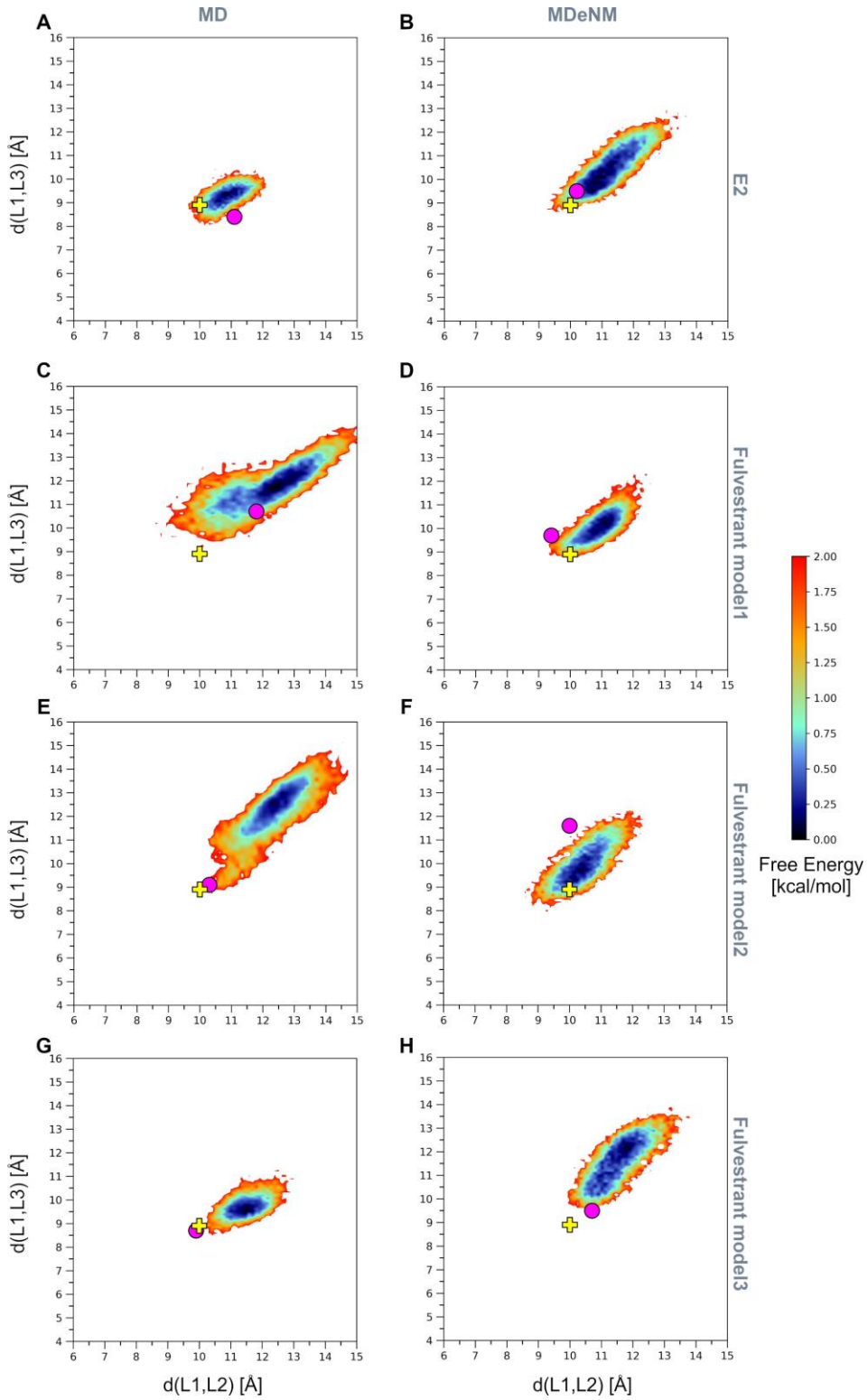

Figure S4. Free Energy Landscapes (FELs) of the complexes substrate-SULT1A1/PAPS in the space defined by the distances d(L1,L2) and d(L1,L3) of the 100 ns long MD simulations starting from an estradiol-bound MD (A) and MDeNM (B) conformations; and from fulvestrant-bound MD (C, E, G) and MDeNM (D, F, H) conformations. All starting complexes were taken after the docking with AutoDock Vina. The crystal structure (4GRA.pdb) is denoted by yellow '+'. The starting conformation for the MD simulations is denoted by a violet circle.

Table S1. Predicted and experimental binding energies for SULT1A1 ligands. Predicted binding energies were calculated by averaging over the best scored Autodock Vina energies obtained for the best 10 MD conformations and the best 10 MDeNM conformations. The experimental binding energies were taken or calculated using ligand affinity constants as reported in the literature.

Compound	*Predicted Binding Energy (kcal/mol)*	*Experimental Binding Energy (kcal/mol)*	*REF*
apomorphine	*-9.9*	*-13.0*	*1*
ethinyl estradiol	*-10.6*	*-12.0*	*2*
p-nitrophenol	*-6.7*	*-5.8*	*3*
4-hydroxytamoxifen	*-8.6*	*-8.3*	*4*
17β-estradiol (E2)	*-10.4*	*-10.0*	*1*
fulvestrant	*-10.2*	*-7.9*	*5*

1. Thomas, N. L., and Coughtrie, M. W. (2003) Sulfation of apomorphine by human sulfotransferases. Evidence of a major role for the polymorphic phenol sulfotransferase, SULT1A1. Xenobiotica 33, 1139–1148
2. Rohn, K. J., Cook, I. T., Leyh, T. S., Kadlubar, S. A., and Falany, C. N. (2012) Potent inhibition of human sulfotransferase 1A1 by 17 - ethinylestradiol. Role of 3 -phosphoadenosine 5 -phosphosulfate binding and structural rearrangements in regulating inhibition and activity. Drug. Metab. Dispos. 40, 1588–1595
3. Lu-Yi Lu, Han-Ping Chiang, Wei-Ti Chen, and Yuh-Shyong Yang Dimerization Is Responsible for the Structural Stability of Human Sulfotransferase 1A1. DRUG METABOLISM AND DISPOSITION. 37:1083–1088, 2009
4. Ting Wang, Ian Cook, and Thomas S. Leyh , 3′-Phosphoadenosine 5′-Phosphosulfate Allosterically Regulates Sulfotransferase Turnover, Biochemistry 2014, 53, 6893−6900
5. Cook, I., Wang, T., Almo, S. C., Kim, J., Falany, C. N., and Leyh, T. S. (2013) The gate that governs sulfotransferase selectivity. Biochemistry 52, 415–424

Table S2. Binding energies (BE) of SULT1A1 substrates calculated with Autodock Vina scoring functionbefore and after MD simulations of 100 ns starting from 8 different substrate-SULT1A1/PAPS structures obtained by docking.

