IN SILICO APPROACHES TO UNDERSTAND AND PREDICT THE LIGAND-BINDING INTERACTIONS OF THE HUMAN P-GLYCOPROTEIN (ABCB1)

Liadys Mora Lagares

Doctoral Dissertation
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

Supervisor: Prof. Marjana Novič, National Institute of Chemistry, Ljubljana, Slovenia
Co-Supervisor: Prof. Emilio Benfenati, Istituto di ricerche farmacologiche "Mario Negri", Milan, Italy

Evaluation Board:
Prof. Dr. Veronika Stoka, Chair, IPS and Jožef Stefan Institute, Ljubljana, Slovenia
Prof. Dr. Janez Mavri, Member, National Institute of Chemistry, Ljubljana, Slovenia
Prof. Dr. Maxime Culot, Member, Faculty of Sciences, University of Artois, Lens, France

MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA
JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL

Liadys Mora Lagares

IN SILICO APPROACHES TO UNDERSTAND AND PREDICT THE LIGAND-BINDING INTERACTIONS OF THE HUMAN P-GLYCOPROTEIN (ABCB1)

Doctoral Dissertation

IN SILICO PRISTOPI ZA RAZUMEVANJE IN NAPOVEDOVANJE INTERAKCIJ ZA VEZAVO LIGANDOV NA ČLOVEŠKI MEMBRANSKI TRANSPORTER P-GLIKOPROTEIN (ABCB1)

Doktorska disertacija

Supervisor: Prof. Marjana Novič

Co-Supervisor: Prof. Emilio Benfenati

Ljubljana, Slovenia, July 2021

To the memory of my beloved grandfather
José Lagares

Acknowledgments

My deepest gratitude goes first to my supervisor, Prof. Dr. Marjana Novič, who gave me the opportunity to join her group at the National Institute of Chemistry and whose expertise was invaluable through my graduate studies. Her insightful feedback pushed me to sharpen my thinking and brought my work to a higher level. I truly appreciate all your support and encouragement over the last 4 years.

My appreciation also extends to my laboratory colleagues whose personal generosity helped make my time at the Institute enjoyable.

I would like to thank the European project “in3” funded by the EU's Marie Skłodowska--Curie Action - Innovative Training Network (MSCA-ITN), of which my PhD makes part. I would especially like to thank Prof. Dr. Paul Jennings, in3 coordinator, for his kind support and enthusiasm during the in3 journey. The in3 project has taught me lessons that have made me a better scientist and person, it has literally changed my life. It has been my pleasure to learn among such an amazing group of scientists with the best human values. I would also like to thank all my in3 colleagues. Together we have pushed the boundaries of knowledge and steered through the rocks of academia. Thank you all for this interdisciplinary experience.

I would also like to express my gratitude to all the collaborators I had the possibility to work with.

A very special thanks goes to Prof. Dr. Yunierkis Perez Castillo from Universidad de Las Américas, Quito, Ecuador, who is the most considerate and helpful person I have ever met. Thank you for all the support and availability during the molecular dynamics simulations, without which this dissertation would not have been possible.

My greatest thanks go to my family, whose value to me only grows with age. Thank you for all the love and support that allowed me to become the person I am right now. You guys have managed to be there for me from afar when I needed it.

And finally, I am immensely grateful to Jernej, who showed me the importance of self-confidence and who was patiently my pillar of support, holding me through the hard times of this PhD journey. From the bottom of my heart, I thank you for your unconditional love.

Abstract

The P-glycoprotein (P-gp) is responsible for the elimination of a wide variety of substances from the cell, thus it is thought to play an important role in detoxification functions. The P-gp also affects the ADMET (absorption, distribution, metabolism, excretion, and toxicity in pharmacokinetics) properties of drugs. It is involved in adverse drug-drug interactions and its overexpression is thought to be responsible for the presence of multidrug resistance (MDR) in cancer cells, which is considered to be the main reason for the failure of cancer therapies. Therefore, *in silico* approaches to understand the ligand binding interactions of the human p-glycoprotein (*h*P-gp) leading to early identification of P-gp active compounds are of great interest for drug development and toxicity assessment.

This dissertation summarises in three different studies how ligand- and structure-based methods can be used to solve the problem of elucidating the ligand–P-gp interactions and the transport mechanism, and to provide a rapid and accurate prediction of potential new P-gp ligands. In the first study, the ligand-based approach is described by developing a classification model to predict the activity of P-gp. The developed multiclass classifier showed a good classification performance and is a useful tool for saving significant time in the drug development pipeline, as it provides a rapid initial screening step for selecting predicted molecules that would interact with P-gp as inhibitors or substrates.

The second and third studies illustrate the structure-based approach of the target protein. The second study describes the construction of a homology model of the *hP*-gp and the use of molecular docking to identify binding modes of different compounds, either active (substrates and inhibitors) or non-active compounds. The study of the ligand–*h*P-gp complexes provided considerable insight into the drug binding mode for the set of investigated compounds. Different modes of interaction for different classes of compounds were revealed and consistency between the predicted interactions and available experimental data was demonstrated.

Finally, a series of molecular dynamics simulations of *h*P-gp in an explicit membrane and water environment were performed to investigate the effects of binding different compounds on the conformational dynamics of P-gp. The results showed a significant difference in the behaviour of P-gp in the presence of an active or non-active compound within the binding pocket. Different motion patterns were identified which could be correlated with the conformational changes leading to the activation of the translocation mechanism.

Povzetek

P-glikoprotein (P-gp) je odgovoren za izločanje najrazličnejših snovi iz celice, zato naj bi imel pomembno vlogo pri razstrupljevanju. P-gp vpliva tudi na lastnosti zdravil (toksičnost v farmakokinetiki – ADMET), vpleten je v škodljive interakcije z zdravili, njegova prekomerna ekspresija pa naj bi bila odgovorna za prisotnost odpornosti na več zdravil (MDR) v rakavih celicah, kar velja za glavni razlog za neuspeh terapij raka. Zato so pristopi *in silico* za razumevanje interakcij vezave ligandov človeškega p-glikoproteina (*h*P–gp), ki vodijo k zgodnji identifikaciji P-gp-aktivnih spojin, zelo zanimivi za razvoj zdravil in oceno toksičnosti.

Ta disertacija v treh različnih študijah povzema, kako se metode, ki temeljijo na ligandu in strukturi, lahko uporabljajo za pojasnjevanje interakcij liganda P-gp in transportnega mehanizma ter za hitro in natančno napoved potencialnih novih ligandov P-gp. V prvi študiji je opisan pristop, ki temelji na podatkih o seriji ligandov in omogoča razvoj klasifikacijskega modela za napovedovanje aktivnosti P-gp. Napovedni model je pokazal dobro učinkovitost pri klasifikaciji spojin v več razredov in predstavlja uporabno orodje za pomemben prihranek časa v linearnem zaporedju specializiranih modulov, ki se uporabljajo za razvoj zdravilnih učinkovin. Večrazredni klasifikator zagotavlja hiter začetni presejalni korak za izbiro predvidenih molekul, ki bi medsebojno vplivale na P-gp kot zaviralci ali substrati.

Druga in tretja študija ponazarjata pristop, ki temelji na strukturi tarčnega proteina. Druga študija opisuje izgradnjo homolognega modela *h*P-gp in uporabo metode molekulskega sidranja za identifikacijo načina vezave različnih spojin, tako aktivnih (substrati in inhibitorji) kot tudi neaktivnih spojin. Študija kompleksov ligand–*h*P-gp je pomembno prispevala k pojasnjevanju načina vezave zdravilnih učinkovin, upoštevajoč preiskovane spojine izbranega podatkovnega niza. Izkazalo se je, da različni razredi spojin kažejo različne načine interakcij, ki so skladni z razpoložljivimi eksperimentalnimi podatki.

Na koncu je prikazan vpliv vezave različnih spojin na konformacijsko dinamiko P-gp s pomočjo simulacij molekulske dinamike *h*P-gp v okolju simulirane lipidne membrane in vode. Rezultati so pokazali pomembno razliko v obnašanju P-gp v prisotnosti aktivne ali neaktivne spojine v vezavnem žepu. Ugotovljeni so bili različni vzorci gibanja, ki bi jih lahko povezali s konformacijskimi spremembami, ki vodijo k aktiviranju mehanizma prehoda spojin preko celične membrane.

Contents

List of Figures

List of Tables

Abbreviations

ABC	... ATP binding cassette
ADME	... Absorption, distribution, metabolism, and elimination
ADMET	... Absorption, distribution, metabolism, elimination, and toxicity
AMI	... Amiodarone
APD	... Pamidronate
ATP	... Adenosine triphosphate
BBB	... Blood–brain barrier
BUS	... Busulfan
CADD	... Computer-aided drug design
CAR	... Constitutive androstane receptor
CAR	... Carvedilol
CPANN	... Counter Propagation Artificial Neural Network
CPU	... Central processing unit
cryo-EM	... Cryogenic electron microscopy
CSA	... Cyclosporine A
DME	... Drug metabolizing enzymes
DOX	... Doxorubicin
ECL	... Extracellular loop
GEN	... Gentamicin
HBA	... Hydrogen bond acceptor
HBD	... Hydrogen bond donor
*h*P-gp	... Human P-glycoprotein
ICL	... Intracellular loop
MD	... Molecular dynamics simulations
MDR	... Multidrug resistance
MM/PBSA	... Molecular Mechanics/Poisson-Boltzmann Surface Area
*m*P-gp	... Mouse P-glycoprotein
NBD	... Nucleotide binding domain
PBC	... Periodic boundary conditions
PC	... Principal component
PCA	... Principal component analysis
PDB	... Protein Data Bank
P-gp	... P-glycoprotein
Pi	... Inorganic phosphate
PME	... Particle-Mesh Ewald
POPC	... 1-palmitoyl-2-oleoyl-phosphatidylcholine
PQT	... Paraquat
PXR	... Pregnane xenobiotic receptor
RMSD	... Root-mean-square deviation
RMSF	... Root-mean-square fluctuation
SASA	... Solvent accessible surface area

SLC	... Solute carrier
TE	... Test
TM	... Transmembrane
TMD	... Transmembrane domain
TR	... Training
V	... Validation
VPA	... Valproic acid
3D	... Three-dimensional

Chapter 1

Introduction

In recent years, *in silico* tools have become a key component in the drug development pipeline (Ou-Yang *et al.*, 2012; Xiang *et al.*, 2012). Trial-and-error approaches to new drugs discovery are both time-consuming and expensive, so computer-based tools are currently used in almost every step of this process, especially in the early stages. On the other hand, the P-glycoprotein (P-gp), an important membrane transporter expressed by the *mdr1* gene and belonging to the ATP binding cassette superfamily (ABC), has been extensively studied; nevertheless, many aspects of its structure–function mechanism remain unresolved. P-gp has a major role in drug absorption and disposition and is involved in the regulation of toxicity and failure of cancer therapies due to multidrug resistance (MDR). Therefore, it seems essential to search for new *in silico* approaches that will allow a better understanding of the P-gp transport mechanism. Efficient *in silico* screening tools that can be easily used in drug discovery and toxicological evaluation are needed to provide a rapid and cost-effective platform for the identification of potential P-gp ligands.

Identification of potential substrates and inhibitors of P-gp is of great importance in overcoming MDR, either by screening for new P-gp inhibitors that might prevent the transporter from expelling drugs out of the cell, or by early identification of P-gp substrates and subsequent outdesign of substrate properties. It is noticeable that although P-gp has been known for more than 30 years and despite numerous experiments, improved and selective P-gp inhibitors have not yet been developed. This can probably be explained by the polyspecificity of the transporter and the lack of high-resolution structural information, which led to limited information on the molecular basis of ligand–transporter interactions.

The *in silico* approaches available in the field of molecular modelling can be divided into two main categories: ligand-based and structure-based approaches; the first category includes Quantitative Structure-Activity Relationships, (QSAR)-modelling (Favia, 2011; Mavromoustakos *et al.*, 2011). Regardless of the success of ligand-based methods for drug discovery, there are still some obstacles; one of their main limitations is that they do not consider the receptor structure during the modelling process. In this sense, structure-based tools allow the analysis of the structure of the molecular target and the ligand–target interactions. Among these tools, molecular docking and molecular dynamics simulations are widely used. These modelling techniques have been used to identify potential drug candidates for a variety of therapeutic categories. However, their application to the study of P-gp has been limited by the use of homology models prior to the availability of the human structure. The high flexibility and polyspecificity of P-gp make the use of structure-based approaches rather difficult, but their application is essential to understand the mechanism of ligand recognition and binding.

In general, work with membrane proteins presents many challenges, ranging from the lack of high-resolution 3D structures to unresolved transport cycles. Therefore, with the resolution of the first crystal structure of mouse P-gp (*m*P-gp) (PDB ID: 3G5U) (Aller *et al.*, 2009) in 2009, the application of structure-based design to this protein became more accessible and promising for the development of this field. Furthermore, in 2019 the cryo-electron microscopy (cryoEM) structure of human P-gp (*hP*-gp) was solved (PDB ID: 6QEX) (Alam *et al.*, 2019), which raised much hope for a better and more efficient development in the study of this protein.

In this context, new approaches are needed to gain a better understanding of the membrane transporter. One of these approaches is the integration of both ligand- and structure-based models, as their integration could increase the success of the virtual screening process in the search for potential ligands of P-gp. In this way, one of the main contributions of this research is the development of our own P-gp classification model. Unlike currently available classifiers, our model is able to differentiate between substrates and inhibitors, an important detail that would extend the use of the multiclass classifier for various purposes. This fast-screening tool would be essential for the initial steps of evaluating the molecules of interest before performing molecular modelling methods such as molecular docking and molecular dynamics simulations, which are more demanding and time-consuming, saving significant time in the pipeline. After the initial fast-screening only those molecules showing a particular response would be subjected to further detailed studies, including a final experimental test of their interactions with the P-gp.

From the perspective of structure-based modelling, we obtained a detailed description of the dynamics of *h*P-gp at the atomistic level, which was previously limited to studies based on the structure of *m*P-gp or homology models of *h*P-gp. Although ligand- and structure-based modelling techniques are widely used in drug discovery and development in both academia and industry, one of our contributions is to combine ligand- and structure-based modelling techniques as an effective approach to study and elucidate the P-glycoprotein ligand recognition and transport mechanism, thereby facilitating the design and discovery of new potential P-glycoprotein ligands.

1.1 Thesis Outline

This dissertation guides you through the challenging task of understanding the ligand binding interactions of the human membrane transporter P-glycoprotein and how this information could be used to provide the community with useful and validated *in silico* models to estimate the likelihood of a new compound to interact with it.

In the introductory Chapter 1, an overview of the purpose of the thesis is presented as well as the structure of the dissertation, the workflow, and the authors contributions.

Chapter 2 presents the biomolecular background of the P-glycoprotein (ABCB1) and the multidrug resistance (MDR) phenomenon. This chapter also reports the most common hypothesis about the transport mechanism of the membrane transporter and a brief depiction of the classification of P-gp interacting compounds depending on the nature of the interaction.

Chapter 3 illustrates the ligand-based modelling approach, which consists of a detailed description of the construction of a classification model that provides a qualitative prediction of P-glycoprotein inhibition/substrate activity. The model was constructed using the Counter Propagation Artificial Neural Network (CPANN) and its output represents the probability that the predicted compound belongs to the class of inhibitors, substrates, or non-active compounds. The P-gp activity model developed as

part of this research can separate substrates from inhibitors unlike other classifiers available until now and has been successfully implemented within the online platform VEGAHUB (Benfenati *et al.*, 2013), which is freely available to the public at https://www.vegahub.eu/portfolio-item/vega-qsar/.

Considering that the ligand-based modelling strategy does not take into account the influence of the protein structure during the modelling procedure and that this is an important factor in modelling any biochemical process, a comprehensive structure-based study of ligand–P-glycoprotein interactions is presented in Chapter 4. The development of a homology model of *h*P-gp and the results of molecular docking calculations on a set of thirteen compounds are presented, including some ligands of P-gp and non-interacting compounds. A good agreement was found between the modelling results and experimental evidence previously reported for this transporter.

In Chapter 5, a 500 ns molecular dynamics simulation of the human P-glycoprotein in complex with various compounds is used to analyse the molecular basis of the ligand–P-glycoprotein interactions. Changes in the conformation of the protein and the volume of the binding pocket are analysed along the entire trajectory along with MM/PBSA energy calculations to predict the stability of ligand–protein complexes of this membrane transporter. This chapter also analyses the relationship between the motion patterns of NBDs and ligand binding.

Chapter 6 summarizes the results of the research and a brief perspective on the future directions in this area of study is presented.

1.2 Workflow

The overall modelling strategy in our research focuses on a ligand-based approach, followed by structure-based approaches (Moro *et al.*, 2007; Prathipati *et al.*, 2007). In summary, we first aim to predict the interaction profile of small molecules towards the P-glycoprotein (P-gp) using a machine learning approach (Ma *et al.*, 2009). We then aim to analyse the ligand–transporter interactions at the atomistic level using molecular docking calculations (Dias *et al.*, 2008) and molecular dynamics simulations (Karplus *et al.*, 2002).

Until 2019, when the cryo-EM structure of *h*P-gp was resolved (PDB ID: 6QEX) (Alam *et al.*, 2019), structure-based studies of P-gp were limited to the use of homology models to study ligand–transporter interactions. At the time of the project initiation, no experimental structure of *h*P-gp was available in the Protein Data Bank (PDB, www.rcsb.org). Therefore, in the first part of the structure-based approach, we also made use of the homology modelling technique. The following figure illustrates the whole workflow of the combined strategy:

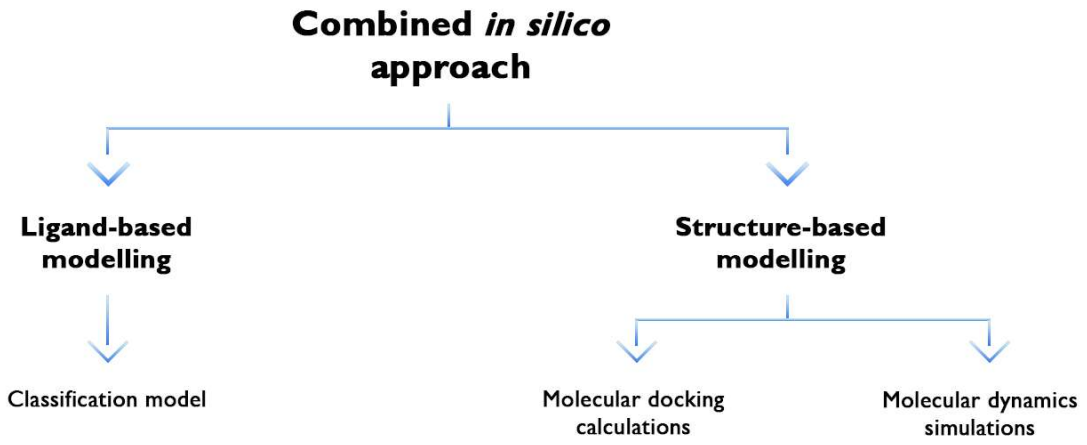


Figure 1.1: Workflow of the study.