Substrate	*Starting SULT1A1/PAPS Conformation taken from*	*Complex No.*	*BE (before MD)* [kcal/mol]	*BE (after MD)* [kcal/mol]
E2	*MD*	*1*	*-11.0*	*-10.8*
	MDeNM	*1*	*-11.4*	*-10.4*
Fulvestrant	*MD*	*1*	*-10.7*	*-8.8*
		2	*-10.6*	*-9.0*
		3	*-9.9*	*-8.5*
	MDeNM	*1*	*-11.1*	*-8.1*
		2	*-10.1*	*-7.6*
		3	*-10.0*	*-9.4*

List of residues forming the binding pocket:
I21, F24, T45, Y46, P47, F81, F84, K85, A86, I89, K106, T107, H108, F142, A146, K147, V148, H149, Y169, Y240, T241, T242, V243, P244, Q245, E246, F247, M248, D249, H250, F255

K. Supporting Information – UGT1A1

Machine learning and structure-based modeling for the prediction of UDP-glucuronosyltransferase inhibition

Balint Dudas[1,2,#], Youcef Bagdad[1,#], Milan Picard[1,Φ] , David Perahia[2], Maria A. Miteva[1,*]

[1]Inserm U1268 MCTR, CiTCoM UMR 8038 CNRS – Université Paris Cité, Paris, France
[2]Laboratoire de Biologie et Pharmacologie Appliquée (LBPA), UMR8113, Ecole Normale Supérieure Paris-Saclay, Gif-sur-Yvette, France
ΦPresent address: Molecular Medicine Department, CHU de Québec Research Center, Université Laval, Québec, Canada

#1st coauthors
*corresponding author: maria.mitev@inserm.fr

Accepted in iScience on 2022 Oct 6

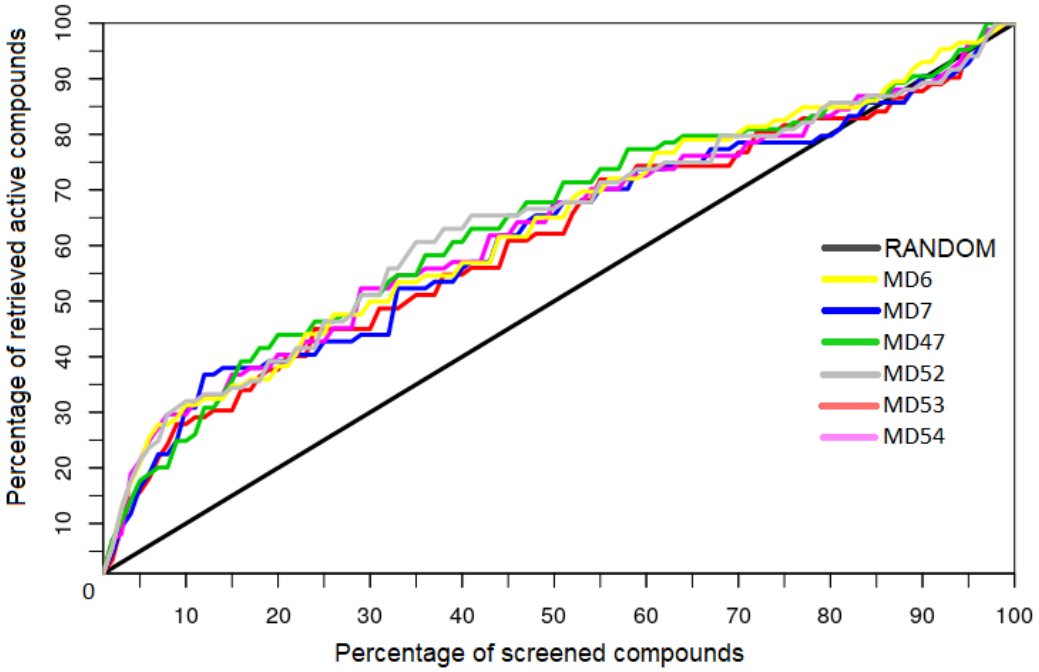

Figure S1. Enrichment curves representing the percentage of actives retrieved at a percentage of the screened compounds including actives and decoys calculated by retaining the best docking score for each compound in each protein conformation. The six UGT1A1 MD conformations showing the best enrichment are presented. 100 % refers to all screened compounds including all actives and decoys.

List of residues to define the volume of the substrate-binding pocket of UGT1A1:
V109, Y113, D151, P152, F153, L172, H173, A174, F181, E182, F190, S191, Y192, V193, P194, F217, S218, F221, D246, S249

Table S1. Parameters of the optimized best RF and SVM models.

Parameters	Default value	56 MOE + 6 IE	162 MOE + 6 IE
RF *ntree*	500	256	256
RF *mtry*	$\sqrt{p}$	15	29
RF *sample size* (actives / decoys)	total number of molecules in each class	53 / 48	58 / 58
SVM *cost*	1	2^4	2^5
SVM *gamma*	1/(*p* x Var)	2^{-8}	2^{-10}
SVM *weight* (actives / decoys)	0.5/0.5	0.91 / 0.09	0.96 / 0.04

p is the total number of variables; Var is the variance of the training dataset.

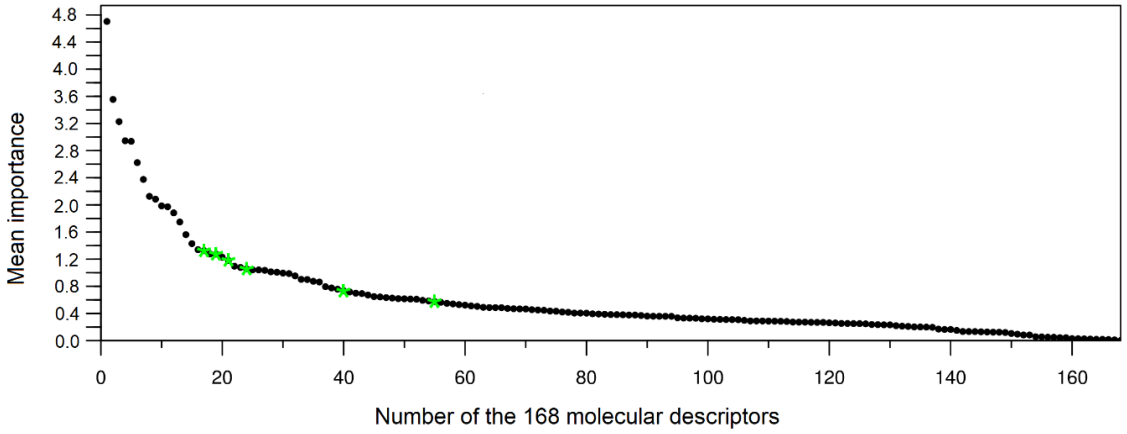

Figure S2. Mean importance inferred from the random forest modeling of the 168 molecular descriptors used to train models for UGT1A1 prediction. Green stars denote the interaction energy descriptors of the six MD conformations.

Table S2. Mean importance inferred from the random forest modeling of the best 21 physicochemical descriptors and 6 IEs (Interaction Energies) discriminating between inhibitors of UGT1A1 and decoys.

Categories	Importance rank	Descriptor name	Description	Importance
MOE 2D molecular descriptors				
Polarity and charges	1	h_pavgQ	average total charge	4.70
	2	wienerPol	Wiener polarity number	3.56
	4	h_pKa	pKa	2.94
	5	PEOE_VSA-6	vdW surface area for atoms having charge < -0.30	2.94
	9	PEOE_PC-	total negative partial charge	2.09
	15	h_pstates	fractional number of protonation states	1.43
	18	a_donacc	number of hydrogen bond donors and acceptors	1.278
	22	PEOE_VSA_POS	positive van der Waals surface area	1.10
Lipophilicity	7	SlogP_VSA0	sum of vdW area for atoms contributing to SlogP <= -0.4	2.37
	10	GCUT_SLOGP_3	GCUT descriptor using atomic contribution to SlogP	1.99
Atom types and bonds	8	a_nN	number of nitrogen atoms	2.13
	12	a_ICM	atom information content (element distribution)	1.89
	13	weinerPath	Wiener path number: half the sum of all the distance matrix entries	1.75
	16	chi1_C	carbon connectivity index	1.34
	20	chiral	number of chiral centers	1.23
	23	b_rotR	fraction of rotatable bonds	1.08
Molar refractivity	11	GCUT_SMR_1	GCUT descriptors using atomic contribution to molar refractivity	1.97
MOE 3D molecular descriptors				
Polar volume	3	vsurf_Wp2	polar volume	3.23
	14	vsurf_D8	hydrophobic volume	1.56
	25	vsurf_Wp3	polar volume	1.04
Potential energy	6	AM1_Eele	electronic energy (kcal/mol) calculated using the AM1 Hamiltonian	2.62
Interaction energies (IE)	17	IE_MD52	IE of the MD52 structure	1.33
	19	IE_MD6	IE of the MD6 structure	1.27
	21	IE_MD54	IE of the MD54 structure	1.18
	24	IE_MD47	IE of the MD47 structure	1.06

	40	IE_MD53	IE of the MD53 structure	0.74
	55	IE_MD7	IE of the MD7 structure	0.58

The GCUT descriptors are calculated from the eigenvalues of a modified graph distance adjacency matrix. (e.g., the diagonal takes atomic contribution to SlogP or molar refractivity.

Table S3. Performances of the non-optimized RF models with MOE descriptors on the internal set (cross-validation CV) and external validation test set.

Descriptors	Data set	BA %	Sensitivity %	Specificity %	MCC %
56 MOE	Internal CV	82.0	65.0	99.0	74.3
	External	88.9	77.8	1.0	86.3
162 MOE	Internal CV	80.2	61.1	99.2	72.2
	External	84.8	70.4	99.3	78.9

Table S4. Performances of the non-optimized SVM models with MOE descriptors on the internal set (cross-validation CV) and external validation set.

Descriptors	Data set	BA %	Sensitivity %	Specificity %	MCC %
56 MOE	Internal CV	83.1	67.4	98.8	75.1
	External	88.9	77.8	1.0	86.3
162 MOE	Internal CV	75.2	50.6	99.8	66.7
	External	83.3	66.7	1.0	79.0

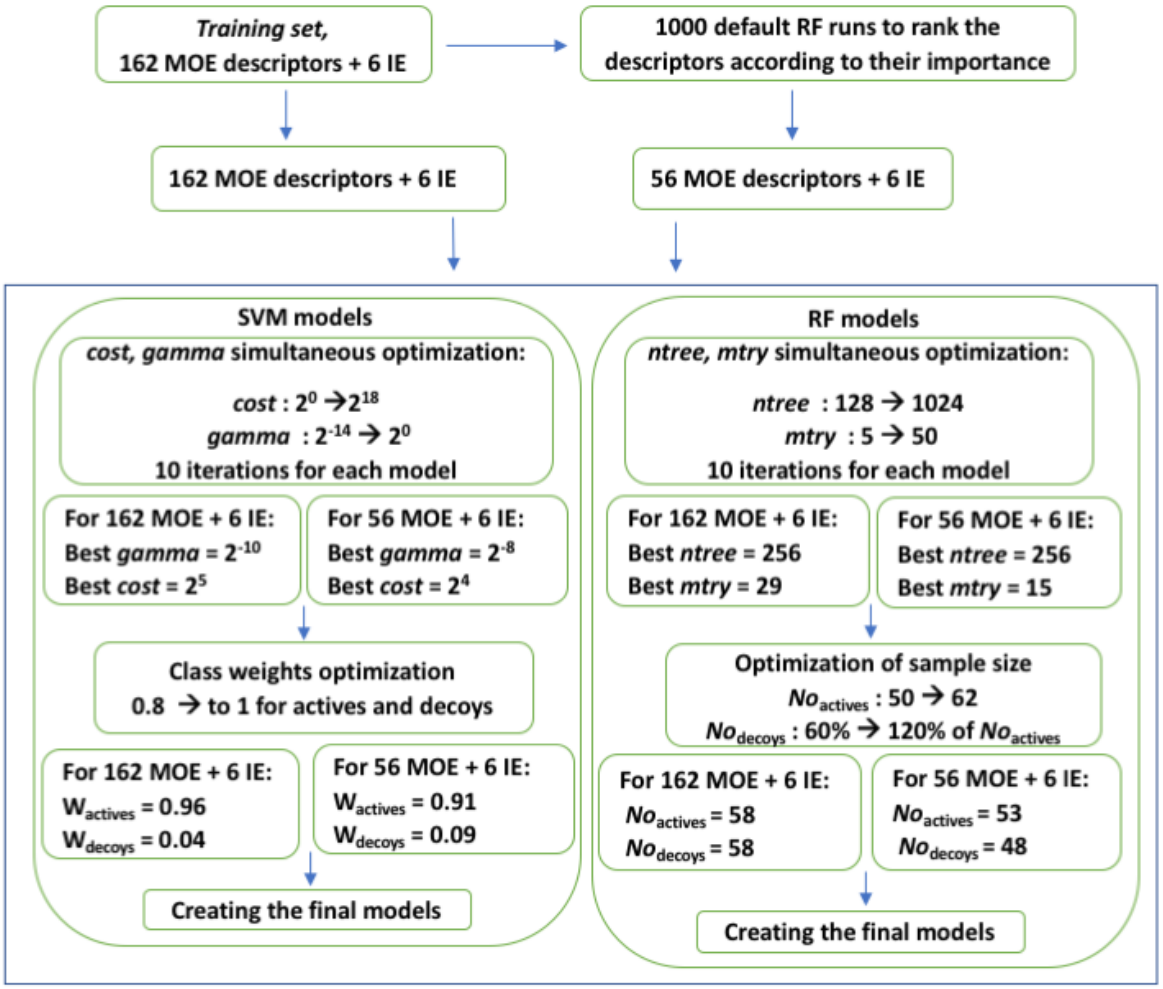

Figure S3. Scheme of the optimization performed to obtain the parameters of the best RF and SVM models.