1.3 Scientific Production and Author Contributions

Part of the results reported here has been previously published in the following journal articles and conference papers:

1. Multiclass Classifier for P-Glycoprotein Substrates, Inhibitors, and Non-Active Compounds (Mora Lagares, Minovski, & Novič, 2019).

Table 1.1: Author contributions.

Author	Contributions
L.M.L. (candidate)	Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing—original draft
N.M.	Conceptualization, Writing—review & editing
M.N.	Conceptualization, Supervision, Writing—review & editing

2. Homology Modelling of the Human P-glycoprotein (ABCB1) and Insights into Ligand Binding through Molecular Docking Studies (Mora Lagares *et al.*, 2020).

Table 1.2: Author contributions.

Author	Contributions
L.M.L. (candidate)	Conceptualization, Formal analysis, Investigation, Methodology, Validation, Writing—original draft
N.M.	Supervision, Writing—review & editing
A.Y.C.A.	Methodology
E.B.	Supervision, Writing—review & editing
S.W.	Methodology
M.C.	Supervision, Writing—review & editing
F.G.	Writing—review & editing
M.N.	Supervision, Writing—review & editing

3. P-gp transport activity in connection to the efflux of toxicants or drugs (Mora Lagares, Minovski, Drgan*, et al.*, 2019).

Table 1.3: Author contributions.

Author	Contributions
L.M.L. (candidate)	Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing—original draft
N.M.	Conceptualization, Writing—review & editing
D.V.	Methodology
M.T.	Methodology
M.N.	Conceptualization, Supervision, Writing—review & editing

4. P-glycoprotein modelling: development of an *in silico* prediction model for substrates, inhibitors and non-interacting compounds (Mora Lagares *et al.*, 2018).

Table 1.4: Author contributions.

Author	Contributions
L.M.L. (candidate)	Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing—original draft
N.M.	Conceptualization, Writing—review & editing
M.N.	Conceptualization, Supervision, Writing—review & editing

Chapter 2

P-glycoprotein

2.1 Introduction

The recognition that membrane transporters play a key role in the absorption, distribution, metabolism and elimination (ADME) of drugs in many organisms, including humans, is critical for the discovery and development of new therapeutic agents (Vrbanac *et al.*, 2017). Membrane transporters pump a variety of nutrients, cofactors, and ions into the cell and mediate the efflux of cellular waste, environmental toxins, and xenobiotics out of the cell; they are also involved in the uptake and elimination of drugs from the cell, which can lead to drug resistance and significant drug–drug interactions. The transport mechanism of these proteins can be passive, facilitating the movement of molecules down concentration gradients into or out of the cell without the need of energy (ATP or reducing equivalents), or active, as many transporters pump compounds against their concentration gradients in a process that requires energy (Borst *et al.*, 2002; Hediger *et al.*, 2004).

There are two major superfamilies of transporters, the ABC (ATP binding cassette) and the SLC (solute carrier) transporter family. ABC transporters are a large and multifunctional family of structurally related membrane proteins located in the plasma membrane of cells or in the membrane of various cellular organelles. Most ABC proteins are active transporters, using the energy of ATP hydrolysis to carry various molecules across the membranes; some others use the ATP to form specific membrane channels. The ABC transporters are involved in various physiological processes, and some inherited diseases also appear to be associated with mutations in the genes that code for them (de Lange, 2007). There are forty-nine known genes that encode for ABC proteins, which can be divided into seven subclasses or families (ABCA to ABCG) (Borst *et al.*, 2002); P-glycoprotein (P-gp) and the cystic fibrosis transmembrane regulator (CFTR) are the best studied transporters in the ABC superfamily. The SLC superfamily includes facilitated transporters and ion-coupled secondary active transporters located in various cell membranes. Forty-three SLC families with approximately 300 transporters have been identified in the human genome (Hediger *et al.*, 2004).

P-glycoprotein, also known as multidrug resistance 1 (MDR1), encoded by the ABCB1 gene, is the best known of the ABC transporters and was the first member of the ATP-binding cassette (ABC) superfamily to be cloned from multidrug-resistant cancer cells (Riordan *et al.*, 1985). Multidrug resistance (MDR) is the ability to simultaneously express cellular resistance to a variety of structurally and functionally unrelated chemotherapeutic agents. In tumour cells, expression of P-gp leads to a reduction in intracellular drug concentrations and thus to a decrease in the cytotoxicity of a broad spectrum of antitumor agents. The identification of P-gp more than three decades ago

facilitated the recognition that reduced intracellular accumulation of anticancer drugs can lead to significant levels of drug resistance; even today, basic research on P-gp is closely linked to clinical oncology (Lockhart *et al.*, 2003; Lum *et al.*, 1993). Over the years, P-gp became the prototype multidrug resistance (MDR) transporter, and studies concluding that drug resistance associated with P-gp leads to major failures of chemotherapy in human cancers (Juliano *et al.*, 1976) determined the use of the terminology "multidrug resistance transporter".

Treatment failure due to MDR was first found in the context of cancer, but was later also connected with other conditions, e.g., some autoimmune disorders and infectious diseases. Modulation of MDR transporters including P-gp has emerged as a pharmacological target for the treatment of cancer. Compounds inhibiting P-gp and related efflux proteins are supposed to increase the intracellular concentration of chemotherapeutic agents and restore their sensitivity.

P-gp is expressed in a variety of tissues, particularly at physiological barriers such as the luminal membrane of brush-border cells in the small intestine, in the epithelial cells, endothelial cells that form the blood–brain barrier (BBB), in proximal tubule epithelia of the kidney, and in the testis (Thiebaut *et al.*, 1987). It plays an important role in intestinal absorption and in biliary and urinary excretion of drugs, while in the cells of the BBB it helps to limit the entry of various drugs into the central nervous system.

The effect of P-gp on intestinal absorption of drugs can be remarkable. For example, the apparent bioavailability of paclitaxel in healthy volunteers increased from 4% to 47% when combined with a single oral dose of cyclosporine A, a potent P-gp inhibitor, resulting in an eightfold increase in systemic exposure (Terwogt *et al.*, 1999). Similarly, systemic exposure to paclitaxel was sixfold higher in P-gp knockout mice compared to wild-type mice following oral administration (Sparreboom *et al.*, 2003). P-gp was one of the first membrane transporters to be associated with drug–drug interactions (DDI) and the first one that was required by a regulatory agency for evaluation of its potential DDI risk during drug development (FDA, 2017).

Many drugs are substrates of P-gp. For this reason, the extent of expression and functionality of this transporter may directly influence the therapeutic efficacy of such agents. Modulation of the expression and function of P-gp through inhibition or induction mechanisms may affect the pharmacokinetics, efficacy, safety, and tissue levels of P-gp substrates (Cascorbi, 2006). Numerous studies have shown that P-gp has an important role in determining the concentration–time profiles of P-gp substrates in different parts of the body, supporting the importance of P-gp in the study of drug pharmacokinetics.

In general, P-gp substrates tend to have a hydrophobic planar structure with positively charged or neutral moieties; these include a wide range of structurally and pharmacologically unrelated compounds. Interestingly, many substrates of P-gp are also substrates of cytochrome P450 (CYP), particularly CYP3A4, an important drug metabolizing enzyme in the human liver and gastrointestinal tract. However, substrates of CYP3A4 are not always transported by P-gp, and not all P-gp substrates interact with CYP enzymes (Marzolini *et al.*, 2004). These two proteins are thought to have evolved to protect the host organism from exposure to environmental or dietary toxins.

In addition to multidrug resistance, P-gp expression is also associated with physiological or pathophysiological conditions. Decreased clearance of amyloid-β peptides from the brain is the cause of Alzheimer's disease; increased amyloid-β levels lead to cerebral amyloid angiopathy and permeability changes at the blood–brain barrier (BBB). The P-gp function may be involved in the pathogenesis of Alzheimer's disease (Vacirca *et al.*, 2011), since it has been demonstrated that amyloid-β is a P-gp substrate *in vitro* and it has been reported that P-gp has an important role in the removal of amyloid-β from

the brain. Inhibition or lack of expression of P-gp leads to increased amyloid-β levels in the interstitial fluid of the brain (Cirrito *et al.*, 2005). Moreover, an interesting role has also been suggested for P-gp in the neuroendocrine function and regulation of the hypothalamic-pituitary-adrenocortical (HPA) axis, since P-gp is involved in the efflux of certain natural and synthetic glucocorticoids from the brain (de Lange, 2007).

The role of P-gp in cancer and other human diseases, drug pharmacokinetics, drug–drug interactions, and toxicity highlight the clinical relevance of this membrane transporter. Understanding the biology, genetics, and biochemistry of P-gp provides clues to improve the treatment of cancer and helps to explain the pharmacokinetics of many commonly used drugs. Much is known about P-gp as a model member of the ABC transporter family, but much remains to be done, particularly to characterize its mechanism of action, and use this information to improve the treatment of human diseases.

2.2 Structure of P-glycoprotein

P-gp contains 1,280 amino acids with a molecular weight of 170 kDa. The structure of the protein consists of two symmetrical halves joined by a highly charged "linker region" ~75 amino-acid-long, which is phosphorylated at multiple sites by protein kinase C (Higgins *et al.*, 1997); each half has a transmembrane domain (TMD) and a cytosolic ATP-binding region called the nucleotide binding domain (NBD). The TMDs are thought to form the pathway by which drug molecules cross the membrane (Zhou, 2008). Each TMD consists of six highly hydrophobic α-helices embedded in the membrane bilayer and extending into the cytosol to form the intracellular loops (ICLs). Although the TMDs are structurally similar throughout the transporter family, they have a large proportion of non-conserved amino acids. In contrast, as with other ABC transporters, the NBDs of P-gp contain three highly conserved sequence motifs at which ATP is hydrolysed: the Walker A and Walker B motifs and the ABC signature motif. The NBDs comprise two subdomains: a catalytic RecA-like subdomain (Ye *et al.*, 2004) containing the Walker A and Walker B motifs, and a smaller and structurally more diverse α-helical subdomain containing the ABC signature motif (LSGGQ) (Zaitseva *et al.*, 2005). It has also been reported that the A, D, H, and Q loops are involved in nucleotide binding (Higgins *et al.*, 1997).

In 2009, the first crystal structure of murine apo-P-gp was resolved (Aller *et al.*, 2009). The X-ray structure suggested a nucleotide-free inward-facing conformation arranged as two halves with a pseudo twofold structure, formed by two bundles of six helices, resulting in a large internal cavity open to both the cytoplasm and the inner leaflet (Figure 2.1). Two portals described in the structure would allow the access of hydrophobic molecules directly from the membrane and an induced fit transport mechanism is proposed in which the transmembrane (TM) segments change to accommodate substrates with different sizes and shapes. According to their findings, the residues that interact with P-gp substrates, e.g., verapamil, are highly conserved, supporting the idea of a common mechanism of polyspecificity drug recognition (Aller *et al.*, 2009).

Despite decades of effort, only structures of the inward-facing conformation of P-gp were available; however, in 2018, the first structure of the human P-glycoprotein (*h*P-gp) in the outward-facing conformation was resolved (Figure 2.2.b) (Y. Kim *et al.*, 2018), determined by cryogenic electron microscopy (cryo-EM) at a resolution of 3.4 Angstroms. In this structure, the two NBDs form a closed dimer enclosing two ATP molecules. The drug-binding cavity is oriented toward the extracellular space and compressed to restrain

substrate binding. According to this observation, it is ATP binding, not hydrolysis, that promotes substrate release. The structure evokes a model in which the dynamic nature of the P-glycoprotein allows translocation of a wide variety of substrates.

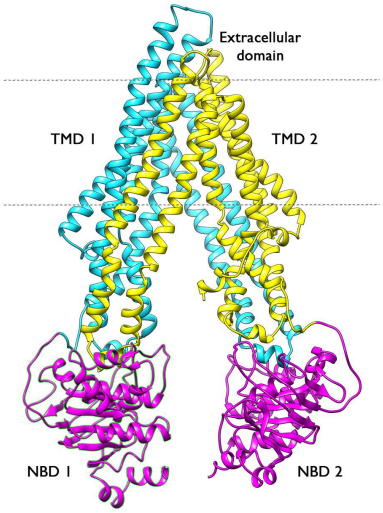
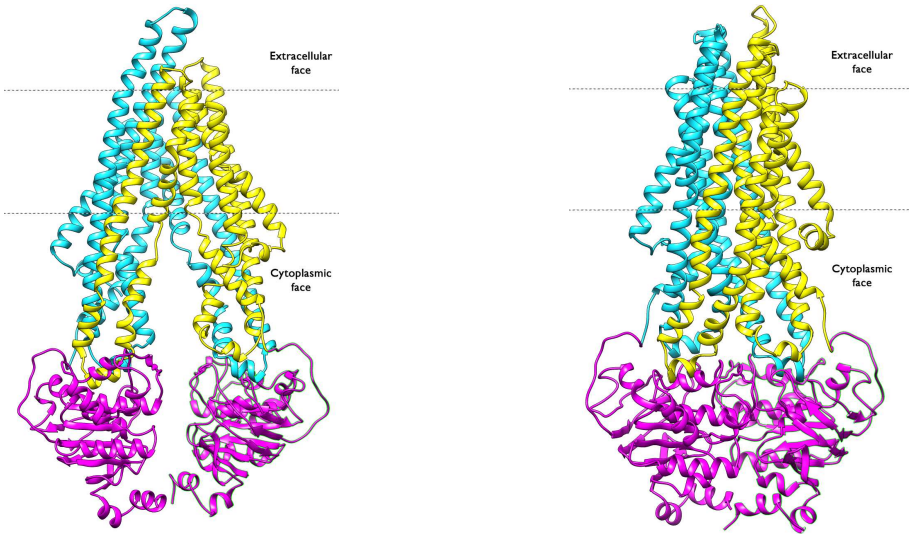


Figure 2.1: Crystal structure of *m*P-gp (PDB ID: 4M1M) in the inward-facing conformation. NBDs are shown in magenta and TMDs are shown in cyan and yellow.

A year later, in 2019, the 3.5 Angstroms cryo-EM structure of the substrate-bound *h*P-gp was determined (Figure 2.2.a) (Alam *et al.*, 2019). This structure reveals that the chemotherapeutic agent paclitaxel (Taxol) is bound in a central occluded pocket. A second structure of inhibited human-mouse chimeric P-gp showed that two molecules of zosuquidar occupy the same drug-binding pocket. This finding also suggests that minor structural differences between substrate- and inhibitor-bound P-gp sites affect the NBD movement and ATPase activity of P-gp.



(a) (b)

Figure 2.2: Cryo-EM structures of *h*P-gp. (a) inward-facing conformation (PDB ID: 6QEX); (b) outward-facing conformation (PDB ID: 6C0V). NBDs are shown in magenta and TMDs are shown in cyan and yellow.

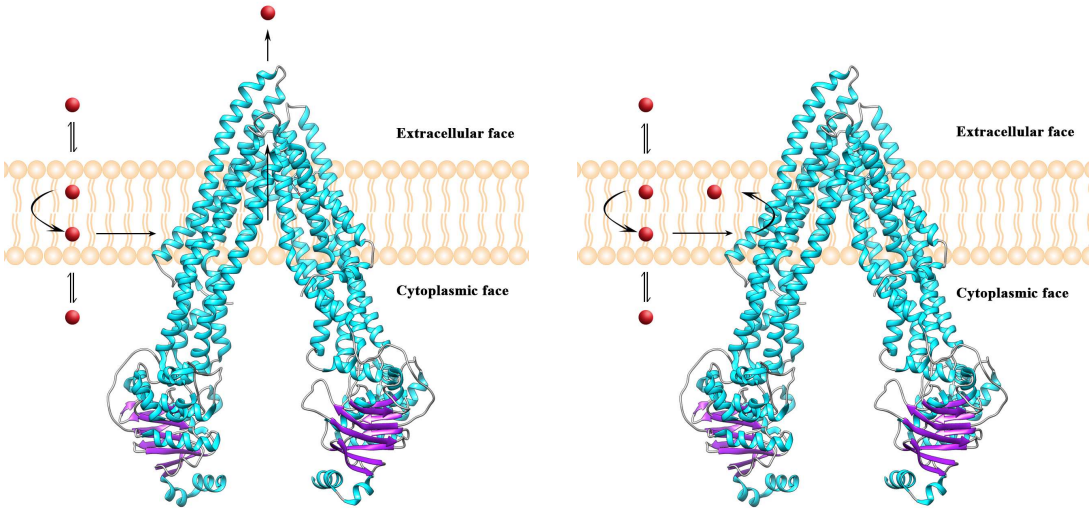
2.3 Mechanism of Transport

To date, the transport mechanism of P-gp remains highly controversial; however, two models have been proposed:

The hydrophobic vacuum cleaner model: according to this model, P-gp transports substrates out of the cell from the intracellular compartment or when they are in the lipid bilayer of the plasma membrane (Figure 2.3.a). This model suggests that P-gp can interact with its substrates within the membrane and subsequently efflux them into the extracellular medium like a hydrophobic "vacuum cleaner", using the energy derived from ATP hydrolysis (Higgins *et al.*, 1992; Sharom, 2014).

The "flippase" model: in this model, the protein functions as a drug translocase or flippase (Figure 2.3.b), moving its substrates from the inner to the outer leaflet of the membrane. Since the hydrophobic moiety of the substrate molecule is aligned with the hydrophobic core of the membrane, the substrate can diffuse laterally to the binding site on the P-gp, which is located in the inner leaflet of the lipid bilayer. The substrate is then flipped from the inner leaflet to the outer leaflet of the lipid bilayer, where it can rapidly partition into the extracellular aqueous phase or migrate by spontaneous flip-flop back to the inner leaflet(Higgins *et al.*, 1992).

The flippase model involves delivery of the drug to the outer leaflet followed by rapid distribution into the extracellular medium, whereas the vacuum cleaner model involves delivery of the drug into the extracellular medium followed by rapid distribution into the outer leaflet. The models are not mutually exclusive, and it is currently not possible to experimentally distinguish between them, since the same equilibrium state is reached in both cases.



(a) (b)

Figure 2.3: P-gp transport mechanism. (a) hydrophobic vacuum cleaner model; (b) flippase model. The small red sphere represents the substrate molecule.

2.3.1 ATP binding

ABC transporters are thought to share a common mechanism of action in which the transport of substrates involves the TMDs that alternately adopt an inward or outward

facing conformation, such that the TM pore alternately opens to one side of the membrane or the other. These transmembrane segments are thought to interact and undergo significant conformational changes during substrate binding or during ATP hydrolysis (E. Crowley *et al.*, 2010; Storm *et al.*, 2007).