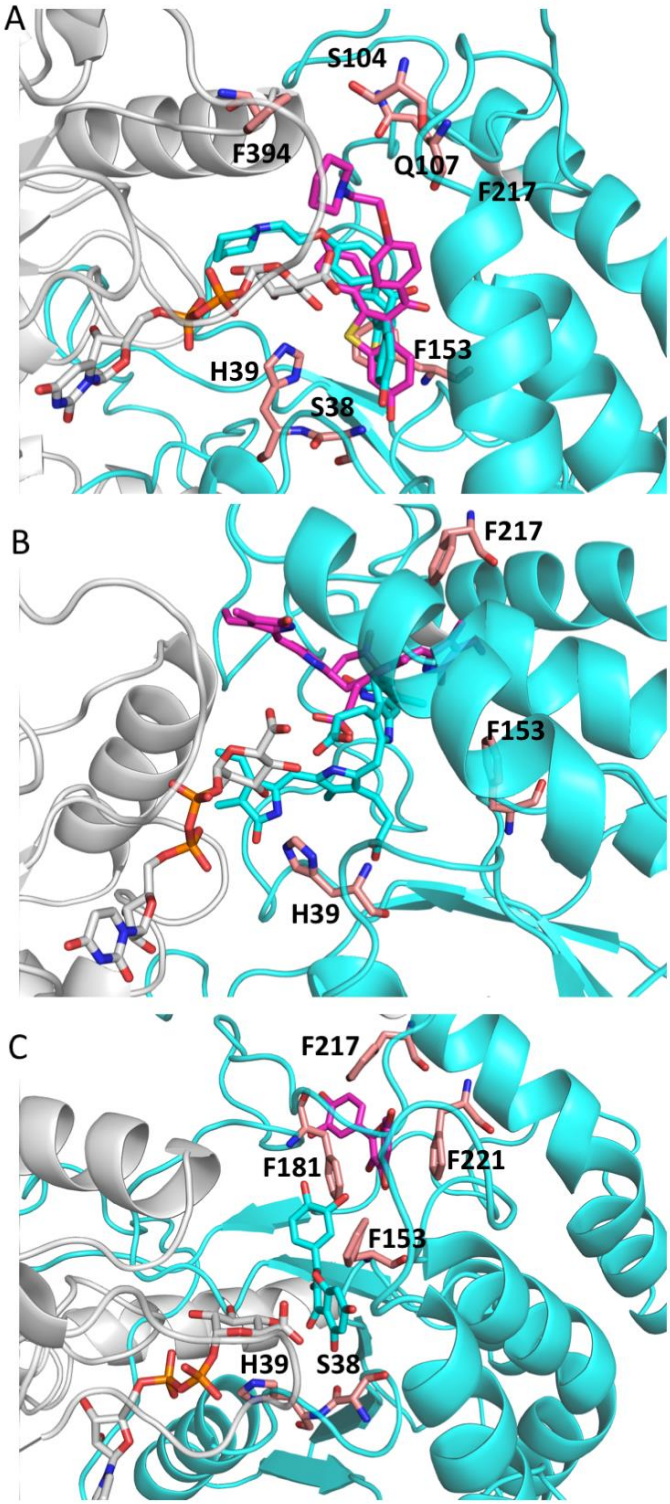

Figure S4. Binding positions of three substrates of UGT1A1 in MD47. The cofactor-binding domain and UDPGA are shown in white, the substrate-binding domain and docked substrates before MD simulations are in cyan. Key residues after 50-ns MD simulations are shown in salmon sticks. A) The binding position of raloxifene after 50-ns MD (in violet stick). B) The binding position of bilirubin after 50-ns MD (in violet stick). C) The binding position of quercetin after 50-ns MD (in violet stick).

ChEMBL IDs of the actives of the training set

CHEMBL108829, CHEMBL121626, CHEMBL14126, CHEMBL14130, CHEMBL16217, CHEMBL163, CHEMBL1908020, CHEMBL1908073, CHEMBL1922235, CHEMBL194014, CHEMBL1951575, CHEMBL281202, CHEMBL294009, CHEMBL325372, CHEMBL3527504, CHEMBL3527578, CHEMBL3527579, CHEMBL42710, CHEMBL4483026, CHEMBL4526403, CHEMBL46740, CHEMBL501680, CHEMBL51628, CHEMBL55814, CHEMBL63323, CHEMBL1064, CHEMBL1164729, CHEMBL117, CHEMBL1272, CHEMBL1289926, CHEMBL129, CHEMBL1401508, CHEMBL1487, CHEMBL157101, CHEMBL1624, CHEMBL1743300, CHEMBL191, CHEMBL1946170, CHEMBL226345, CHEMBL2364638, CHEMBL255863, CHEMBL374975, CHEMBL3813873, CHEMBL4066936, CHEMBL429, CHEMBL435298, CHEMBL44793, CHEMBL457, CHEMBL475251, CHEMBL477772, CHEMBL502835, CHEMBL553, CHEMBL80, CHEMBL1329, CHEMBL3222137, CHEMBL3527329, CHEMBL487805, CHEMBL55415, CHEMBL691, CHEMBL370963, CHEMBL2403238, CHEMBL10359050

ChEMBL IDs of the actives of the test set

CHEMBL1096146, CHEMBL1412489, CHEMBL1477, CHEMBL152295, CHEMBL154, CHEMBL186179, CHEMBL25308, CHEMBL254316, CHEMBL277346, CHEMBL289277, CHEMBL298398, CHEMBL307145, CHEMBL3187723, CHEMBL32749, CHEMBL338604, CHEMBL4229237, CHEMBL50, CHEMBL71851, CHEMBL723, CHEMBL73930, CHEMBL76398, CHEMBL81, CHEMBL8145, CHEMBL837, CHEMBL898, CHEMBL9352, CHEMBL956

L. Supporting Information – ABCG2

ABCG2/BCRP transport mechanism
revealed through kinetically excited targeted molecular dynamics simulations

B. Dudas[a,b], X. Decleves[c,d], S. Cisternino[c,e], D. Perahia[b,*], M. A. Miteva[a,*]

[a]Inserm U1268 MCTR, CiTCoM UMR 8038 CNRS - Université Paris Cité, Paris, France
[b]Laboratoire de Biologie et Pharmacologie Appliquée, Ecole Normale Supérieure Paris-Saclay, Gif-sur-Yvette, France
[c]Inserm UMRS 1144, Optimisation Thérapeutique en Neuropsychopharmacologie - Université Paris Cité, Paris, France
[d]Biologie du Médicament et Toxicologie, Assistance Publique Hôpitaux de Paris, AP-HP, Hôpital Universitaire Cochin, Paris, France
[e]Service Pharmacie, Assistance Publique Hôpitaux de Paris, AP-HP, Hôpital Universitaire Necker-Enfants Malades, Paris, France

[*]corresponding authors: maria.mitev@inserm.fr, david.perahia@ens-paris-saclay.fr

Published in Computational Structural Biotechnology Journal on 2022 Jul 29
doi: 10.1016/j.csbj.2022.07.035

State	Name of System	Ligands			Initial PDB ID	Details on System Assembly
		E_1S	ATP-Mg^{2+}	ADP		
IFS	apo IFS				6HCO	E_1S removed from cavity 1
	E_1S bound IFS	✓				
	E_1S & ATP-Mg^{2+} bound IFS *	✓	✓			ATP-Mg^{2+} positions taken from 6HBU after overlapping on residues 80-94
OFS	ATP-Mg^{2+} bound OFS		✓		6HBU	
	ADP bound OFS			✓		ATP γ-phosphates cleaved, Mg^{2+} ions removed
	nucleotide-free OFS					ATPs and Mg^{2+} ions removed

Table S1: Details on the different systems used in the MD simulations and NMA.
(*) An E_1S and ATP-Mg^{2+}-bound IFS was also constructed using the structure PDB 7OJ8.