The activity of P-gp is completely dependent on the presence of ATP (Romsicki *et al.*, 1998). The ATP-binding domains function as an ATPase that converts ATP to ADP to provide the energy needed to transport substrates across the membrane, often against steep concentration gradients. ATP must bind to both NBDs to enable P-gp activity; however, it is uncertain whether hydrolysis of both ATP molecules is necessary to generate this activity.

The ATP binding site is formed by the Walker A motif of one NBD and the signature motif of the other NBD. A Mg^{2+} ion, which is necessary for the hydrolysis of ATP, binds to the Walker B motif; according to some molecular dynamics studies, it binds to the conserved glutamate and aspartate residues (O'Mara *et al.*, 2012). The two NBDs form a dimer in which the two Mg^{2+} ions and ATP molecules are embedded between the Walker A motif of one NBD and the ABC signature motif of the other (J. Chen *et al.*, 2003; Smith *et al.*, 2002; Zaitseva *et al.*, 2005). Residues of the Walker B motif, A-loop, D-loop, H-loop, and Q-loop also contribute to the binding and hydrolysis of Mg^{2+} and ATP, e.g., the histidine residue in the H-loop was proposed to be responsible for the tight coupling of ATP binding and dimerization (P. M. Jones *et al.*, 2012). This mechanism is called "*nucleotide sandwich dimer*", first proposed by Jones *et al.* (P. M. Jones *et al.*, 1999, 2002) and later identified in crystallographic studies (Rosenberg *et al.*, 2001; Smith *et al.*, 2002). Dimerization of the NBDs is fundamental to the catalytic cycle of the ABC transporter by coupling with rearrangements in the TMDs and subsequent transport of the substrate across the membrane; therefore, both NBDs are essential for the correct function of P-gp.

The binding of ATP to the NBDs is stabilized by a series of interactions, such as, π-stacking interactions between a conserved aromatic residue preceding the Walker A motif and the adenosine ring of ATP; hydrogen bonding between the conserved lysine residue in the Walker A motif and oxygen atoms of the β- and γ-phosphates of ATP; coordination of the β- and γ-phosphates and residues in the Walker A motif with a Mg^{2+} ion; and the interaction of the γ-phosphate with the side chain of serine and the backbone amide groups of glycine residues in the signature motif (Ambudkar *et al.*, 2006; P. M. Jones *et al.*, 1999, 2002; I.-W. Kim *et al.*, 2006).

Although it has often been assumed that ATP hydrolysis drives the transport process, major conformational changes in P-gp occur after ATP binding and not after ATP hydrolysis. Similarly, the reduction in binding affinity of drugs to P-gp is due to ATP binding rather than hydrolysis (Martin *et al.*, 2000; Martin *et al.*, 2001; Rosenberg *et al.*, 2001). Therefore, ATP binding seems to cause the major conformational changes that reduce the drug binding affinity and expose the binding site to the extracellular medium; accordingly, ATP hydrolysis may simply "reset" the transporter to the initial conformation (Sauna *et al.*, 2007).

2.3.2 ATP hydrolysis and substrate translocation

The exact molecular mechanism by which ATP hydrolysis is coupled to substrate transport is still unknown (Qu *et al.*, 2003). However, two models have been proposed: the *switch model* (Higgins *et al.*, 2004) and the *constant contact model* (P. M. Jones *et al.*, 2009). In the *switch model*, the ATP molecule binds first to one of the two NBDs but is not hydrolysed. When the second ATP molecule binds to the second site, the NBDs

form a closed dimer conformation with the two ATP molecules at the dimer interface (Figure 2.4). The dimerization of the NBDs induces a conformational change in the TMDs, which decreases the affinity of the bound substrate and consequently causes its release across the membrane. ATP hydrolysis follows, and the release of ADP and Pi (inorganic phosphate) leads to the dissociation of the dimerized NBDs.

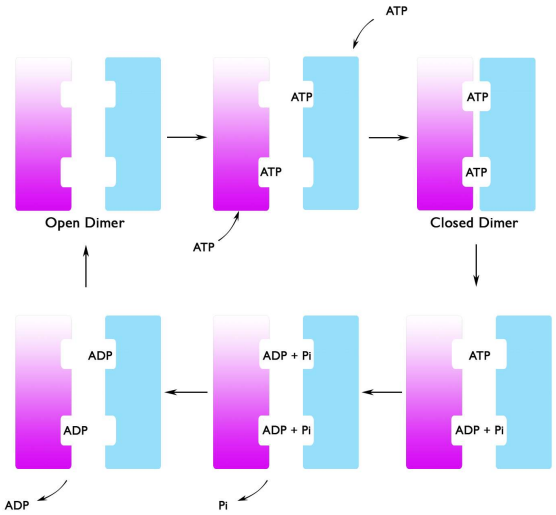


Figure 2.4: Switch model of the ATP hydrolysis mechanism. The NBDs are represented as magenta- and cyan-coloured shapes.

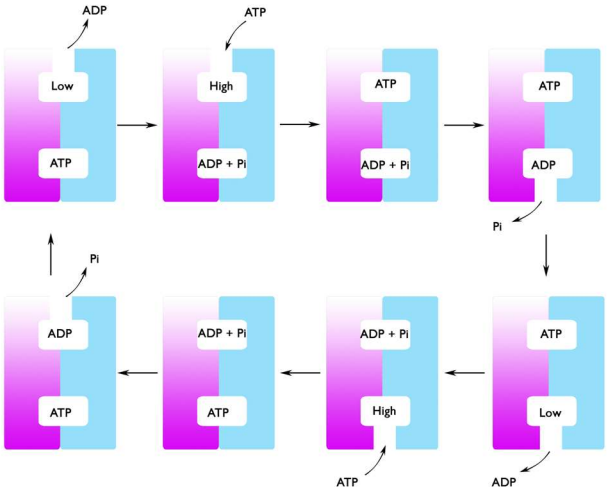


Figure 2.5: Constant contact model of the ATP hydrolysis mechanism. The NBDs are represented as magenta- and cyan- coloured shapes.

On the other hand, the *constant contact* model, based on the alternating site model developed by Senior *et al.* (Senior *et al.*, 1995), suggests alternating ATP hydrolysis in each NBD, where one ATP-binding site opens during ATP hydrolysis while the second ATP-binding site dimerizes and remains closed (Figure 2.5). In this model, an ATP molecule is occluded and hydrolysed at one site, while the opposite site is empty. The ATP hydrolysis at one site, with the hydrolysis products still bound in the occluded active site, induces the empty low affinity open site to switch to high affinity, allowing

the binding of a second ATP molecule. The ATP binding to the empty open core subdomain causes its inward rotation and the occlusion of the ATP. Occlusion of the second ATP molecule induces the opening of the occluded site containing the hydrolysis products, releasing Pi and ADP. The ATP-binding site opens enough to release the nucleotide without full separation of the NBDs. When this site closes with a new bound and occluded ATP molecule, the opposite site becomes ready for hydrolysis, and the process repeats in alternating cycles. The NBDs remain in contact throughout the cycle, with site opening and closing occurring through intrasubunit conformational changes within the NBD monomers.

2.4 P-gp Substrates, Inhibitors/Modulators, and Inducers

Compounds that interact with P-gp have been classified according to the nature of their interactions into substrates, inhibitors, modulators, and inducers (Colabufo *et al.*, 2010). The compounds that are actively transported by P-gp are considered substrates and therefore have a higher concentration outside the cell compared to the cytosol, while compounds that interfere with the function of the transporter are classified as inhibitors.

Substrates of P-gp include compounds of various unrelated structures such as anthracyclines, e.g., doxorubicin and daunorubicin; alkaloids such as reserpine, vincristine, and vinblastine; specific peptides such as valinomycin and cyclosporine; cardiac antidysrhythmic agents such as digoxin; steroid hormones such as aldosterone and hydrocortisone; and local anaesthetics such as dibucaine. Most P-gp substrates are amphipathic hydrophobic organic cations at pH 7.4, although they can also be anionic or uncharged molecules with molecular weights ranging from 300 to 2,000 Da. Some studies have tried to characterize the structural features that are required for the P-gp–substrate interaction, e.g., a hydrogen bond acceptor, two hydrophobic moieties and an aromatic ring centre were considered as the main features for the interaction with P-gp in a representative pharmacophore model (Schmid *et al.*, 1999). In general, the P-gp–substrate interaction seems to be correlated with the substrate lipophilicity and the number of hydrogen bonds present.

Inhibitors are the compounds that interfere with the substrate or nucleotide binding step, thereby blocking P-gp translocation. Based on the specificity and potency of inhibition, P-gp inhibitors are divided into three generations (Palmeira *et al.*, 2012). The first generation of inhibitors is less potent and interacts with multiple transporters. They were not designed to specifically inhibit P-gp but have other pharmacological properties, as well as rather low affinity for MDR transporters; an example of this first generation of P-gp inhibitors are verapamil and cyclosporine. One of the main problems with the first generation of P-gp inhibitors is the predominance of the original therapeutic activity of the drug, which makes the use of most of these compounds as P-gp inhibitors impossible; in phase I clinical trials, these drugs were either too toxic or not active enough.

The second generation of inhibitors have higher potency for P-gp inhibition and fewer side effects; their potency is about ten times that of cyclosporine. An example of a second-generation P-gp inhibitor is valspodar, a cyclosporine structural analog. The approach taken in the development of the second generation of P-gp inhibitors was to identify analogs of the first generation that do not have the pharmacological properties of the original molecule, but can specifically inhibit P-gp, with less toxicity and greater potency. However, some of the second-generation compounds lack P-gp selectivity because these inhibitors also inhibit several other transporters. The third generation of P-gp inhibitors, e.g., zosuquidar and tariquidar, are more potent than the first and second generations, and more specific for P-gp, avoiding the risk of blocking other transporters

and causing altered bioavailability or excretion of chemotherapeutic agents. They do not affect the cytochrome P450 3A4 (Leitner *et al.*, 2011; Wandel *et al.*, 1999), so they usually do not alter the plasma pharmacokinetics of the simultaneously administered antitumor drug and consequently do not require a dose reduction in chemotherapy. Despite all the advances, the third generation of P-gp inhibitors showed unexpected toxic effects in clinical trials.

Modulators interact with binding sites that are different from substrate binding sites; radioligand studies have shown that modulators reduce substrate binding in a non-competitive manner, suggesting negative allosteric communication between substrate and modulator binding sites (Colabufo *et al.*, 2010). Modulators and inhibitors exert the same biological end of restoring cell sensitivity to chemotherapeutic agents, for this reason the terms inhibitor and modulator have often been used as synonymous. In oncology, chemotherapeutic agents are co-administered with a modulator or an inhibitor with the aim of reversing MDR.

Induction of P-gp in the intestine, kidney, and peripheral tissues could reduce drug bioavailability, increase renal clearance, and decrease peripheral tissue distribution; the compounds that induce P-gp expression are called inducers. The induction mechanism is thought to occur through interaction with the nuclear receptors, pregnane xenobiotic receptor (PXR) and constitutive androstane receptor (CAR), which mediate xenobiotic drug-induced changes by increasing transcription of genes involved in drug clearance and disposition; they are activated by a structurally diverse spectrum of xenobiotics (Elmeliegy *et al.*, 2020). PXR and CAR can regulate the expression of a variety of metabolizing enzymes and drug transporters including P-gp and MRP2 (Y.-M. Wang *et al.*, 2012). Upon their activation, PXR and CAR bind to transcriptional binding sites for several drug metabolizing enzymes (DMEs) (Goodwin *et al.*, 1999; Kliewer *et al.*, 1998) and drug transporters (Geick *et al.*, 2001), resulting in increased expression of these proteins. Binding of PXR and CAR to the DR4 motif in the human P-gp promoter leads to increased transcription of P-gp.

In an effort to find safer and more effective P-gp inhibitors capable of overcoming MDR and to identify substrates in the early stages of drug development, computational models such as 2D-QSAR (two-dimensional quantitative structure-activity relationship), 3D-QSAR (three-dimensional quantitative structure-activity relationship), molecular docking, and pharmacophores have been widely used; considerable progress has been made in the predictability and accuracy of P-glycoprotein activity towards new drugs (L. Chen *et al.*, 2012).

Chapter 3

Ligand-Based Modelling Approach[1]

3.1 Introduction

In the last decades, molecular modelling and computational chemistry have taken an important role in understanding the basis of drug–receptor interactions and in assisting researchers to design new therapeutic agents, leading to the development of the field of computer-aided drug design (CADD).

CADD methods have been used as an alternative screening tool and have become a critical component of many drug discovery programmes; approaches such as virtual screening techniques are currently in widespread use (Jain, 2004; Stahura *et al.*, 2005). The role of CADD in drug discovery consists in screening large compound libraries into smaller clusters of predicted active compounds (hit identification), which enables lead compound optimization (hit optimization) by improving biological properties such as affinity for the target receptor and ADMET. The advances in computational chemistry algorithms and tools have driven the development of this field, helping to shorten the drug design process and reduce costs.

The CADD can be divided into ligand-based or structure-based techniques (Jorgensen, 2004). In the first case, only information from known active compounds (ligands) is used to identify other compounds with similar properties in one or more databases (Prathipati *et al.*, 2007). Ligand-based approaches are essential tools when structural information of the biological target is missing or when precise knowledge of the mechanism of action of the molecules are not known, including cases where the 3D structure of the target is known but the active site on the receptor is unknown. The ligand-based approach has been the most commonly used either as structure–activity relationship (SAR) or quantitative structure–activity relationship (QSAR). On the other hand, when the three-dimensional (3D) structure of the biological target is available, other more complex computational techniques such as molecular docking and molecular dynamics simulations can be used (Hansson *et al.*, 2002; G. M. Morris *et al.*, 2008). The basic idea behind ligand-based approaches is that the analysis of a set of molecules with experimentally determined activities can reveal the chemical features responsible for that activity. Ligand-based approaches were formulated before structure-based methods and still play an important role in drug design, either alone or in conjunction with structure-based techniques (Moro *et al.*, 2007). This chapter aims to provide a general overview of

[1] The chapter is based on the author's manuscript: Mora Lagares, L., Minovski, N., & Novič, M. (2019). Multiclass Classifier for P-Glycoprotein Substrates, Inhibitors, and Non-Active Compounds. *Molecules, 24*(10). doi: 10.3390/molecules24102006

the ligand-based modelling technique. The structure-based modelling approach is discussed in more detail in chapters 4 and 5.

In general, the ligand-based modelling approach consists of two elements: molecular descriptors and mathematical methods for deriving predictive models such as linear models, artificial neural networks, Support Vector Machines (SVMs), and others. The ligand-based methods require the creation of an appropriate model from a training dataset in the first step. Then, the model is applied to the screening set of molecules to make predictions for the property under study (Moro *et al.*, 2007).

Ligand-based methods include fingerprint-based similarity search, 2D-QSAR, pharmacophores, 3D-QSAR, and others (Vedani *et al.*, 2000; Vedani *et al.*, 2002). These approaches are faster and relatively easy to implement compared to structure-based drug design methods; one of their main advantages are the low CPU requirements. Ligand-based approaches allow the use of generalized descriptors, features, and fingerprints, making them effective virtual screening queries, and have evolved along with the advances in statistics and other machine learning algorithms such as regression, pattern recognition, and neural networks. One of their limitations is that ligand-based approaches are based on the widely accepted principles of molecular similarity, which state that structurally similar molecules should elicit similar biological responses; therefore, they are limited to the chemical space used in the formulation of the models, i.e., to their domain of applicability. Efforts are needed to develop methods that can be universally applicable (Prathipati *et al.*, 2007).

The P-glycoprotein (P-gp) ligand-based drug design relies on the knowledge of compounds known to interact with the membrane transporter. Structure-activity relationship (SAR) (Seelig *et al.*, 2000), quantitative structure-activity relationship (QSAR) (Dearden *et al.*, 2003), three-dimensional quantitative structure-activity relationship (3D-QSAR) (K. H. Kim, 2001), and pharmacophore models (Pajeva *et al.*, 2002) have been used to predict the activity of new compounds towards P-gp. The amount of literature dealing with the ligand-based approaches of P-gp is enormous and dates back several decades. Since the discovery of verapamil as a multidrug resistance (MDR) reversing agent (Tsuruo *et al.*, 1981), many SAR and QSAR studies have been published. One of the first studies on MDR modulation was conducted in 1988 using a series of vinca alkaloids with different structures and properties (Beck *et al.*, 1988).

SAR studies examine how structural variations in individual molecules affect their ability to interact with P-gp. QSAR studies, on the other hand, quantify the observed relationship between molecular descriptors and activity, it relates numerical properties of the molecular structure to the activity through a mathematical model (Perkins *et al.*, 2003). A variation of QSAR methods is the so-called 3D-QSAR, which refers to the application of molecular field calculations that require 3D structures. This involves the study of steric and electrostatic fields, and other atomic or molecular properties, which are then used as unique parameter sets for QSAR (Cianchetta *et al.*, 2005). This method was called Comparative Molecular field Analysis (CoMFA) (Cramer *et al.*, 1988) and was the first method widely accepted and used for 3D-QSAR. Comparative Molecular Similarity Index Analysis (CoMSIA) is another one of the most common 3D-QSAR algorithms (Myint *et al.*, 2010). In 1997, the first P-gp 3D-QSAR analysis was performed using structurally related thioxanthenes (M Wiese *et al.*, 1997).

The first attempt to characterize key pharmacophoric features for P-gp modulation was performed by Seelig in 1998. Seelig proposed a pharmacophore based on specific arrangements of electron-donor groups (Seelig, 1998). The pharmacophore approach tries to define the minimal structural features that a molecule must possess in order to bind to P-gp.

In this chapter, we present a detailed description of a ligand-based method for modelling the P-gp activity: the construction of a classification model that provides a qualitative prediction of P-glycoprotein inhibition/substrate activity. The developed P-gp activity model has been successfully implemented within the online platform VEGAHUB (Benfenati *et al.*, 2013), which is freely available to the public at https://www.vegahub.eu/portfolio-item/vega-qsar/.

3.2 Multiclass Classifier for P-Glycoprotein Substrates, Inhibitors, and Non-Active Compounds

3.2.1 Abstract

P-glycoprotein (P-gp) is a transmembrane protein that actively transports a wide variety of chemically diverse compounds out of the cell. It is closely related to the ADMET properties (absorption, distribution, metabolism, excretion, and toxicity) of drugs/drug candidates and contributes to reducing toxicity by eliminating compounds from cells, thus preventing intracellular accumulation. Therefore, during drug discovery and toxicological evaluation, it is advisable to pay attention to whether a compound under development could be transported by P-gp.

In this study, an *in silico* multiclass classification model capable of predicting the probability of a compound to interact with P-gp was developed using a counter-propagation artificial neural network (CP ANN) based on a set of 2D molecular descriptors, as well as an extensive dataset of 2,512 compounds (1,178 P-gp inhibitors, 477 P-gp substrates and857 P-gp non-active compounds).

The model provided a good classification performance, producing non error rate (NER) values of 0.93 for the training set and 0.85 for the test set, while the average precision (AvPr) was 0.93 for the training set and 0.87 for the test set. The performance of the model was challenged on an external validation set of 385 compounds, producing NER and AvPr values of 0.70 for both indices. We believe that this *in silico* classifier could be effectively used as a reliable virtual screening tool for identifying potential P-gp ligands.

3.2.2 Introduction

The P-glycoprotein (P-gp or ABCB1) is a transmembrane protein that belongs to the ATP-binding cassette family of transporters (ABC-transporters). It is an efflux pump that actively transports a large number of compounds (structurally diverse) out of the cell using the energy provided by the hydrolysis of ATP (Sauna *et al.*, 2007; Vasiliou *et al.*, 2009). Efflux pumps are considered as a first line of defence against toxicants; they decrease the toxicity due to xenobiotic exposure by restricting their absorption and transport, thus preventing their intracellular accumulation.

The P-glycoprotein has a great influence on the ADMET properties (absorption, distribution, metabolism, excretion, and toxicity) of drugs and toxins. This is evidenced by its widespread expression in the small intestine, liver, colon, kidneys, placenta and the blood-brain-barrier (BBB)(Sharom, 2008), i.e., tissues that perform an excretory or barrier function. P-gp is also involved in the multidrug resistance (MDR) phenomenon (Fromm, 2000; Leslie *et al.*, 2005), whereby drugs are pumped out of the cell and their concentration is lowered at the intracellular target site. This was assessed to be a pivotal reason for chemotherapeutic failure in the treatment of various cancers, as P-gp is commonly over-expressed in tumour cell lines (Kartner *et al.*, 1983).

Given the role of P-gp in drug absorption and disposition, particular attention should be paid to both substrate and inhibitory properties of new compounds. Inhibitors are of particular interest in drug-drug interactions; several cases have been reported in which co-administration of a P-gp inhibitor and a P-gp substrate often resulted in increased blood concentrations of the substrate, which in turn can cause serious side effects (Greiner *et al.*, 1999; Montanari *et al.*, 2015). Considering the important role of ABC transporters in reducing the toxic events associated with drug–drug interactions, the elucidation of ligand–transporter interactions for effectively predicting the toxicity and safety of a ligand requires much more attention and complex strategies.

Nowadays, the Food and Drug Administration (FDA) recommends a standardized set of experiments to assess the likelihood of a compound to interact with P-gp and the breast cancer resistance protein (BCRP/ABCG2) (OECD, 2007). According to the FDA recommendations, "all investigational drugs should be evaluated *in vitro* to determine if they are a potential substrate of P-glycoprotein" (FDA, 2017). This demonstrates that *in silico* models are important for predicting the interaction of a compound with P-gp in the early stages of drug discovery, development, and toxicological assessment.

Many studies have been performed with the aim of identifying P-gp substrates or developing more potent, selective and specific P-gp inhibitors; however, the wide variety of ligands (polyspecificity or promiscuity of P-gp) complicates the effective design of new P-gp interacting compounds (Demel *et al.*, 2009). In other words, most P-gp inhibitors interact with the same binding site as P-gp substrates, even when substrates and inhibitors have different biological purposes, suggesting that the two classes may share numerous structural similarities, making it extremely difficult to distinguish compounds belonging to one class from the other. A further complication is that P-gp inhibitors and substrates are measured using different assay protocols. For example, if a bidirectional transport assay based on a polarized epithelial monolayer overexpressing P-gp is used, P-gp substrates can be distinguished from inhibitors by a net flux above or below 2 (Wessler *et al.*, 2013). However, most data available in the literature often lack this information. Consequently, one must be aware of the promiscuous nature of P-gp and the different assay protocols when pooling data from different sources.

The first P-gp studies started with ligand-based approaches (e.g., Beck *et al.* (Beck *et al.*, 1988), Seeling *et al.* (Seelig *et al.*, 2000), Ichiro *et al.*, (Ichiro *et al.*, 1989) and Dearden *et al.* (Dearden *et al.*, 2003; Ramu *et al.*)), which extended to the structure-based level when the first X-ray structure of a mouse P-gp became available (PDB ID: 3G5U) (Aller *et al.*, 2009). Several methods available for ligand-based design have been applied, including ligand-based pharmacophore modelling (Güner, 2000), linear and non-linear classification algorithms (Freeman *et al.*, 1991), as well as supervised and unsupervised artificial neural networks (Zupan *et al.*, 1993). These methods were mainly applied for constructing *in silico* models relating the P-gp inhibitory activity of the compounds and a set of considerably basic physicochemical parameters (e.g., lipophilicity, H-bonding, aromatic rings, and charge).

In the last 10 years, a variety of P-gp classification models were developed focusing on large datasets with the purpose of screening compound libraries. Some of the most relevant contributions are related to the work by Broccatelli *et al.* (Broccatelli *et al.*, 2011) and Chen *et al.* (L. Chen *et al.*, 2011). In both studies, large datasets with thousands of compounds were used to develop models for the classification and prediction of P-gp inhibitors. According to their results, the compound's lipophilicity expressed as the logP proved to be a crucial parameter for distinguishing between P-gp inhibitors and non-inhibitors.

In addition to the P-gp inhibitor/non-inhibitor classification models, various pharmacophore models have also been developed. It is widely known that pharmacophore

models can contribute significantly to the understanding of probable P-gp-ligand interactions by assessing the relevant pharmacophoric features. However, considering the high structural diversity of known P-gp ligands, the concept of pharmacophore modelling did not lead to a better understanding of their polyspecificity. Nevertheless, some of the models provided good results in identifying new ligands as in the study by Palmeira *et al.* (Palmeira *et al.*, 2011) in which starting from a DrugBank screening, 12 compounds were identified to increase the intracellular accumulation of Rhodamine-123, an *in vitro* biologically confirmed P-gp substrate.

In the case of P-gp substrate and non-substrate classification, the studies by Wang *et al.* (Z. Wang *et al.*, 2011), and Li *et al.* (D. Li *et al.*, 2014) can be considered as the studies with the largest datasets reported so far (consisting of 332 and 723 compounds, respectively). Their results link substrate activity to a set of quite common physicochemical properties, such as molecular weight and water solubility.

For structure-based P-gp modelling, several molecular docking studies of ligands were performed based on the available murine crystallographic data (PDB ID: 3G5U) (Aller *et al.*, 2009) with the aim of understanding the molecular interactions that determine the binding of the ligands to P-gp, as shown in the studies by Dolghih *et al.* (Dolghih *et al.*, 2011) and Ferreira *et al.* (Ricardo J. Ferreira *et al.*, 2013b).

Compared to experimental drug design methods, *in silico* methods are faster, cheaper, more efficient and high-throughput in screening, with reduced labour and use of animals (Wongrattanakamon *et al.*, 2017). Therefore, the aim of the present study was to develop an efficient *in silico* screening tool capable of providing a rapid and cost-effective platform for the identification of potential P-gp ligands that can be effortlessly used in drug discovery and toxicological assessment. Here, we show the development and performance of a multiclass classifier capable of distinguishing between substrates and inhibitors, rather than just active/non-active compounds. The identification of potential P-gp substrates and inhibitors is of great importance, not only for the development of agents that can counteract the mechanisms of multidrug resistance (MDR reversal agents), but also for the identification of P-gp substrates from drug candidates.

3.2.3 Materials and Methods

3.2.3.1 Dataset

The compounds present in the dataset were collected mainly from the admetSAR database (Cheng *et al.*, 2012). This database is a compilation of diverse chemicals gathered from different literature sources associated with known ADMET profile. The data was extracted from the original studies (Broccatelli *et al.*, 2011; L. Chen *et al.*, 2011; Z. Wang *et al.*, 2011) as SMILES notations together with their corresponding experimentally determined P-gp class (inhibitors, non-inhibitors, substrates, and non-substrates). An additional set of P-gp substrates was retrieved from the work of Li *et al.* (D. Li *et al.*, 2014).

To increase the size of the dataset and to extend the chemical space, we collected and included compounds derived from different references which use different types of experimental assays to assess the P-gp class. Therefore, before constructing the model, pre-processing of the data was required to detect duplicate compounds and compounds with both experimental classes (or overlapping classified compounds). The P-gp non-inhibitor and non-substrate compounds were merged into the non-active class. The overlap of negative compounds in both sets was desirable (See Appendix A, Figure A.1.), so they were included in the non-active class, while all other overlaps that might introduce uncertainty into the model (S/I = 42 S/NI = 29 I/NS = 10; S: substrate I:

inhibitor NI: non-inhibitor NS: non-substrate) were removed. The final dataset includes 2,512 structurally diverse compounds, e.g., acridone derivatives, flavonoids, azoles, antidepressants of the selective serotonin reuptake inhibitor (SSRI) class, persistent organic pollutants (POPs), β-lactam antibiotics, and benzodiazepines, among others, which can be divided into three main classes, i.e., 1,178 P-gp inhibitors, 477 substrates, and 857 non-active compounds.

In this study, 42 compounds were found in both the substrate and inhibitor classes. These compounds were removed from the dataset because for the purposes of our study, we wanted to include only well-defined compounds in each class. Some drugs may belong to multiple P-gp classes (Wessler *et al.*, 2013), but the available experimental assays use different criteria to classify P-gp-interacting compounds, resulting in different reports of their class. Differences in reports of P-gp class are quite common and could be due to the diversity of available assays and the criteria used to determine P-gp activity (threshold used) in each assay. The promiscuity of the P-gp transporter itself and its interacting ligands is another possible reason.

The data curation was mainly performed utilizing the software Pipeline Pilot 9.2 (Accelrys, 2014). Since the dataset was collected from different literature sources, the existing SMILES notations showed a high level of heterogeneity. To facilitate the data curation, it was necessary to convert the original SMILES notations into a uniform representation, running a Pipeline Pilot protocol. The protocol used includes the Canonical SMILES component, which adds canonical smiles as a new property to the dataset. All newly generated SMILES were then combined into a single SDF file format along with their P-gp class notation. Duplicate compounds were identified and removed from further analysis by running a pipeline pilot protocol that includes the Remove Duplicate Molecules component. In this component, the canonical SMILES was set as a filter to find duplicates. Additionally, compounds classified as belonging to more than one class, defined as overlapping compounds, were also discarded from the analysis. After running the pipeline pilot protocol, some duplicate compounds were still present in the dataset. Removal of the remaining duplicates and overlapping compounds was performed manually based on the descriptor values for the molecules.

Prior to the modelling, the entire dataset was divided into training (TR), test (TE) and validation (V) sets, comprising 1,786, 341, and 385 compounds, respectively, utilizing the Kohonen mapping as implemented in the CPANNatNIC software (Drgan *et al.*, 2017).

3.2.3.2 Descriptors Calculation

2D molecular descriptors were calculated for the entire dataset using the software Dragon 7.0 (Mauri *et al.*, 2017). Initially, a total of 1,229 molecular descriptors were calculated, and their values were normalized according to Eq. (3.1):

$$x_i^{norm} = \frac{x_i - \overline{x}}{s_x} \tag{3.1}$$

where x_i^{norm} represents the normalized value of descriptor x_i for the i^{th} molecule, the $\overline{x}$ is the average of all descriptor values x_i in the dataset, while s_x is the standard deviation.

With the intention to eliminate the uninformative descriptors (noise) as well as to prevent over-fitting of the model, a variable reduction was performed on the initial set of descriptors before the modelling. For this reason, the descriptors with constant values as

well as those with a standard deviation of less than 0.0001 were removed, as they provide little information for the construction of the model. In addition, descriptors that are orthogonal to each other were identified by pair-wise correlations using the Pearson correlation coefficient; if two descriptors have an absolute correlation coefficient above the desired threshold, only one of them is retained, i.e., redundancy is avoided. Descriptors with an absolute pair correlation coefficient value greater than or equal to 0.95 were removed.

To further reduce the likelihood of correlations between descriptors, a Kohonen top-map was used (Drgan *et al.*, 2017). In this way, the remaining descriptors were mapped onto a network with a 7 by 7 architecture of neurons using the transpose of the descriptor matrix; two descriptors were selected from each neuron, those with the largest and the shortest Euclidean distance to the central neuron, yielding a final set of 96 molecular descriptors for further use.

3.2.3.3 Selection of Training, Test and Validation Sets

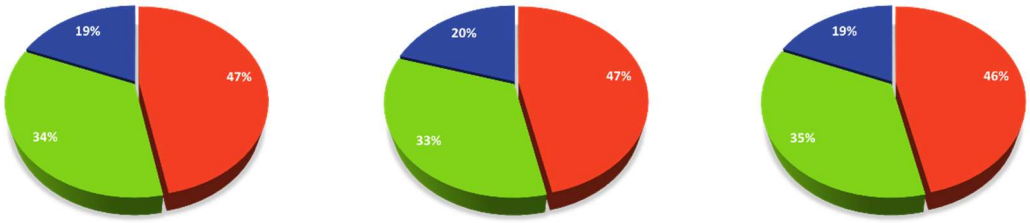
The methodology used for splitting the dataset is fundamental to obtain consistent results; it must guarantee that the TR set incorporates all sources of expected variability. The most diverse samples should be included into the TR set and be selected in such a way that they are as representative as possible of the global dataset.

The global dataset was divided into TR, TE and V set based on the clusters formed in the top-map of the Kohonen neural network. For this purpose, the entire dataset was mapped onto the network using the 1,229 calculated 2D molecular descriptors. The information space covered by the whole map should be well represented in each subset and to fulfil this requirement, the selected compounds were distributed over the entire Kohonen top-map.

In order to obtain the best distribution of the objects in the Kohonen top-map, the technical parameters of the network were adjusted, including the network size, the number of learning epochs (training iterations), and the maximum and minimum learning rates. Fixed parameters of the network were non-toroidal boundary conditions and triangular neighbourhood function. The selection criterion of the best network for splitting purposes was the minimum average error on one object at the maximum neuron occupancy.

The global dataset was mapped onto the network. The V set containing 385 compounds was selected and not used during the model construction and optimization procedures. The TR and TE sets were selected from the remaining compounds; 1,786 compounds were chosen for TR set and 341 for TE set (Figure 3.1). The network parameters used for mapping the dataset were: 20 $\times$ 20 neurons, 100 learning epochs, maximum learning rate 0.5 and minimum learning rate 0.01.

Since the formation of the clusters in the Kohonen top-map is based on an unsupervised learning methodology, the resulting distribution in the top-map is only influenced by the structural descriptors used, e.g., percentage of H atoms and the number of secondary amides (aromatic); the clusters form as a result of the structural similarity of the objects.



(a) (b) (c)

Figure 3.1: Dataset distribution. (a) Training set (TR), (b) Test set (TE), and (c) Validation set (V). Red slice represents P-gp inhibitors; blue slice represents P-gp substrates and green slice represents non-active compounds.

3.2.3.4 Feature Selection

With the intention of constructing a classification model based on the most significant variables (molecular descriptors), and consequently to optimize its predictive ability, robustness, and reliability, the feature selection was performed on the entire final set of 96 descriptors using the genetic algorithm (GA) (Leardi, 2003) coupled with counter-propagation artificial neural networks (CP-ANNs) (Zupan *et al.*, 1993).

A population of 95 chromosomes (binary vectors) evolving in 150 generations (iteration steps) was considered in several combinations of different networks and GA parameters. Several GA runs with different random origins were performed. Fixed parameters were maximal (0.6) and minimal (0.001) learning rate, number of survivals (20) and per cent of mutations (0.02). On the other hand, the number of neurons, the number of learning epochs, and the number of (initial) genes were modified to optimize the selection of variables.

3.2.3.5 Construction of the Model

To construct the classifier, the following steps were executed:

(1) The dataset was unified into a single SDF file format that contained the structural information and experimentally determined class (substrate, inhibitor, or non-active) for each molecule.
(2) 2D molecular descriptors were calculated for the entire dataset using the software Dragon 7.0 (Mauri *et al.*, 2017).
(3) The dataset was divided into TR, TE, and V sets using the Kohonen map clustering method. The V set was excluded and not considered in the training process.
(4) An initial reduction in the number of descriptors was made using a Kohonen top-map.
(5) The TR set was used to construct and train several CP-ANN models; the TE set was used to tune the hyperparameters of the classifier.
(6) The models were optimized using the GA to select the most informative variables to include. The model with the best predictive ability was selected based on the product of the Mathew correlation coefficient (MCC) values (Figure 3.2) of the TE and TR sets (MCC_{TR} * MCC_{TE}).

$$\textbf{Sensitivity} = \frac{TP}{TP + FN}$$

$$\textbf{Specificity} = \frac{TN}{TN + FP}$$

$$\textbf{MCC} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Figure 3.2: Primary measures related to single classes. TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives and FN is the number of false negatives.

3.2.3.6 Methods

3.2.3.6.1 Self-Organizing Maps (SOM) or Kohonen Artificial Neural Networks

Kohonen artificial neural networks (KANNs) (Zupan *et al.*, 1993) were used for the selection of the TR, TE, and V sets, as well as for the variable reduction (descriptors). KANNs, widely known as self-organizing maps (SOM), are a type of artificial neural network (ANN) that can be used for clustering and visualization tasks. The outcome of the KANN is the mapping of multi-dimensional information into a two-dimensional plane of neurons, based on the similarity among the objects.

The input for KANNs is a vector of independent variables (descriptors) $X_s = (x_{s1}, x_{s2}, x_{si}, \dots, x_{sm})$ and the winning neuron W_c is the neuron with the weights closest to the input according to the calculated Euclidean distance Eq. (3.2):

$$d_j = \sqrt{\sum_{i=1}^{m} (x_{si} - w_{ji})^2} \tag{3.2}$$

The next step consists of the correction of the weights of the winning neuron and all the neurons in the range of the topological distance $0 > D < D_{max}$, in order to make them more similar to the input variable, using the following Equations (3.3)–(3.5):

$$w_{ji}^{new} = w_{ji}^{old} + \Delta w_{ji} \tag{3.3}$$

$$\Delta w_{ji} = \eta(t) a (D_c - D_j)(x_i - w_{ji}^{old}) \tag{3.4}$$

$$\eta(t) = (a_{max} - a_{min}) \frac{t_{max} - t}{t_{max} - 1} + a_{min} \tag{3.5}$$

where parameter η is the learning rate that has a maximum value at the beginning, i.e., a minimum value at the end of the learning process; $(D_c - D_j)$ is the topological distance between the central neuron W_c and the current neuron j function.

The Kohonen type of net is based on a single layer of neurons arranged in a two-dimensional plane that has a well-defined topology. A defined topology means that each neuron has a defined number of neurons as nearest neighbours, second-nearest neighbours, etc. (Figure 3.3). The neighbourhood of a neuron is usually arranged either in squares or in hexagons, which means that each neuron has either four or six nearest neighbours.