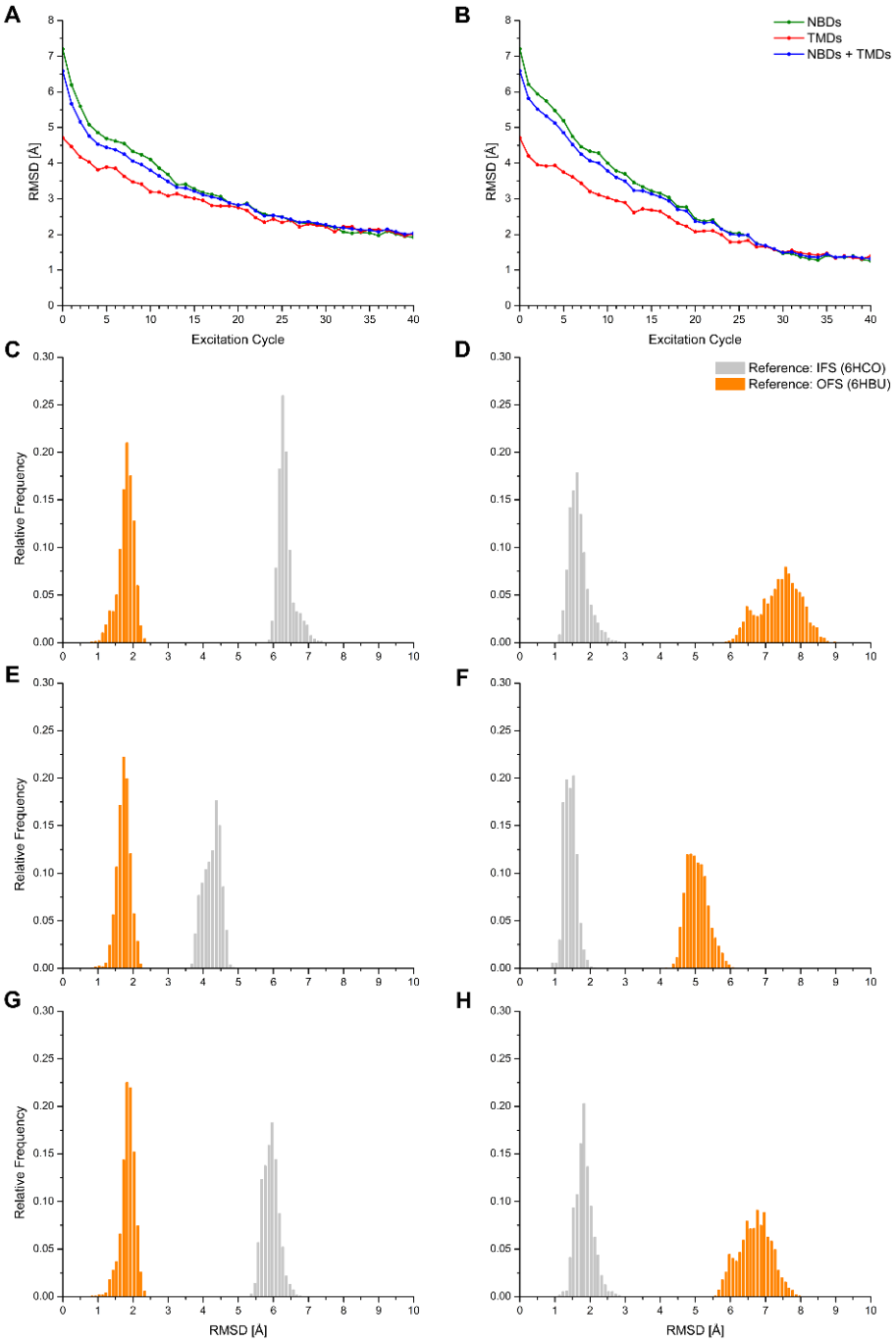

*Figure S1: Root Mean Square Deviation (RMSD) from experimental reference structures. Evolution of the backbone RMSD during the ketMD simulation **(A)** of transition 1 and **(B)** of transition 2 with respect to their corresponding experimental target structures (PDB 6HBU (OFS) for transition 1 and PDB 6HCO (IFS) for transition 2); RMSD calculated on the NBD dimer is in olive, on the TMD dimer in red, and on the whole transporter in blue. RMSD distribution of the three 100-ns-long classical MD generated conformations, calculated on the NBD dimer for **(C)** the OFS and **(D)** the IFS, calculated on the TMD dimer for the **(E)** OFS and **(F)** the IFS, and calculated on the whole transporter for the **(G)** OFS and the **(H)** IFS MD conformations. RMSD with respect to PDB 6HCO (IFS) is in light gray, with respect to PDB 6HBU (OFS) in orange.*

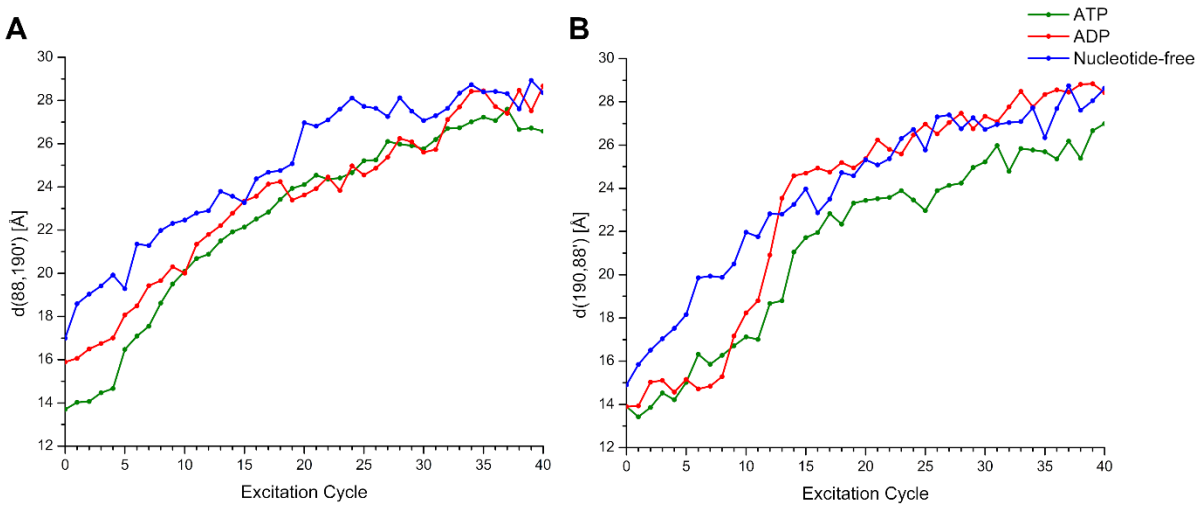

*Figure S2: The evolution of the distances corresponding to the two catalytic ATP-binding sites, symmetrically between the P-loop of one monomer and the signature sequence of the other, represented by the distance **(A)** between the Cα atoms of residues S88 and E190' and **(B)** between the Cα atoms of residues E190 and S88' during the ketMD simulations of transition 2. The ketMD simulation in the presence of ATP-Mg^{2+} is shown in olive, in the presence of ADP in red, and in the absence of bound nucleotides in blue. The protein initial conformation of all three cases correspond to the experimental structure PDB 6HBU.*

PDB	Ligand*	Nucleotide	State	Reference
5NJ3	-	-	IFS	[1]
6HIJ	MZ29 (*I*)	-	IFS	[2]
6FFC	MZ29 (*I*)	-	IFS	
6FEQ	FKo143 (*I*)	-	IFS	
6ETI	MZ29 (*I*)	-	IFS	
6HCO	estrone 3-sulfate (*S*)	-	IFS	[3]
6HBU	-	ATP-Mg^{2+}	OFS	
6HZM	-	ATP-Mg^{2+}	OFS	
6VXF	-	-	IFS/OFS**	[4]
6VXJ	SN38 (*S*)	-	IFS	
6VXI	mitoxantrone (*S*)	-	IFS	
6VXH	imatinib (*S*)	-	IFS	
7NFD	mitoxantrone (*S*)	-	IFS	[5]
7NEZ	topotecan (*S*)	-	IFS	
7NEQ	tariquidar (*S*)	-	IFS	
7OJI	topotecan (*S*)	ATP	IFS/OFS***	[6]
7OJH	topotecan (*S*)	ATP	IFS	
7OJ8	estrone 3-sulfate (*S*)	ATP	IFS/OFS***	

Table S2: Available ABCG2 experimental structures.
(*) S = substrate, I = inhibitor
(**) 'apo-closed state', the arrangement of TM helices more closely resembles that seen in the outward facing ATP bound state, whereas the lack of NBD dimerization more closely resembles that of the inward facing state [4]
(***) 'turnover-2 state', semi-closed NBDs and an almost fully occluded substrate cavity [6]

1. Taylor, N.M.I., et al. (2017) *Structure of the human multidrug transporter ABCG2.* Nature. 546:504-509.
2. Jackson, S.M., et al. (2018) *Structural basis of small-molecule inhibition of human multidrug transporter ABCG2.* Nat Struct Mol Biol. 25:333-340.
3. Manolaridis, I., et al. (2018) *Cryo-EM structures of a human ABCG2 mutant trapped in ATP-bound and substrate-bound states.* Nature. 563:426-430.
4. Orlando, B.J. and M. Liao (2020) *ABCG2 transports anticancer drugs via a closed-to-open switch.* Nat Commun. 11:2264.
5. Kowal, J., et al. (2021) *Structural Basis of Drug Recognition by the Multidrug Transporter ABCG2.* J Mol Biol. 433:166980.
6. Yu, Q., et al. (2021) *Structures of ABCG2 under turnover conditions reveal a key step in the drug transport mechanism.* Nat Commun. 12:4376.

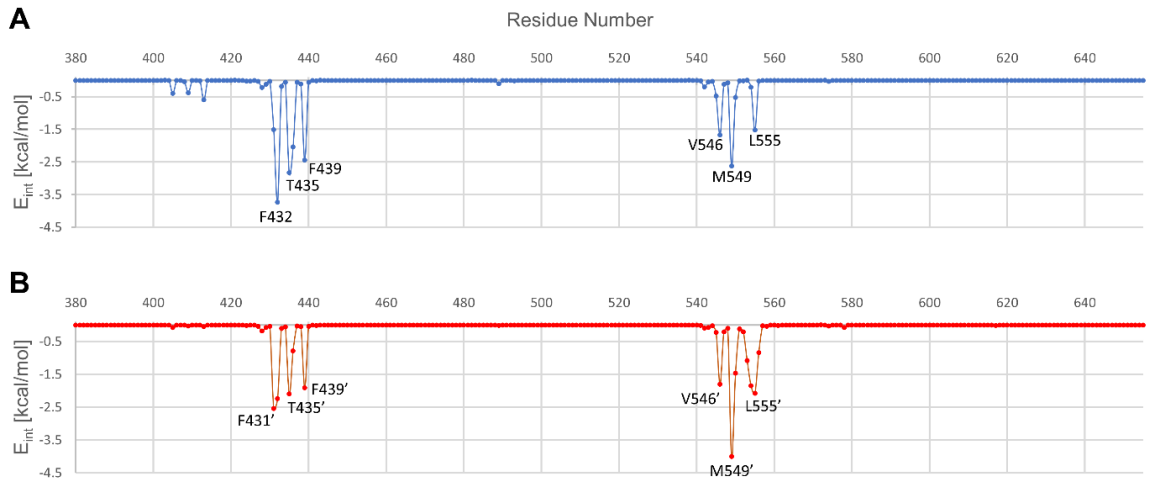

*Figure S3: Interaction Energy between the residues of the transporter and the E_1S substrate when trapped in the pocket-like formation between the F439 valve and the leucine plug, based on the classical MD simulations starting from the ketMD generated transient conformations for (**A**) one monomer and (**B**) the other.*

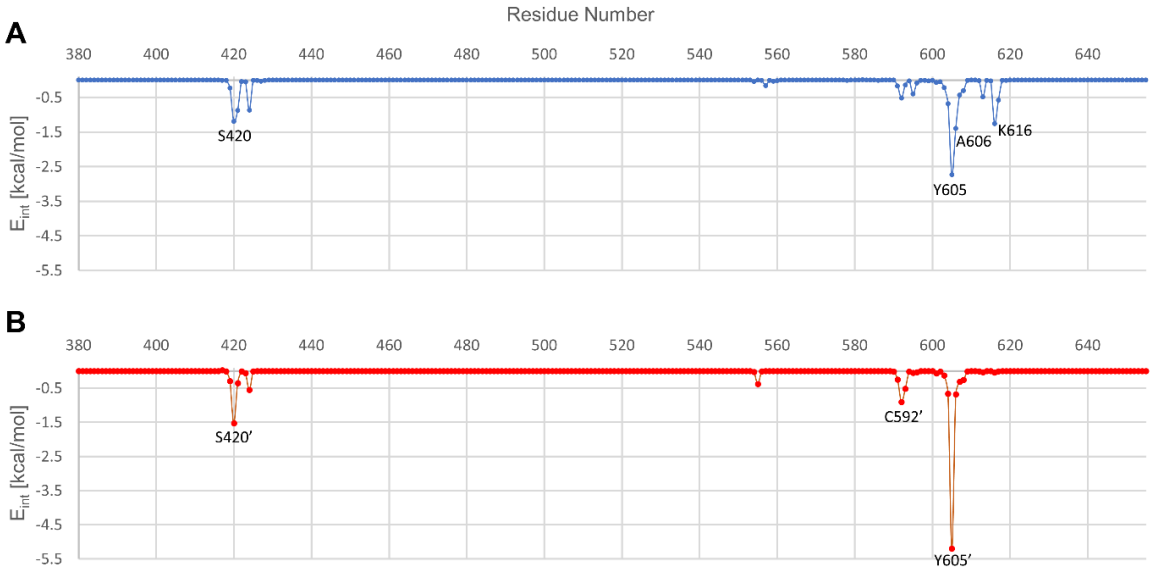

*Figure S4: Interaction Energy between the residues of the transporter and the E_1S substrate in cavity 2, based on the classical MD simulations starting from the ketMD generated transient conformations for (**A**) one monomer and (**B**) the other.*

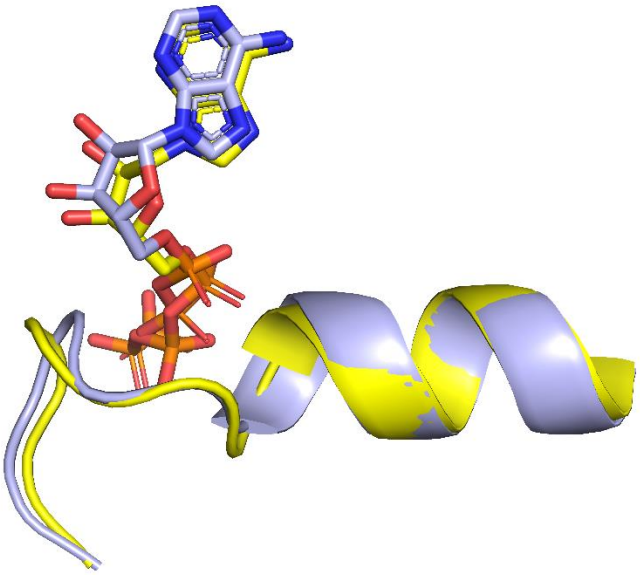

Figure S5: The superposition of residues 80-94 of the PDB 7OJ8 structure (in light blue) with bound ATP and the model (in yellow) that was constructed using the IFS structure (PDB 6HCO) with the nucleotide from the OFS structure (PDB 6HBU). The ATPs are in licorice representation.