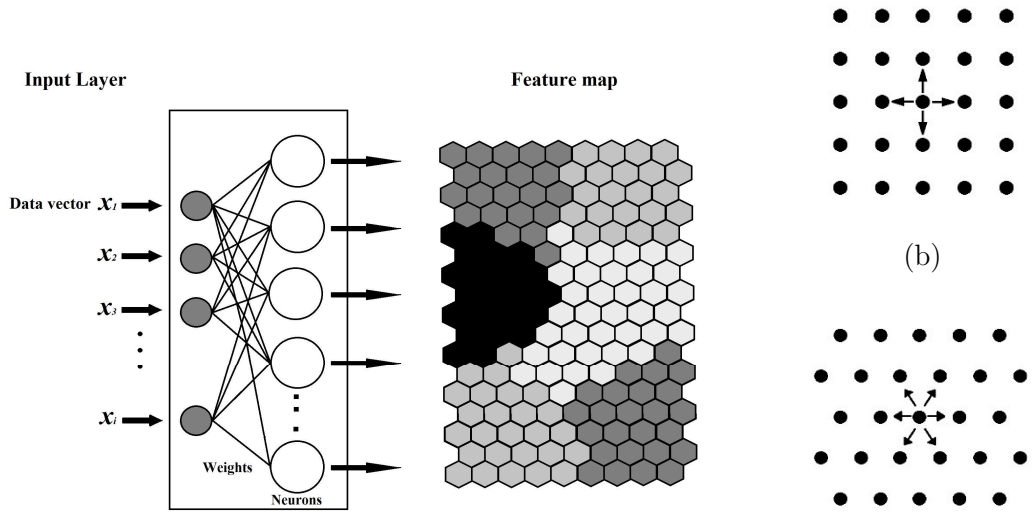


Figure 3.3: Diagram of a Kohonen artificial neural network. (a) Diagram, (b) Square layout of neighbours (S), (c) Hexagonal layout of neighbours.

3.2.3.6.2 Counter-Propagation Artificial Neural Network The counter-propagation artificial neural networks (CP ANNs) (Zupan *et al.*, 1993) method was used for the construction of the model. This learning strategy is based on a supervised competitive learning, and it requires a set of object-target pairs of data for training and verification.

The architecture is basically a two-layer network consisting of a Kohonen layer (influenced by the inputs, i.e., independent variables) and an output layer (influenced by the targets, i.e., dependent variables) (Figure 3.4). The inputs are fully connected to the Kohonen network, where competitive learning is performed, i.e., each unit in the input layer is linked to all neurons in the Kohonen Layer.

In Figure 3.4, the weights connecting the input unit i with the Kohonen neuron j are labelled as w_{ji}, each neuron in the Kohonen layer is described by a weight vector W_j. The neurons of the Kohonen layer are connected to the neurons in the output layer. This is a full connection. However, after each input, only a certain neighbourhood of a given neuron is connected to the output neurons, and only the weights linking these neurons are allowed to change. The weights connecting the *j-th* neuron in the Kohonen layer with th *k-th* neuron in the output layer are labelled u_{kj}, and the weight vector belonging to a given output neuron is labelled R_k. A given answer is not stored as a set of weights in one neuron, but as one component of the weights of all the output neurons. This kind of organization requires the number of neurons in the Kohonen layer to be equal to the number of answers to be stored, and the number of neurons in the output layer to be equal to the number of variables comprising the output answer, e.g., one thousand answers, each consisting of four variables, requires a Kohonen network with one thousand

neurons, and an output layer with four. The input layer should have the same number of units as input variables.

The training process is very similar to the KANNs. First the objects are mapped in the Kohonen layer in an unsupervised manner and then the supervised learning is used for the correction of weights in the output layer. The weights are adapted by comparing the actual output with an ideal output. The weights in the output layer are influenced by the position of the winning neuron in the input layer, as it defines the neighbourhood in the output layer, and by the target values. Once the winning neuron has been selected, two types of correction are made: first, the correction of weights w_{ji} within the neurons of the Kohonen layer, and secondly, the correction of weights in the output layer u_{kj} according to Eq. (3.6):

$$\Delta u_{kj} = \eta(t)a(D_c - D_j)(y_i - u_{kj}^{old}) \tag{3.6}$$

The outcome of the learning process is an arrangement of objects in a two-dimension map that corresponds exactly to the maps generated in the Kohonen layer. The output is taken from all weights between one Kohonen neuron and all the output neurons and there is a one-to-one correspondence between the neurons in the Kohonen map and those in the output map. CP ANNs can be used for building models able to predict unknown properties of new objects. It is also a suitable tool for clustering and classification tasks.

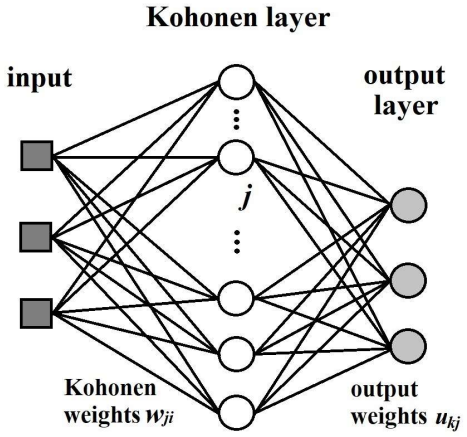


Figure 3.4: The layout of a counter-propagation artificial neural network (CP ANNS).

3.2.3.6.3 Genetic Algorithm The genetic algorithm (GA) (Leardi, 2003) was used in combination with CP ANNs for descriptors selection to improve the performance of the model. This algorithm of heuristic search mirrors the process of natural selection, where the fittest individuals are selected for reproduction to generate offspring of the next generation.

By implementing this in a multidimensional descriptor space, GA can be used effectively for descriptor selection and/or the optimal adjustment of parameters, which must be passed to a function that evaluates how well they solve the problem. The general steps for descriptor selection using GA are as follows:

1) Create an initial population of genetic vectors and calculate their fitness.
2) Choose two members of this population based on their fitness to become parents.
3) Use a mating operator to construct a new genetic vector from the parents.
4) Use a mutation operator to probabilistically change the genetic vector.

5) Calculate the fitness of this offspring and have it replace the weakest member in the population.
6) Return to step 2 until a sufficient number of offspring have been produced.

3.2.3.6.4 Applicability Domain The applicability domain (AD) of (Q)SAR models is defined by the Organization for Economic Co-Operation and Development (OECD) as the response and chemical structure space in which the model makes predictions with a given reliability. It is defined as a "physicochemical, structural, or biological space, knowledge or information on which the training set of the model has been developed, and for which it is applicable to make predictions for new compounds" (OECD, 2007). The purpose of defining the AD for a given (Q)SAR model is to determine the prediction accuracy of a new unknown compound independently of our naïve interpretation of its similarity to the molecules from the set used to construct and validate the model.

There are different approaches for describing the AD of a model: methods based on ranges of molecular descriptors, geometrical methods, distance-based methods, and probability distribution-based methods. Within distance-based approaches, three are widely used in (Q)SAR research: Euclidean (ED), Mahalanobis and city-block distance. The AD of the model in this study was analysed using the ED between objects (molecules) and the central neuron of the neural network.

The ED between the molecules and the central neuron (in the Kohonen layer of CP ANN models) is the fundamental characteristic of the neural network. It represents the interval between a central node (*ci*) in the Kohonen layer and an input pattern (X). The ED can be expressed using Eq. (3.6):

$$ED(X, w_{ci}) = sqrt\left((x_1^T - w_{ci1})^2 + (x_2^T - w_{ci2})^2 + \cdots + (x_m^T - w_{cim})^2\right) \tag{3.6}$$

where w_{ci1}, w_{ci2},..., w_{cim} are the weights to the neuron ***ci*** corresponding to a particular descriptor, and ***m*** is the number of descriptors or levels corresponding to a particular descriptor in the Kohonen layer. Input patterns for each level can be expressed as transposed matrixes. Each transpose matrix $(x_1^T, x_2^T, \ldots, x_m^T)$ includes the values of descriptors $D_1, D_2, \ldots, D_m$ respectively, calculated for each molecule. The distances are unitless because the descriptors have been previously autoscaled.

The goal of an AD is to set up boundaries whereby the obtained predicted values can be trusted with confidence. However, there is not a clear consensus about the determination of thresholds in AD for non-linear classification models (Fjodorova *et al.*, 2011). Since our model is a non-linear model, we did not set a warning threshold, but investigated the prediction accuracy of the models in the chemical and descriptors space and tried to find out the space in which the models provide reliable predictions.

A particularly important point to consider when determining the AD is the uncertainty. There may be input uncertainties as well as variability and structural (model) uncertainties resulting from simplifications of the reality due to limited systematic knowledge. As a result of uncertainties associated with individual (Q)SAR predictions, some predictions may fall within the defined AD of the model but be unreliable due to properties and features not accounted for by the model. On the other hand, a chemical that falls outside the defined AD may still exhibit the response being modelled because it brings out this response by a mechanism not accounted by the model under study. The essential problem of the AD definition is to find out the uncertainty areas into which less reliable predictions fall.

3.2.4 Results and Discussion

The dataset used consists of compounds that were tested for interactions with P-gp. A total of 2,512 compounds were grouped into three classes: P-gp inhibitors, substrates, and non-active molecules. Figure 3.5 shows an example of the P-gp 3D structure with a bound ligand.

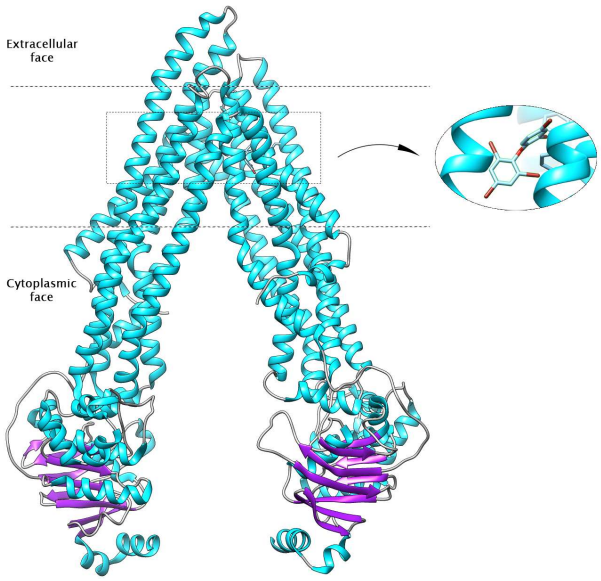


Figure 3.5: Crystal structure of *m*P-gp with the ligand PBDE-100 (PDB ID: 4XWK). The structure is coloured according to the secondary structure: helices are shown in cyan, beta sheets in purple, and coils in gray.

3.2.4.1 Descriptors

The optimization procedure based on genetic algorithm (Leardi, 2003) resulted in the selection of 26 molecular descriptors (Table 3.1) that were used in the construction of the classification model. Among these 26 descriptors, there are six 2D atom pairs; three CATS 2D descriptors, three functional group counts and three 2D autocorrelations; two ring descriptors and two P_VSA-like descriptors; one atom-type E-state index, one atom-centred fragment, one 2D matrix-based descriptor, one information index, one connectivity index, one walk and path count and one constitutional index.

Most of the selected descriptors (more than half) are count descriptors that provide information about occurrences or specify the presence/absence of predefined structural features in the molecule, such as functional groups, augmented atoms, pharmacophore point pairs, atom pairs, presence of rings, and walk and path counts. Some of these count descriptors help in discriminating cyclic compounds from acyclic ones and reflect the local geometrical environment in complex cyclic systems. Therefore, they contribute to a deeper understanding of the structural complexity of the molecules.

Other molecular descriptors that are more prevalent in the selected set are the autocorrelation descriptors. They describe how the property under consideration is distributed along the topological structure of the molecule, e.g., the ATS descriptor corresponds to a decomposition of the square molecular property in different atomic contributions. On the other hand, the included P_VSA-like descriptors provide information about the amount of Van der Waals surface area (VSA) having a particular

property P in a certain range. The properties considered in this case are the log P and ppp (potential pharmacophore points) hydrogen-bond donor. The selection of these two properties is consistent with previous *in silico* studies suggesting that essential chemical properties such as lipophilicity (Klepsch *et al.*, 2011; Ramu *et al.*, 1992) and hydrogen bond acceptor/donor (Seelig *et al.*, 2004) play an important role in the interaction of ligands with P-gp.

Table 3.1: 2D Dragon descriptors selected for the model.

Symbol	**Definition**	**Block Description**
H%	percentage of H atoms	Constitutional indices
nR07	number of 7-membered rings	Ring descriptors
D/Dtr11	distance/detour ring index of order 11	Ring descriptors
MWC01	molecular walk count of order 1	Walk and path counts
X2A	average connectivity index of order 2	Connectivity indices
SIC3	Structural Information Content index (neighbourhood symmetry of 3-order)	Information indices
VE1sign_B(s)	coefficient sum of the last eigenvector from Burden matrix weighted by I-state	2D matrix-based descriptors
ATSC7m	Centred Broto-Moreau autocorrelation of lag 7 weighted by mass	2D autocorrelations
MATS6v	Moran autocorrelation of lag 6 weighted by van der Waals volume	2D autocorrelations
GATS4s	Geary autocorrelation of lag 4 weighted by I-state	2D autocorrelations
P_VSA_LogP_3	P_VSA-like on LogP, bin 3	P_VSA-like descriptors
P_VSA_ppp_D	P_VSA-like potential pharmacophore points, D-hydrogen-bond donor	P_VSA-like descriptors
nRCOOR	number of esters (aliphatic)	Functional group counts
nArCONHR	number of secondary amides (aromatic)	Functional group counts
nArCO	number of ketones (aromatic)	Functional group counts
H-048	H attached to C2(sp3)/C1(sp2)/C0(sp)	Atom-centred fragments
SdsCH	Sum of dsCH E-states	Atom-type E-state indices
CATS2D_01_DN	CATS2D Donor-Negative at lag 01	CATS 2D
CATS2D_05_PP	CATS2D Positive-Positive at lag 05	CATS 2D
CATS2D_02_PL	CATS2D Positive-Lipophilic at lag 02	CATS 2D
B07[O-F]	Presence/absence of O - F at topological distance 7	2D Atom Pairs
F01[C-C]	Frequency of C - C at topological distance 1	2D Atom Pairs
F02[C-O]	Frequency of C - O at topological distance 2	2D Atom Pairs
F04[C-P]	Frequency of C - P at topological distance 4	2D Atom Pairs
F04[C-Br]	Frequency of C - Br at topological distance 4	2D Atom Pairs
F07[O-F]	Frequency of O - F at topological distance 7	2D Atom Pairs

The selected set of descriptors also includes one simple constitutional descriptor, i.e., the percentage of H atoms, as well as some relevant indices, such as the average connectivity index and the information index, which give information about the shape of the molecule, and the atom-type E-state index that provides topological and electronic information related to specific atom types in the molecule.

3.2.4.2 Classification Model

Sensitivity, specificity, and precision are commonly used to evaluate classifiers, but in the case of multiclass problems they do not give a global evaluation of the classification quality; they only give information about the classifier performances on each specific class. Therefore, global indices derived from primary class measures have been proposed (Ballabio *et al.*, 2018), namely sensitivity and precision. The average sensitivity, also known as non error rate (NER), and average precision (AvPr) are calculated as the arithmetic mean of the sensitivity and precision values of the G classes. An analogous global measure based on specificity values has never been proposed, probably due to the specificity bias in relation to the number of classes G.

$$NER = \frac{\sum_{g=1}^{G} Sn_g}{G} \tag{3.7}$$

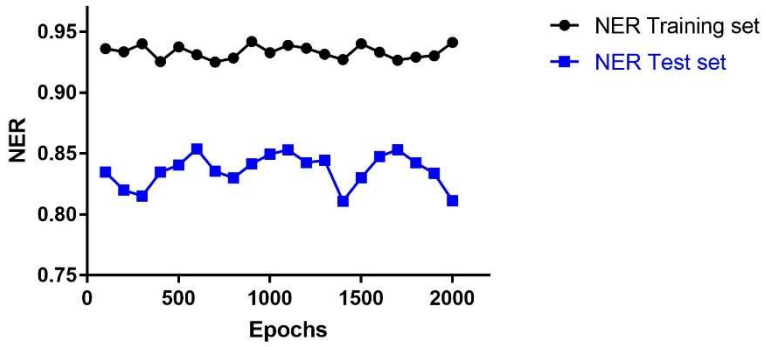
$$AvPr = \frac{\sum_{g=1}^{G} Pr_g}{G} \tag{3.8}$$

It is known that accuracy, one of the most commonly used classification indices in the literature, is affected by the presence of unbalanced classes, as in this case; it is biased towards the most numerous class and for this reason it is not considered in this study to evaluate the performance of the classifier.

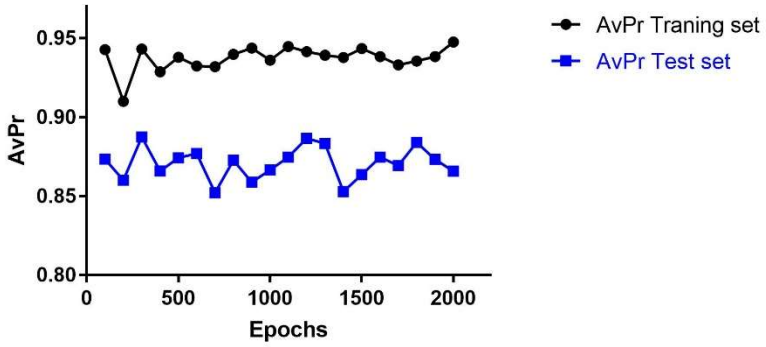
Different models were built modifying the network size, from 20 by 20 to 45 by 45 neurons, and the number of learning epochs from 100 to 2000. The minimum and maximum learning rates were set to 0.001 and 0.6, respectively. The best classification performance was obtained for models with a network dimension of 43 by 43 neurons.

The statistical performance of the models with dimension 43 by 43 neurons as a function of the number of learning epochs is shown in Figure 3.6. The highest NER value for the Test set (TE) ($NER_{TE} = 0.85$) was obtained using 600 epochs, while the highest AvPr value for the TE set ($AvPr_{TE} = 0.88$) was obtained using 300 epochs; the second highest AvPr value (0.87) was obtained using 600 epochs; therefore, the optimal number of learning epochs for the model was set at 600, since the NER value at 300 epochs has the minimum value in the curve for the TE set.

After optimizing the models, the model with the best classification ability was selected. The statistical performance of this model is shown in Table 3.2, the confusion matrix in Table 3.3 and the network parameters in Table 3.4. The global performance of the model based on the training set (TR) showed a NER value of 93.10% and AvPr of 93.22%. For the TE set, the NER was 85.37% and AvPr was 87.68%. For the external validation set, the NER and AvPr values were 70.21% and 70.08%, respectively, demonstrating a good classification performance.