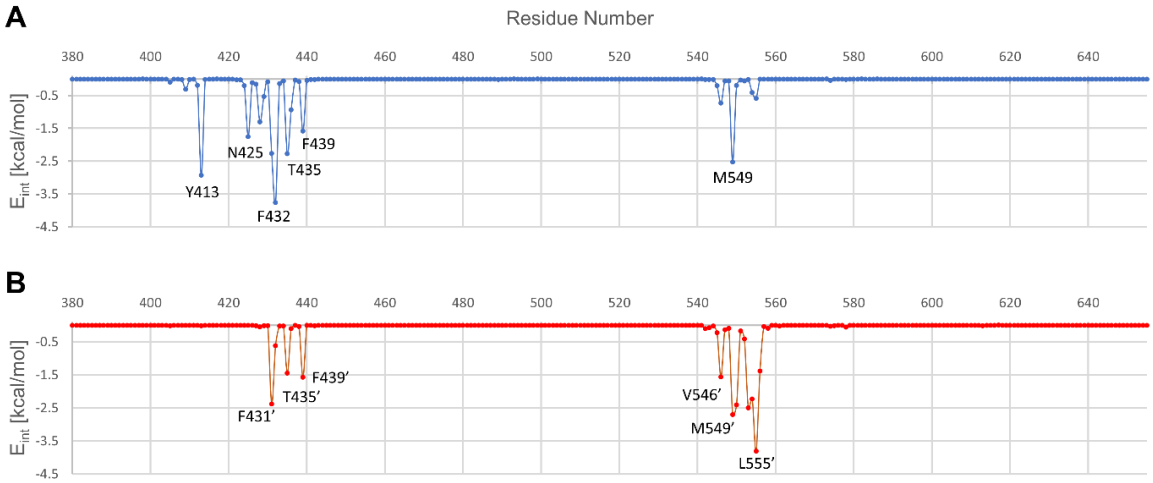

*Figure S6: Interaction Energy between the residues of the transporter and the E_1S substrate when trapped in the pocket-like formation between the F439 valve and the leucine plug for (**A**) one monomer and (**B**) the other. The interaction energies are based on the classical MD simulations that were performed starting from the transient conformations of the ketMD simulation using the initial structure PDB 7OJ8.*

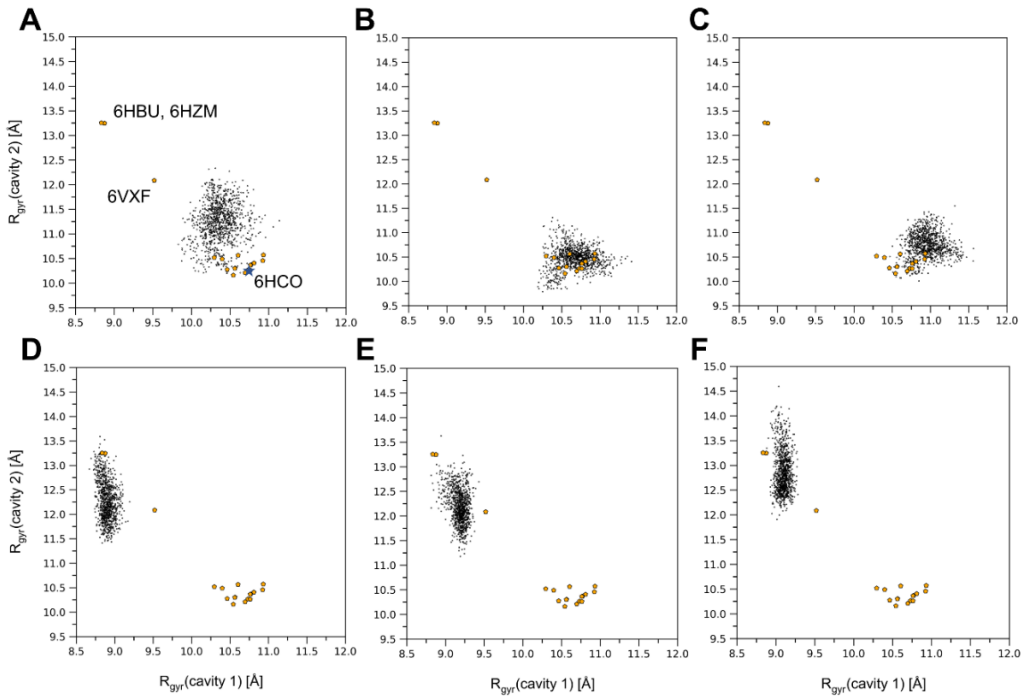

*Figure S7: Changes in the substrate-binding cavities represented by the radius of gyration (R_{gyr}) of the helical structures bordering the cavities during classical MD simulations. Conformations of (**A**) the apo IFS, (**B**) the substrate-bound IFS, and (**C**) the substrate- and ATP-Mg^{2+}-bound IFS transporter, and (**D**) the ATP-Mg^{2+}-bound OFS, (**E**) the ADP-bound OFS, and (**F**) the nucleotide-free OFS simulations. Available experimental structures are marked with orange pentagons.*

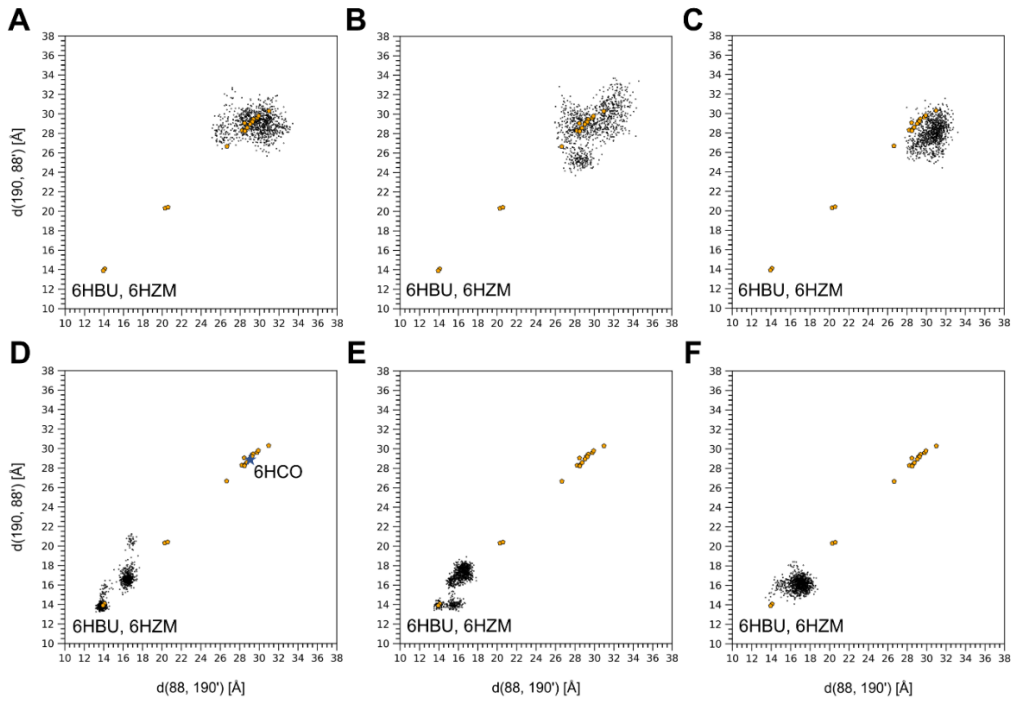

*Figure S8: The openness at the catalytic ATP-binding site of the MD generated conformations, represented by the distance between the CA atoms of residues S88 of one monomer and E190 of the other. Simulations on (**A**) the apo IFS, (**B**) the substrate bound IFS, and (**C**) the substrate and ATP-Mg^{2+} bound IFS transporter, (**D**) the ATP-Mg^{2+} bound OFS, (**E**) the ADP bound OFS, and (**F**) the nucleotide-free OFS ABCG2. Available experimental structures are marked with orange pentagons.*