(a)



(b)

Figure 3.6: Statistical performance of models with dimension 43 × 43 neurons as a function of the number of learning epochs: (a) NER versus number of Epochs; (b) AvPr versus number of Epochs.

Table 3.2: Statistical performance of the selected model.

Global indices	**Training set**			**Test set**			**Validation set**		
NER	0.93			0.85			0.70		
AvPr	0.93			0.87			0.70		
	Training set			**Test set**			**Validation set**		
	I[1]	**S**[2]	**NA**[3]	**I**	**S**	**NA**	**I**	**S**	**NA**
Sensitivity	95.1	92.5	91.6	90.5	79.7	85.8	75.9	65.2	69.4
Specificity	95.6	97.8	96.7	91.2	97.4	92.1	83.0	90.7	83.2
MCC	0.90	0.89	0.88	0.79	0.78	0.76	0.63	0.54	0.52

[1] Inhibitor. [2] Substrate. [3] Non-active.

Table 3.3: Confusion matrix for the Validation set of 385 compounds.

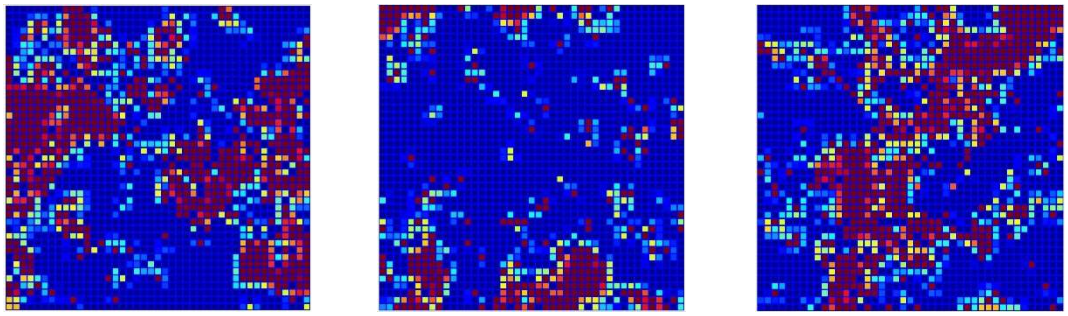
		Experimental class	
		Inhibitors	**Non-Inhibitors**
Predicted class	**Inhibitors**	136	35
	Non-Inhibitors	43	171
		Substrates	**Non-Substrates**
	Substrates	47	29
	Non-Substrates	25	284
		Non-Active	**Active**
	Non-Active	93	42
	Active	41	209

Looking at the performance of the classifier in each specific class, the results for the sensitivity were quite good with values above 91% in the TR set and above 80% in the TE set. The specificity was also good with values of over 90% in both sets, TR and TE. In addition, the MCC values showed the capability of the model for classifying both positive and negative objects, with values greater than 0.77 in the TR and TE sets. The model was challenged using the V set, which provided sensitivity values greater than 65% and specificity values greater than 83%. The MCC values were around 0.50 indicating a good predictive performance of the model.

Table 3.4: CP-ANN parameters for the selected model.

Parameters of CP-ANN			
Network dimension	**Learning epochs**	**Max. Learning rate**	**Min. Learning rate**
43×43	600	0.6	0.001

Figure 3.7 shows the distribution of the 1,786 structurally diverse compounds of the TR set in the Kohonen top-map. The separation of P-gp substrates, inhibitors and non-active compounds is quite good. P-gp inhibitors are mainly clustered on the right and left sides of the network, while P-gp substrates are clustered in the upper and inner parts of the top-map. P-gp non-active compounds are located in the central part of the network drawing a diagonal from the inner left part to the upper right part of the top-map; however, some of the compounds from each class are located outside the formed clusters. A complete separation of P-gp substrates, inhibitors, and non-active compounds was not possible even at larger dimensions of the network due to the high structural similarities between compounds belonging to different classes.



(a) (b) (c)

Figure 3.7: Distribution of objects in the Kohonen top-map. Neurons coloured red represent the position of the objects belonging to the corresponding class: (a) Inhibitors, (b) Substrates, and (c) Non-active compounds.

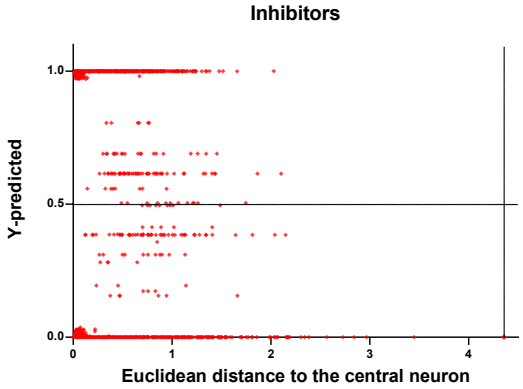
3.2.4.3 Applicability Domain of the Model

In this study, the AD of the model was analysed using the ED between the molecules and the central neuron of the neural network. This metric gives the possibility to compare the TR and TE sets chemical coverage in terms of the incorrectly predicted.

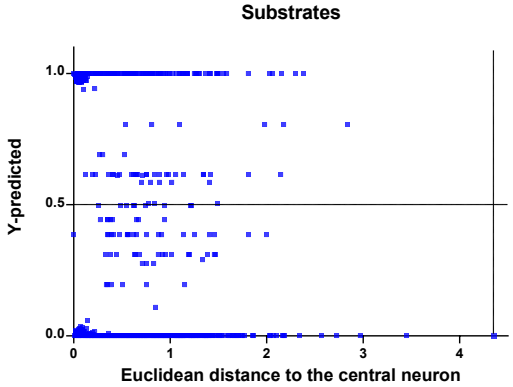
In the developed model, the target value was set to 1 for compounds belonging to a class and zero for compounds not belonging to that class. The predicted response values are expressed as continuous values in the interval between 0 and 1. Therefore, the threshold for separating the objects that belong or do not belong to a particular class was set equal to 0.5. If a compound has a predicted value > 0.5, then the compound is classified as belonging to the class under study; if it is < 0.5, then it is classified as not belonging to the class under study. The data closer to the threshold can be determined as correctly predicted by our model, but they are less reliable. The area closer to the threshold is the uncertainty area, because here the results of the model contain a very high uncertainty in the prediction.

In Figures 3.8–3.10, the ED to the central neuron is plotted against the predicted values in each class, for the TR, TE and V sets. The ED here indicates the similarity or dissimilarity between the compounds. The maximum value of the ED characterizes the boundaries of the model under study. In these graphs, the area near the threshold (0.5) represents the uncertainty zone of the prediction. The chemicals predicted as false or with high level of uncertainty are grouped here. Therefore, the predicted P-gp inhibitors (Figures 3.8.a, 3.9.a, 3.10.a), substrates (Figures 3.8.b, 3.9.b, 3.10.b), and non-active compounds (Figures 3.8.c, 3.9.c, 3.10.c) closer to the edge of the classes (Y = 1) and (Y = 0) are considered to have better prediction accuracy than those in the middle, near the threshold 0.5.

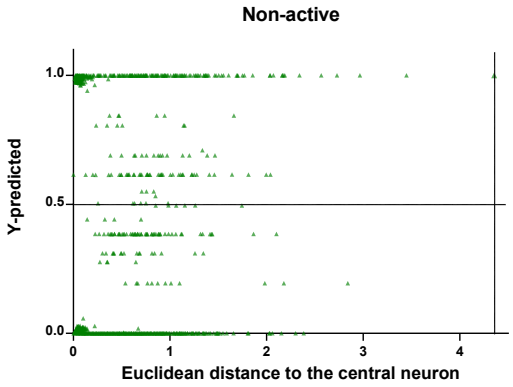
In Figures 3.8 and 3.9, each graph has a vertical dashed line indicating the maximum values of ED obtained in the TR and TE sets, respectively. In the TR set, the EDs were mostly kept within low values. Therefore, the largest ED distance to the central neuron (5.45) obtained with the TE set is the distance considered for the analysis in this discussion. In Figure 3.10, which corresponds to the results of the V set, the dashed line has been placed on the ED of 5.45, since it is the largest distance obtained in the construction of the model and it can be used as a reference point for a better analysis of the boundaries of the model.



(a)

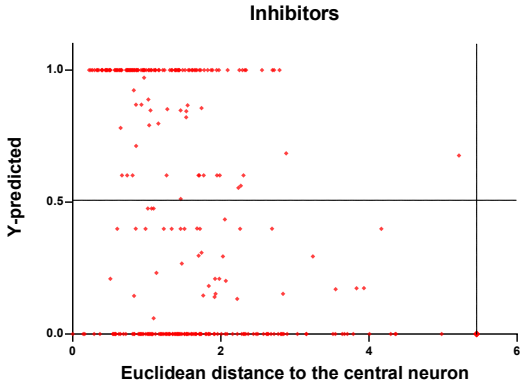


(b)

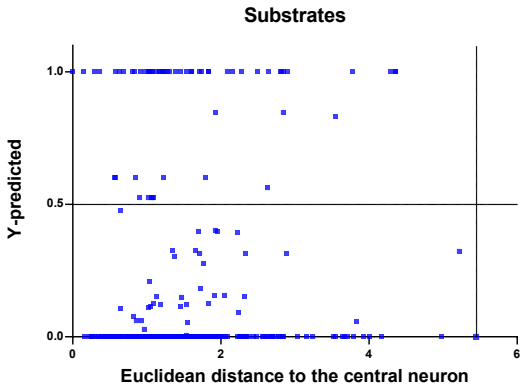


(c)

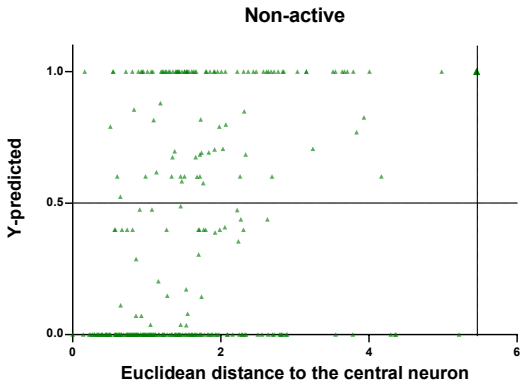
Figure 3.8: Qualitative assessment of the applicability domain for the selected model: Training set (TR): (a) Inhibitors, (b) Substrates and (c) Non-active.



(a)

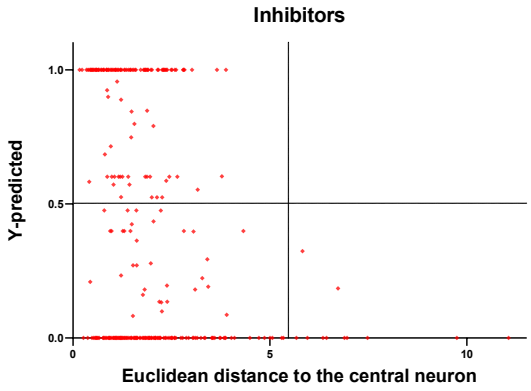


(b)

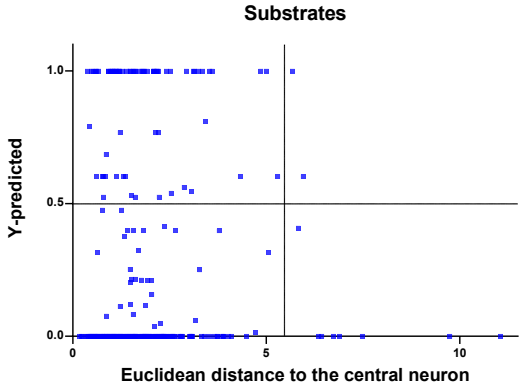


(c)

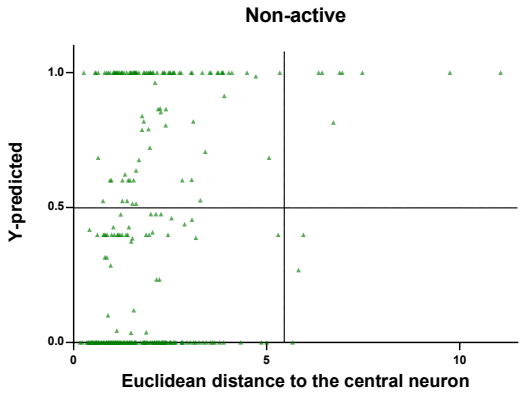
Figure 3.9: Qualitative assessment of the applicability domain for the selected model: Test set (TE): (a) Inhibitors, (b) Substrates and (c) Non-active.



(a)



(b)



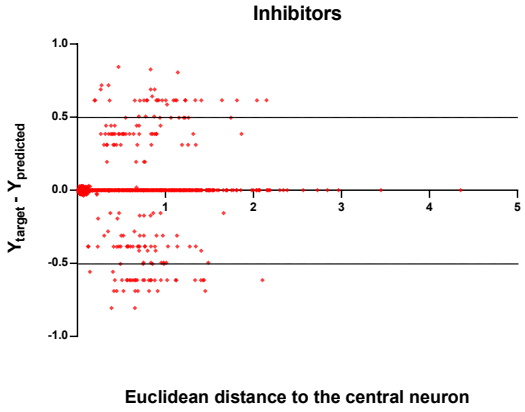
(c)

Figure 3.10: Qualitative assessment of the applicability domain for the selected model: Validation set (V): (a) Inhibitors, (b) Substrates and (c) Non-active.

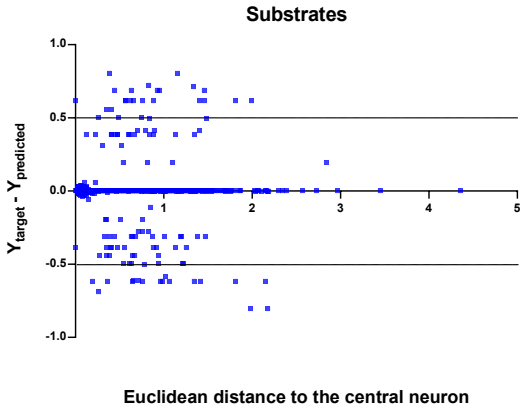
Looking at the descriptor values of the compounds with the largest ED to the central neuron in the TE, it can be noticed that one count descriptor, one atom-centred fragment descriptor (F04[C-P], H-048) and two autocorrelation descriptors (MATS6v, GATS4s) have the highest values in the dataset. The two compounds with the largest

ED in the TE set are phosmet and triphenylphosphane, both compounds share the presence of a phosphorous atom in their structure (see Appendix A, Figure A.2), which can explain the high values for the descriptor Frequency of C – P at topological distance 4. Nevertheless, the large values of ED in the model are not evidence of incorrect prediction.

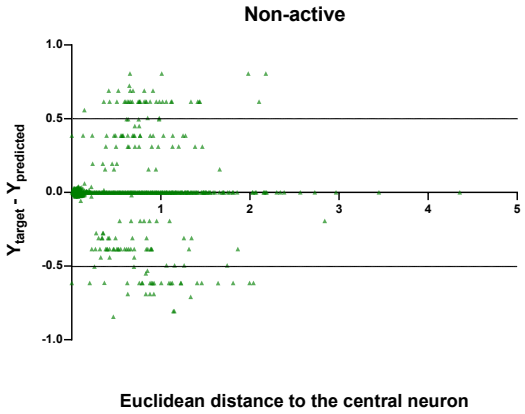
In Figures 3.11–3.13, the graphs show the distribution of the true predicted and false predicted compounds relative to the threshold 0.5. All values above and below the threshold are the false predicted by the model. In the plots it can be noted that some of the false predicted compounds are within the shortest ED to the central neuron.



(a)

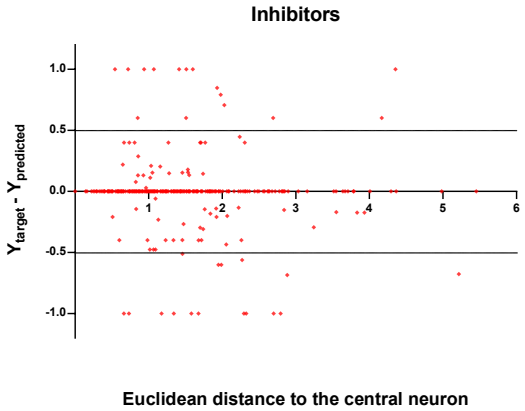


(b)

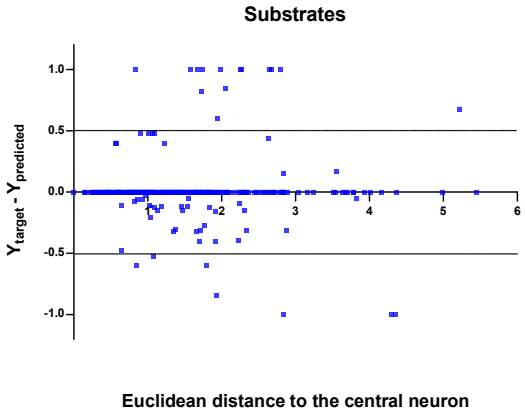


(c)

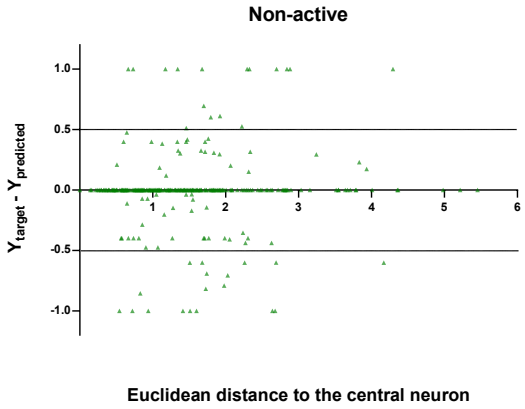
Figure 3.11: Plot of the EDs to the central neuron versus (Ytarget–Ypredicted) for the Training set (TR): (a) Inhibitors, (b) Substrates and (c) Non-active.



(a)

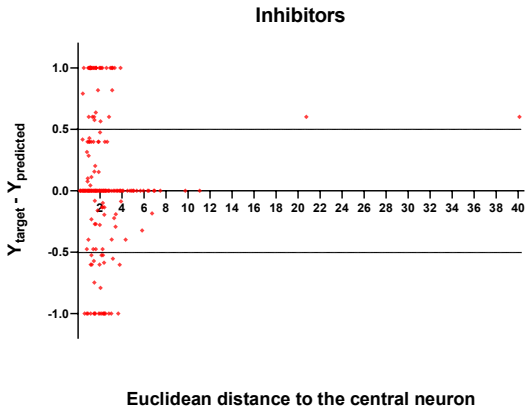


(b)

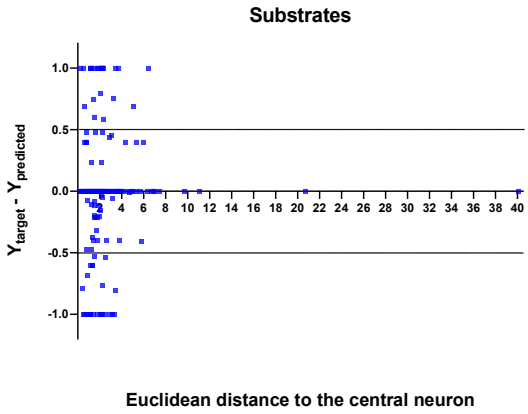


(c)

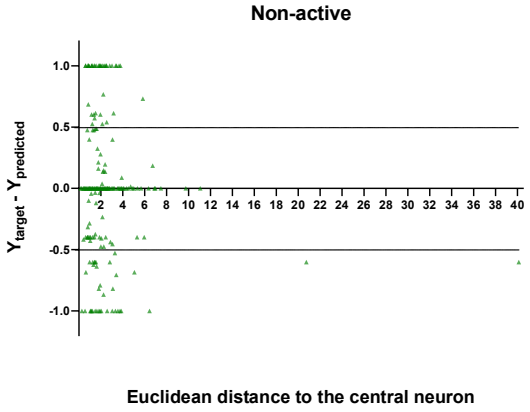
Figure 3.12: Plot of the EDs to the central neuron versus (Ytarget—Ypredicted) for the Test set (TE): (a) Inhibitors, (b) Substrates and (c) Non-active.



(a)



(b)



(c)

Figure 3.13: Plot of the EDs to the central neuron versus (Ytarget – Ypredicted) for the Validation set (V): (a) Inhibitors, (b) Substrates and (c) Non-active.

In the V set, thirteen compounds have ED, which are greater than the ED value taken as the reference point. These thirteen compounds have in common the presence of highly electronegative atoms in their structure such as chlorine, fluorine, bromine and sulphur (see Appendix A, Figure A.3). However, only four of the thirteen were incorrectly predicted by the model. This could be due to the uncertainties present in the predictions. Some predictions may fall within the defined AD of the model but be unreliable due to properties and features not accounted for by the model. On the other hand, a chemical that falls outside the defined AD may still exhibit the modelled response because it brings out this response by a mechanism not accounted by the model under study.

In these graphs, most correctly predicted compounds are concentrated within the shorter ED, suggesting that this is the region in which the predictions are expected to be reliable.

3.2.5 Conclusions

In summary, a multi-class classification model was developed for P-glycoprotein inhibitors, substrates, and non-active compounds, which provided a good classification performance as evidenced by the values of the global indices non error rate (NER) and average precision (AvPr) for the training (TR), test (TE), and validation (V) sets. Therefore, the presented classifier could be used as a reliable *in silico* screening tool to identify potential ligands of P-glycoprotein (P-gp).

Unlike currently available classifiers, the one developed here not only separates active/non-active compounds, but also distinguishes between substrates and inhibitors. This is an important detail that would extend the use of the multi-class classifier for different purposes. The fast-screening tool would be used as an initial step in evaluating a set of molecules of interest, even before implementing a molecular modelling method (e.g., molecular docking and molecular dynamics simulations, which are more demanding and time consuming). After the initial fast screening, the molecules showing a desirable response would be subjected to further detailed studies, including final experimental testing of their interactions with the P-gp.

Chapter 4

Structure-Based Modelling Approach[2]

4.1 Introduction

In recent years, significant improvements in proteomics and chemical genomics (Scapin, 2006), protein chemistry, structure elucidation and refinement techniques have led to an exponential increase in the number of available three-dimensional (3D) structures of proteins. Structure-based drug design approaches depend on the knowledge of the 3D structure of the biological target; therefore, a greater availability of 3D structures of proteins has enabled an increase in the application of this method. The 3D structure of the targets can be obtained by methods such as X-ray crystallography, NMR spectroscopy, or cryo-electron microscopy (cryo-EM) (Verlinde *et al.*, 1994). However, if no experimental structure of the target is available, it is also possible to build a homology model of the target based on the experimental structure of related proteins.

Current methods for structure-based drug design can be divided into three main categories: *virtual screening*, which consists in identifying new ligands in large databases of 3D structures of small molecules, with the goal of finding those that fit into the binding pocket of the receptor, using fast, approximate docking programs. The second method, *de novo* design of new ligands, attempts to build a new structure directly from the analysis of the binding site; ligand molecules are constructed by taking into account the constraints of the binding pocket, assembling small pieces step by step. These pieces can be either single atoms or molecular fragments. The third method consists in optimising known ligands by evaluating proposed analogs within the binding pocket (Klebe, 2000).

Structure-based methods are useful for predicting binding modes of small molecules and their relative affinities (Śledź *et al.*, 2018). They can be used to identify ligands by high-throughput docking of small molecules to a rigid target protein, and scoring the hits obtained based on an implicit solvent force field. Alternatively, one can use a low-throughput explicit solvent technique, such as molecular dynamics (MD), to characterise flexible binding sites and accurately evaluate binding pathways, kinetics, and thermodynamics. In this chapter, we will focus on the molecular docking approach.

In general, molecular docking refers to the prediction of ligand conformation and orientation (pose) within the target binding site, with the aim of obtaining an accurate structural modelling and a correct prediction of the activity. The docking algorithms pose

[2] The chapter is based on the author's manuscript: Mora Lagares, L., Minovski, N., Caballero Alfonso, A. Y., Benfenati, E., Wellens, S., Culot, M., . . . Novič, M. (2020). Homology modeling of the human P-glycoprotein (ABCB1) and insights into ligand binding through molecular docking studies. *International journal of molecular sciences, 21*(11), 4058. doi: 10.3390/ijms21114058

small molecules in the active site, which is challenging due to the conformational degrees of freedom of the molecules; they must be accurate in order to identify the ligand conformation that best matches the receptor structure, and fast enough to allow the evaluation of thousands of compounds in one docking run. Docking algorithms are complemented by scoring functions that rank the poses and provide a theoretical estimate of the binding affinity, allowing the prediction of the more active compounds (Cole *et al.*, 2005; Kitchen *et al.*, 2004). The scoring functions are based on approximations and simplifications and do not fully account for some physical phenomena that govern molecular recognition, such as entropic effects, when scoring the modelled complexes; for this reason, consensus scoring schemes were introduced to address the limitations of the scoring functions. Consensus scoring combines information from different scores to compensate for the errors in individual scores and improve the probability of identifying ligands (J.-M. Yang *et al.*, 2005).

The main advantage of the structure-based methods is the possibility to explore new structural prototypes and large virtual libraries in a rather short time. They serve as a good filter to remove inactive compounds or to distinguish between more and less active compounds. Nevertheless, there are some limitations associated with these techniques; for example, not all hit compounds may be present in a physical compound collection, which slows down the possibility of evaluation in biochemical assays; some molecules cannot be readily synthesised. When structural information is available, *de novo* generation and molecular docking can be used either as an alternative to, or in parallel with, traditional high-throughput screening methods to predict which compounds will have affinity for a particular target (Moro *et al.*, 2007).

In the study of P-gp, various structure-based approaches have been used to classify and understand ligand–P-gp interactions. Prior to the publication of the *m*P-gp structure in 2009 (Aller *et al.*, 2009), structural studies of *h*P-gp relied heavily on homology models based on bacterial transporters (Palmeira *et al.*, 2012). The access to an X-ray structure of P-gp, albeit not at perfect resolution, represented a major advance for structure-based studies on this transporter. In recent years, the number of available 3D structures of ABC proteins (Klepsch *et al.*, 2010) and the power of experimental approaches (Winter *et al.*, 2008), have facilitated the application of structure-based methods to predict ligand–transporter interactions. For example, Bikadi *et al.* (Bikadi *et al.*, 2011) built a free web server for predicting P-gp substrate activity and binding modes based on a support vector machine (SVM) classification model and molecular docking. They used molecular docking into the crystal structure and homology model of *m*P-gp to generate protein–ligand complexes of the predicted compounds, but not for classification purposes. On the other side, Dolghih *et al.* (Dolghih *et al.*, 2011) used induced fit docking into the crystal structure of *m*P-gp to separate P-gp binders from non-binders based on their docking score. Similarly, Chen *et al.* (L. Chen *et al.*, 2012) used a set of 245 P-gp substrates and non-substrates to evaluate the predictive ability of docking, and more recently, Kazemi *et al.* (Kazemi *et al.*, 2021) used the *m*P-gp structure to evaluate the activity of some lignanamides from *Cannabis sativa* against P-gp using molecular docking. The 3D structure of *h*P-gp in the active conformation (open to the cytoplasm) was recently solved with cryo-electron microscopy (Alam *et al.*, 2019), fact that will greatly improve the structure-based studies of this interesting membrane transporter.

In this chapter, we present a detailed report of a structure-based approach applied to the modelling of P-glycoprotein–ligand binding interactions, using molecular docking tools and a homology model of the human membrane transporter.

4.2 Homology Modelling of the Human P-glycoprotein (ABCB1) and Insights into Ligand Binding through Molecular Docking Studies

4.2.1 Abstract

The ABCB1 transporter also known as P-glycoprotein (P-gp) is a transmembrane protein belonging to the ATP binding cassette super-family of transporters; it is a xenobiotic efflux pump that limits intracellular drug accumulation by pumping the compounds out of cells. P-gp contributes to a decrease of toxicity and possesses broad substrate specificity. It is involved in the failure of numerous anticancer and antiviral chemotherapies due to the multidrug resistance (MDR) phenomenon, where it removes the chemotherapeutics out of the targeted cells. Understanding the details of the ligand–P-gp interaction is therefore crucial for the development of drugs that might overcome the MDR phenomenon and for obtaining a more effective prediction of the toxicity of certain compounds.

In this work, an *in silico* modelling using homology modelling and molecular docking was performed with the aim of better understanding the ligand–P-gp interactions. Based on different mouse P-gp structural templates from the PDB repository, a 3D model of the human P-gp (*h*P-gp) was constructed using protein homology modelling. The homology model was then used to perform molecular docking calculations on a set of thirteen compounds, including some well-known compounds that interact with P-gp as substrates, inhibitors, or both. The sum of ranking differences (SRD) was employed to compare the different scoring functions used in the docking calculations. A consensus-ranking scheme was used to select the top-ranked pose for each docked ligand.

The docking results showed that a high number of π interactions, mainly π–sigma, π–alkyl, and π–π interactions, together with the simultaneous presence of hydrogen bond interactions contribute to the stability of the ligand–protein complex in the binding site. It was also observed that some interacting residues in *h*P-gp are the same when compared to those observed in a co-crystallized ligand (PBDE-100) with mouse P-gp (PDB ID: 4XWK). Our *in silico* approach is consistent with the available experimental results regarding the P-gp efflux transport assay; therefore, it could be useful in predicting the role of new compounds in systemic toxicity.

4.2.2 Introduction

The ATP-binding cassette (ABC) transporter ABCB1, known as P-glycoprotein (P-gp) is a transmembrane efflux transporter with a broad substrate specificity that limits intracellular drug accumulation and contributes to a decrease of toxicity. It is present in normal tissues linked to excretory or barrier functions, as well as in tumour cells, where it is responsible for resistance to a large variety of chemotherapeutic drugs; a phenomenon known as the multidrug resistance (MDR) (Nobili *et al.*, 2012). The structure of this transporter consists of two symmetrical and homologous halves that act in a coordinated manner as a unit, each with six transmembrane domains (TMD) and a nucleotide binding domain (NBD) located on the cytosolic surface (Fardel *et al.*, 1996) responsible for ATP binding and hydrolysis.

P-gp can interact with large numbers of structurally diverse compounds, which according to their interactions can be classified as substrates, inhibitors, and modulators (L. Chen *et al.*, 2012). Compounds actively transported by P-gp are known as substrates,

whereas those that compromise the transporting function of P-gp are known as inhibitors. Modulators interact with P-gp reducing substrate binding through a negative allosteric modulation.

Studies have been done trying to understand the nature of the ability of P-gp for binding so many different compounds and to elucidate the attributes of the drug binding pocket. A study conducted using photoaffinity labelling of P-gp with azidopine showed that there are two distinct binding sites for this drug (Bruggemann *et al.*, 1989). In subsequent years, three different binding sites were suggested: the H-site interacting with Hoechst 33342 and colchicine, the R-site interacting with rhodamine 123 (R123) and anthracyclines, and a third binding site exerting an allosteric interaction with the previous two (Safa, 2004); over time, the number of binding sites has increased up to seven (Safa, 2004). In addition, the mechanism of "substrate induced-fit" has also been proposed, which states that a substrate, depending on its size and shape, is able to induce conformational changes in the transmembrane segments (TM), allowing the substrate to accommodate within P-gp and successively be transported (Loo *et al.*, 2003b).

Due to the importance of P-gp in MDR and absorption, distribution, metabolism, excretion, and toxicity (ADMET), many studies have been conducted with the aim of identifying P-gp substrates and developing more effective P-gp inhibitors (Polli *et al.*, 2001). For this purpose, *in silico* models are recognized as valuable tools (S Ekins *et al.*, 2007; Van De Waterbeemd *et al.*, 2003) and the methodologies used are either ligand-based or structure-based prediction methods (Bikadi *et al.*, 2011).

Quantitative structure–activity relationship (QSAR), a traditional ligand-based method, has been used extensively in predicting the biological activity providing rapid and cost-effective screening platforms for identifying P-gp inhibitors or substrates. A P-gp classification model was also developed in the authors' previous study (Mora Lagares, Minovski, & Novič, 2019). On the other hand, structure-based methods (e.g., molecular docking) allow the investigation of ligand–receptor interactions at the atomistic level when high-resolution structures of the receptors are available. Until 2019, when the cryoEM structure of human ABCB1 was solved (PDB ID: 6QEX) (Alam *et al.*, 2019), docking studies on P-gp to understand binding site interaction profiles were limited due to the availability of the experimentally solved structure of human P-gp (*h*P-gp) (Becker *et al.*, 2009; Klepsch *et al.*, 2011; Sirisha *et al.*, 2011), necessitating the use of homology models for studying ligand–*h*P-gp interactions.

The first homology models developed and utilized in molecular docking studies relied on bacterial homologues used as templates, such as the bacterial transporters Sav1866 and MsbA structures (O'Mara *et al.*, 2007; Pleban *et al.*, 2005; Ravna *et al.*, 2007), representing different catalytic states of the transport cycle. In 2009, the crystal structure of the mouse P-gp (*m*P-gp) in complex with a cyclic tetrapeptide (PDB ID: 3G5U) (Aller *et al.*, 2009) was resolved representing a ligand binding competent conformation of the protein. With 87% sequence identity, *m*P-gp is a well-suited template for homology modelling of the *h*P-gp and provides a better model for structure-based approaches.

In the present study, we developed a *h*P-gp homology model based on *m*P-gp multiple templates that can be used in further docking simulations. We used the available knowledge on the interaction of substrates and inhibitors with P-gp to apply a molecular docking approach, first, to clarify whether molecular docking is able to discriminate between active and non-active P-gp compounds; and second, to determine the extent to which the amino acids predicted by molecular docking are consistent with the available experimental data. To this end, we performed molecular docking simulations on a set of thirteen compounds belonging to the above-mentioned classes, and the homology model

of *h*P-gp. Ligand–protein binding energies, number and type of interactions were analysed to assess whether there are significant differences between the compounds under study.

4.2.3 Materials and Methods

4.2.3.1 Protein Homology Modelling

The 3D protein homology model of *h*P-gp was constructed using three different tools, SWISS-MODEL (Schwede *et al.*, 2003), I-TASSER (Roy *et al.*, 2010; J. Yang *et al.*, 2015; Y. Zhang, 2008) and Discovery Studio 4.1/Modeler 9.12 (Accelrys, 2017; Šali *et al.*, 1993). The complete *h*P-gp protein sequence, which consists of 1,280 amino acids, was retrieved from the UniProtKB database (accession number P08183).

4.2.3.1.1 Template Selection and Alignment The selection of the templates was based on sequence similarity with known protein structures (homologous) from the Protein Data Bank (PDB) repository. The SWISS-MODEL and Discovery Studio 4.1 protocols identified suitable templates based on BLAST (Camacho *et al.*, 2009), while I-TASSER structure templates were identified using LOMETS (S. Wu *et al.*, 2007; Zheng *et al.*, 2019).

The target and template sequences were aligned in order to analyse the sequence conservation. Insertions and deletions were made to obtain the best alignment. In general, it is preferable to include more than one template in the alignment, as this might allow a better fitting of regions where the percentage of identity using a single template is very low.

4.2.3.1.2 Model Generation The overall structure of the *h*P-gp (based on full-length sequence as retrieved from UniProtKB accession number P08183) was modelled including the nucleotide binding domains (NBDs) and the flexible linker region. Since the quality of the constructed *h*P-gp model is directly dependent on the quality of the template used, the complete primary sequence and secondary structure information related to the *h*P-gp were used, including NBDs. In addition, the linker region appears to be important for the stabilization of NBDs, acting as a "damper" that reduces the movements of the cytoplasmic regions of P-gp (Ricardo J. Ferreira *et al.*, 2013a); therefore, it was also included in the model.

In general, the steps for generating a protein homology model involve the creation of the target backbone by copying the coordinates of the template-backbone to the target. When the residues are identical, also the protein side-chain coordinates are copied. The gaps in alignment due to insertions and deletions are modelled by loop modelling. The side chains can be built by searching every possible conformation for every torsion angle of the side chain and selecting the one that has the lowest interaction energy with neighbouring atoms. A rotamer library can also be used for this purpose, which has all the favourable side chain torsion angles extracted from known protein crystal structures. Finally, the geometry of the resulting model is minimized by using a force field.

The various tools used here to model the 3D structure of *h*P-gp differ in the algorithms used for building the model and in the methods utilized for the model refinement.

4.2.3.1.2.1 SWISS-MODEL This tool extracts the initial structural information from the template structure. Insertions and deletions are resolved looking for viable candidates in a structural database. Final candidates are selected using statistical

potentials of mean force scoring methods. If no candidates can be found, then a conformational space search is performed using Monte Carlo conformational techniques. Non-conserved side chains are modelled using an in-house backbone-dependent rotamer library. The optimal configuration of rotamers is estimated using the graph-based TreePack algorithm (Xu *et al.*, 2005) by minimising the SCWRL4 energy function (Krivov *et al.*, 2009). As a final step, small structural distortions, unfavourable interactions or clashes introduced during the modelling process are resolved by energy minimisation. SWISS-MODEL uses CHARMM27 force field for parameterisation.

SWISS-MODEL assesses the quality of the model through the GMQE (Global model quality estimation) and the QMEAN (Benkert *et al.*, 2011) score. The GMQE has values between 0 and 1, reflecting the accuracy of the model built with that specific alignment and template. The higher the number, the higher the reliability of the model. The QMEAN score indicates the degree of nativeness of the structure in the model. Values around 0 mean good quality agreement between the modelled structure and experimental structures of similar size. Values less than -4 indicate models of low quality. In addition to the previous scores, the expected similarity to the native structure for each residue in the model can be checked through the Local Quality plot. Usually, residues showing a score below 0.6 are expected to be of low quality.