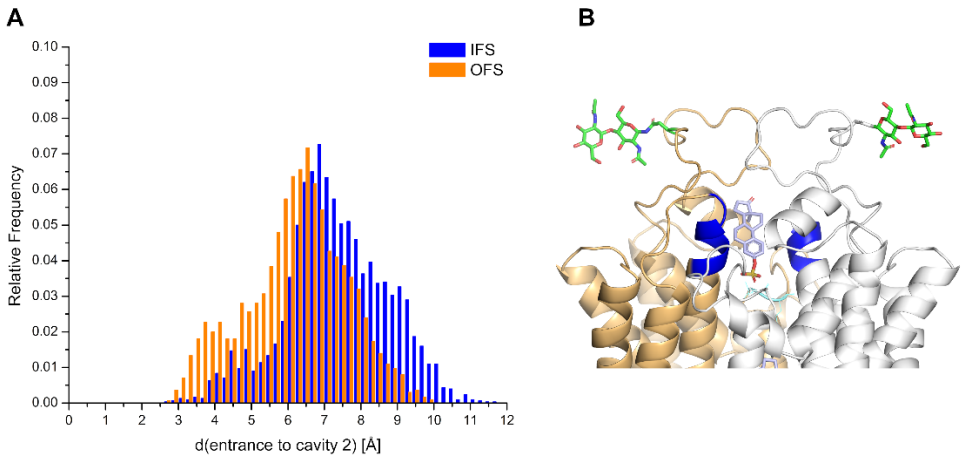

Figure S9: The tightest part of cavity 2 after passing through the leucine gate. ***(A)*** *The population distribution of the minimum distance between the upper tip of TM3 (residues 420-425) and TM3' (residues 420'-425') heavy atoms. The distribution of the IFS is shown in blue, and the OFS classical MD simulations in orange.* ***(B)*** *The regions at the upper tips of TM3 and TM3' which were used to calculate the minimum distance in panel A.*

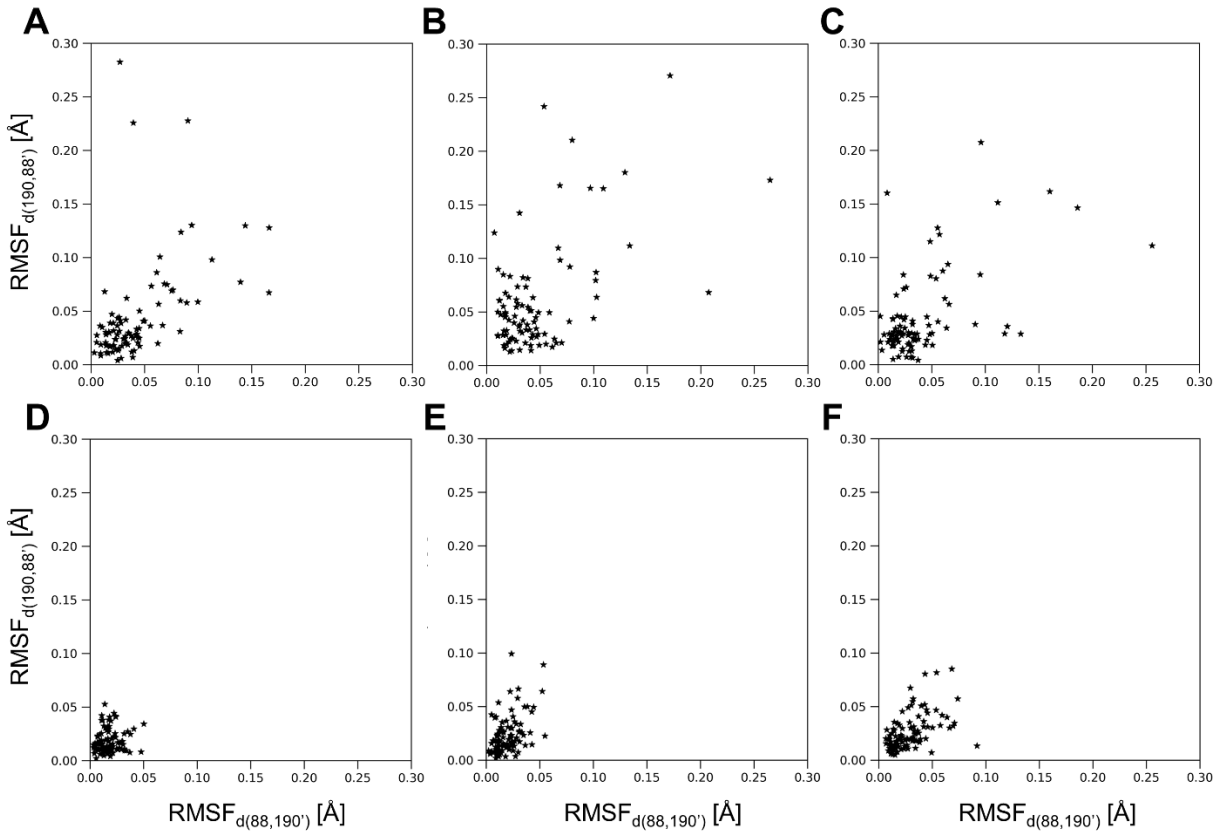

*Figure S10: Amplitudes of NM contributions to the fluctuations of the distances between the CA atoms of residues S88 of one monomer and E190 of the other at 300 K, in the case of (**A**) the apo IFS, (**B**) the substrate-bound IFS, and (**C**) the substrate- and ATP-Mg^{2+}-bound IFS transporter, (**D**) the ATP-Mg^{2+}-bound OFS, (**E**) the ADP-bound OFS, and (**F**) the nucleotide-free OFS ABCG2.*

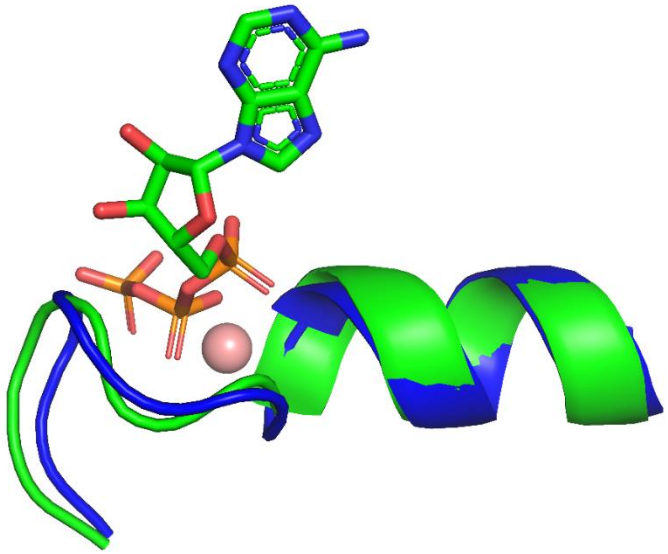

Figure S11: The superposition of residues 80-94 of the IFS (6HCO, blue) and the OFS structure (6HBU, green). The ATP is in licorice, the Mg^{2+} ion in sphere representation.

Domain	Region	Residue(s)
NBD	A-loop	F52
	P-loop (Walker A)	80-88
	Q-loop	Q126
	Signature sequence	186-193
	Walker B	206-211
	D-loop	L216, D217
	H-loop	H243
TMD	Elbow helix (TM1a)	373-391
	TM1b	393-413
	TM2	421-448
	CpH	451-461
	TM3	466-496
	TM4	503-528
	TM5a	535-552
	TM5b	565-571
	TM5c	573-585
	TM6a	610-617
	TM6b	623-650

Table S3: Structural elements of ABCG2.