4.2.3.1.2.1 I-TASSER This tool uses fragments excised from the PDB templates, reassembles them into full-length models using Monte Carlo simulations, and builds the loops using ab initio modelling. The large ensemble of structural conformations, called decoys, is then clustered by the SPICKER (Y. Zhang *et al.*, 2004b) program based on pairwise structural similarity. The final full-atomic models are obtained with REMO (Y. Li *et al.*, 2009), which builds up the atomic details from the selected I-TASSER decoys by optimizing the hydrogen-bonding network; for this purpose, it uses the CHARMM22 force field parameters. Five models corresponding to the five largest structural clusters are reported.

For assessing the global accuracy of the model, I-TASSER employs the C-Score, the TM-Score (Y. Zhang *et al.*, 2004a), and the RMSD. The C-Score is a confidence score calculated based on the significance of the threading templates alignments and the convergence parameters of the structure assembly simulations. It has scores between -5 and 2, with higher scores indicating a high confidence model and vice versa. The TM-score and the RMSD are predicted based on the C-Score because they are highly correlated. The correlation coefficient of the C-score of the first model with TM-score and RMSD is 0.91 and 0.75, respectively. The TM-score is a measure of the structural similarity between the predicted model and the native structure; unlike the RMSD, it is insensitive to local modelling error, because it weights the small distance stronger than the large one; in the RMSD, a local error leads to a large RMSD value even if the global topology is correct. The TM-score has values ranging from 0 to 1, i.e., values larger than 0.5 indicate a correct topology of the model, while values smaller than 0.17 indicate a random similarity.

The local accuracy of the model can be visualized in the Estimated Local Accuracy Plot, which shows the distance deviation between the residue positions in the model and the estimated native structure.

4.2.3.1.2.1 Discovery Studio 4.1/Modeler 9.12 Modeler uses restraints on the spatial structure of the amino acid sequence and ligands being modelled. The output is a 3D structure that satisfies these restraints as much as possible. The program automatically derives the restraints only from the known related structures and their

alignment with the target sequence. The restraints can be on distances, angles, dihedral angles, pairs of dihedral angles and some other spatial features. During the refinement of the model, conjugate gradient and simulated annealing molecular dynamics (MDs) are used to optimize the positions of heavy atoms. Modeler utilizes the CHARMM22 force field for parameterisation.

Discovery Studio uses the Verify Score to evaluate the validity of the modelled 3D structure, which measures the compatibility of each residue in the current 3D environment. The Verify Expected High and Low score for a protein of the same size is given as a reference point. If the calculated Verify Score is greater than the Expected High Score, the structure is likely correct. If, on the other hand, it is lower than the Expected Low Score, the structure is almost certainly misfolded. In general, the closest the Verify Score is to the Verify Expected High Score, the better the quality of the model. Discovery Studio also reports the Probability Density Function (PDF) Energy and the Discrete Optimized Protein Energy (DOPE) (Shen *et al.*, 2006) scores. The lower these values are, the better the model is. The local accuracy of the model can be visualized in the Verify Score Plot, which gives the compatibility score for each residue in the given 3D structure.

4.2.3.1.2.2 Assessment of the Models The validation of the models developed was performed using the programs PROCHECK (Laskowski *et al.*, 1993), VERIFY 3D (Bowie *et al.*, 1991), ERRAT (Colovos *et al.*, 1993), and PROVE (Pontius *et al.*, 1996), which are available at the Structural Analysis and Verification Server (SAVEs) (UCLA-DOE).

PROCHECK was used for assessing the stereochemical quality of the protein structure. It checks conformation of the protein backbone by analysing the torsion angles phi (φ) and psi (ψ) of the amino acid residues in the modelled protein using the Ramachandran plot.

VERIFY 3D program analyses the compatibility of the assembled atomic model (3D) with its corresponding amino acid primary sequence (1D). It classifies each residue in the protein into one of the 18 classes according to the structural environment of the residue in the input model. The propensity of each amino acid to exist in each structural environment class is calculated according to statistics collected from structures in the PDB repository. The final score given to the protein structure is the sum of propensities of the individual residues. If at least 80% of the amino acids in the 3D/1D profile have a score greater than or equal to 0.2, the test is considered passed.

ERRAT is an algorithm that analyses the statistics of non-bonded interactions between different atom types (CC, CN, CO, NN, NO, and OO). The ERRAT score is expressed as the percentage of the protein for which the calculated error value falls below the 95% rejection limit. Good high-resolution structures generally produce values around 95% or higher. For lower resolutions (2.5–3.0 Å) the average overall quality factor is around 91%. The generally accepted ERRAT score for considering a model of good quality is at least 50% of the structure below the 95% confidence limit. This is an especially useful tool for assessing the reliability of a model.

PROVE calculates the volumes of atoms in macromolecules using an algorithm that treats the atoms as hard spheres and calculates a statistical Z-score deviation of the model from highly resolved and refined structures deposited in the PDB. If the percentage of buried atoms in the structure is less than 1%, the test is considered passed, otherwise, if it is between 1% and 5%, a warning is issued for the structure. When the percentage of buried atoms is greater than 5%, it is considered that there are some errors in the structure.

4.2.3.2 Molecular Docking Calculations

Molecular docking calculations were performed on a set of thirteen compounds, including eight well-known molecules that interact with P-gp as substrates, inhibitors, or both: cyclosporine A (CsA), a high-affinity substrate (Saeki *et al.*, 1993) and inhibitor (Wigler, 1999) of P-gp, amiodarone (AM), a known substrate (Jouan *et al.*, 2016) of P-gp, doxorubincin (DOX), a substrate of P-gp (Gao *et al.*, 2001; Takara *et al.*, 1999), digoxine (DIG), a high-affinity P-gp substrate (Taipalensuu *et al.*, 2004), loperamide (LPM), a known substrate (Wandel *et al.*, 2002) of P-gp, rifampin (RMP), a substrate (Collett *et al.*, 2004) and inducer (Geick *et al.*, 2001) of P-gp, verapamil (VER), a well-known substrate of P-gp, carvedilol (CAR), a well-known substrate of P-gp (Jouan *et al.*, 2016), and five compounds that do not interact with P-gp: valproic acid (VPA), busulfan (BU), gentamicin (GEN), pamidronate (APD), and paraquat (PQ) (Lacher *et al.*, 2014). We used the available knowledge on the interaction of substrates and inhibitors with P-gp to clarify whether molecular docking is able to discriminate between active and non-active P-gp compounds, and to determine the extent to which amino acids predicted by molecular docking are consistent with the available experimental data.

The ligand–P-gp complexes were built by docking the ligand into the homology model of *h*P-gp and the cryo–EM structure of *h*P-gp (PDB ID: 6QEX) using the Dock Ligands protocol in Discovery Studio 4.1 (Accelrys, 2017). Two algorithms, CDOCKER (G. Wu *et al.*, 2003) and GOLD (G. Jones *et al.*, 1995; G. Jones *et al.*, 1997), were utilized to investigate the binding affinities and conformations of the thirteen compounds. The docking studies were performed without considering the flexible linker region present in the *h*P-gp structure. This region is more than 30 Å away from the binding pocket, therefore it does not appear to be involved in drug binding.

The selection of the binding site and the settings for the molecular docking simulations using the constructed homology model and cryoEM structure of *h*P-gp were validated by the procedure called "self-docking" or "re-docking". The re-docking procedure is a widely used method to validate all docking settings done prior to performing the molecular docking calculations. The quality of all performed structure-based settings (molecular docking) was first checked by docking the co-crystallized ligand PBDE-100 and the cryoEM ligand Taxol into their defined binding pockets (Kirchmair *et al.*, 2008) and comparing them with their experimental conformation, i.e., the reproduction of their spatial conformation and orientation (see Appendix B, Tables B.1, B.2, Figures B.1, B.2). The experimental coordinates of PBDE-100 and Taxol, as well as the surrounding amino acid residues were used to define the binding cavity, for our model and the cryoEM *h*P-gp separately. During the re-docking validation procedure, PBDE-100 (our *h*P-gp model) and Taxol (cryoEM *h*P-gp) were first removed from their binding site and re-docked three times. As an evaluation criteria for a successfully performed re-docking validation, the heavy-atoms RMSD values (RMSD $\leq$ 2 Å) between each pose obtained by re-docking and the natively present ligand (PBDE-100 and Taxol, respectively) were calculated (Verdonk *et al.*, 2003).

After proper validation, each investigated ligand was docked up to ten times with the same docking parameters obtained by the re-docking validation, while the quality of the obtained docking poses was quantified by the –CDOCKER Energy and GoldScore fitness functions, i.e., the main scoring functions of CDOCKER and GOLD, respectively. The resulting docking poses were then re-scored using fourteen scoring functions implemented in Discovery Studio 4.1. A total of two docking runs were performed, each of which was scored using 15 different fitness functions.

The following general steps were followed for the docking simulations:

1. Ligand preparation: The CHARMm force field from the simulation tool was applied to the ligands and a minimization protocol was performed.
2. Protein preparation: The CHARMm force field from the simulation tool was applied to the target protein, a minimization protocol was performed, and the binding site based on ligand coordinates was defined.
3. Running CDOCKER or GOLD protocol.
4. Scoring of docked Ligand Poses.
5. Calculation of ligand Binding Energies.

4.2.3.2.1 Docking with CDOCKER CDOCKER is a grid-based molecular docking method that employs CHARMm force field to dock ligands into a receptor binding site. The receptor is held rigid, while the ligands are allowed to be flexible during the refinement. Random ligand conformations are generated from the initial ligand structure by high temperature molecular dynamics, followed by random rotations. When these conformations are translated to the active site, candidate poses are refined by grid-based (GRID1) simulated annealing and a final grid-based or full force field minimization. CDOCKER uses soft core potentials, which have been shown to be effective in exploring the conformational space of small organics and macromolecules. The non-bonded interactions which involve van der Waals (vdW) and electrostatics are softened at different levels, except during the final minimization step (G. Wu *et al.*, 2003).

Ten conformations for each ligand were generated in the binding site of the *h*P-gp, which was created as a spherical region defined by the atoms within a 15 Å radius around the co-crystallized ligand of *m*P-gp (PDB ID: 4XWK). The selection of the binding site was validated by the re-docking procedure. Random conformations were generated with specific molecular dynamics steps while the system was heated to 700 K in 2,000 steps and then cooled to 300 K in 5,000 steps. The final minimization refinement step was performed using the full potential. The minimized docking poses were then clustered based on a heavy atom RMSD approach. The final ranking was based on the total docking energy, which is composed of the intramolecular energy of the ligand and the ligand–receptor interaction.

4.2.3.2.2 Docking with GOLD GOLD (Genetic Optimization for Ligand Docking) uses a genetic algorithm to explore the full range of ligand conformational flexibility with partial flexibility of the protein active site while searching for favourable ligand poses. A population of chromosomes is manipulated during a genetic algorithm run, with each chromosome representing a trial docking. A chromosome contains all the information necessary to completely define a trial ligand pose and is associated with a fitness value computed from the scoring function. Different values of the genetic algorithm parameters may be used to control the balance between the speed of GOLD and the reliability of its predictions.

Different conformations for each ligand were generated in the binding site of the *h*P-gp, which was created as a spherical region defined by the atoms within a 24.7 Å radius around the co-crystallized ligand of *m*P-gp (PDB ID: 4XWK). The radius of the sphere is significantly higher than in CDOCKER because some of the residues in the binding site are allowed to move. The selection of the binding site was validated by the re-docking procedure.

Flexible docking was performed, meaning that the side chains of some amino acids in the binding site were able to rotate continuously about single bonds during the docking

simulation. Ten residues in the binding pocket were selected to be flexible based on some drug binding residues that have been experimentally reported (Aller *et al.*, 2009; Nicklisch *et al.*, 2016). The selected residues were F303, Y307, Y310, F314, Q725, F728, F732, F759, F983, and Q990 with hydrophobic, aromatic, and polar properties.

4.2.3.2.3 Scoring of Docked Ligand Poses and Calculation of Binding Energies The docked ligand poses obtained from each algorithm were re-scored using various scoring functions through the Score Ligand Poses protocol in Discovery Studio 4.1. Then, the Sum of Ranking Differences (SRD) methodology (Héberger *et al.*, 2011) was used to compare the performance of the computed scoring functions. The SRD is a robust statistical method, specifically developed for method comparison tasks. It evaluates the Manhattan distances of a set of rank-transformed vectors, in this case the different fitness functions, from a reference vector which corresponds to a hypothetical ideal reference method. The reference vector can be a "gold standard" or experimental values, where available, or a consensus method based on data fusion. In this case, we have defined the reference vector as the average value of the scoring functions.

A consensus ranking scheme was used to select the top-ranked pose of each docked ligand. Instead of combining the raw scoring values generated by different scoring functions, the ranks produced by these scoring functions were combined in the following way: first, the rank derived from each scoring function was produced. Then, for a given combination of scoring functions, a fused rank was computed as the geometric mean (Bajusz *et al.*, 2019) of the compound's rank in the individual models. The scoring functions selected for combination were those that showed better performance in the SRD results.

For each ligand, the average binding energy was calculated over the set of related poses. The binding free energies between the *h*P-gp 3D model and the obtained set of docked ligand poses were estimated using the Calculate Binding Energies protocol in Discovery Studio 4.1, in which the binding free energy for a receptor–ligand complex is calculated from the free energies of the complex, receptor, and ligand according to Equation (4.1). The protocol uses CHARMm based energies and implicit solvation methods to estimate the free energies and calculate an estimate of the total binding free energy.

$$Energy_{binding} = Energy_{Complex} - Energy_{Ligand} - Energy_{Recepor} \tag{4.1}$$

4.2.3.3 Caco-2 Pump Out Assay

The interaction with the efflux pump P-gp of nine compounds included in the docking set (CsA, AM, DOX, VER, VPA, BU, GEN, ADP, PQ) was characterized with a new protocol based on the use of Caco-2 cells seeded directly on 96-well plates and the use of fluorescent substrates for efflux pumps, Rhodamine-123 (R123) for the case of P-gp. The experiment was performed as described in the reference (Sevin *et al.*, 2019).

Shortly, Caco-2 cells were washed with HEPES-buffered Ringer's (RH) solution (NaCl 150 mM, KCl 5.2 mM, $CaCl_2$ 2.2 mM, $MgCl_2$ 0.2 mM, $NaHCO_3$ 6 mM, Glucose 2.8 mM, HEPES 5 mM, water for injection), pH = 7.4 and incubated for 120 min with 10 µM R123. After incubation, cells were washed with RH and incubated with test compounds for one hour during which the rate out (K_{out}) of R123 (λ_{ex} = 485 nm and λ_{em} = 538 nm) was monitored every 2 minutes with a microplate fluorescence reader (Fluoroskan Ascent

FL, Thermo Labsystems, Issy-Les-Moulineaux, France) at 37 °C, after which the cell viability was assessed using MTT cytotoxicity assay kit. Rhodamine K_{out} was calculated as the slope of the curve of the cumulative R123 fluorescence against the time. Verapamil was used as a positive control and diazepam as a negative control.

Amiodarone, busulfan, cerium dioxide nanoparticles, cyclosporine A, diazepam, gentamicin sulfate, lead (II) chloride, paraquat dichloride, rhodamine 123, valproic acid and verapamil were obtained from Sigma-Aldrich (Saint Quentin Fallavier, France); doxorubicin hydrochloride was obtained from J&K Scientific (Lommel, Belgium) and pamidronate was obtained from Tebu-bio (Heerhugowaard, The Netherlands).

Cerium dioxide, doxorubicin, gentamicin C, lead (II) chloride, pamidronate and paraquat dichloride were dissolved in water; cyclosporine A, valproic acid and verapamil were dissolved in DMSO; amiodarone was dissolved in methanol, and busulfan in acetone.

MTT ((3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide) tetrazolium reduction assay), Cell Proliferation, and Cytotoxicity Assay Kit were obtained from Alphabioregen (Boston, MA, USA).

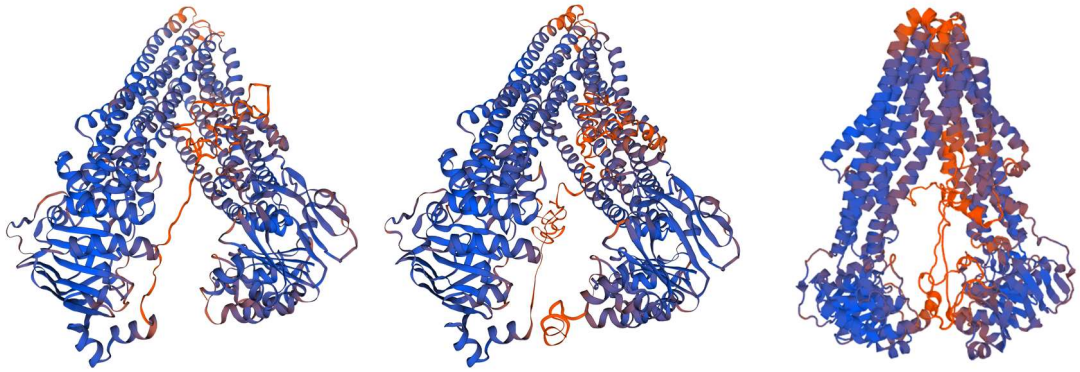
4.2.4 Results and Discussion

4.2.4.1 Homology Modelling of *h*P-gp

The crystal structure of *m*P-gp (PDB ID: 4M1M), which has an 87% sequence identity to the *h*P-gp, was selected as the most suitable template for developing the homology models of the 3D structure of *h*P-gp. However, the models built using the tools Discovery Studio 4.1/Modeler 9.12 (Accelrys, 2017; Šali *et al.*, 1993) and I-TASSER (Roy *et al.*, 2010; J. Yang *et al.*, 2015; Y. Zhang, 2008) were based on the alignment utilizing more than one template, among these the crystal structure of *Caenorhabditis elegans* P-gp (PDB ID: 4F4C), which was included in the model generated with the tool Discovery Studio 4.1 /Modeler 9.12 (Accelrys, 2017; Šali *et al.*, 1993).

The alignments between the *h*P-gp sequence and the templates can be found in Appendix B, Figure B.3. Several models were created using each of the tools and one model per tool was selected for further evaluation and validation.

4.2.4.1.1 SWISS-MODEL A total of five initial *h*P-gp models were created with the tool SWISS-MODEL (Schwede *et al.*, 2003) using as template the crystal structure of the *m*P-gp (PDB ID: 4M1M). The obtained models were evaluated using the scoring functions GMQE (global model quality estimation) and QMEAN (Benkert *et al.*, 2011). Model 1 (Figure 4.1.a) was selected for further evaluation and validation because it had the best quality factors in the set of models developed (Table 4.1). Although, in general terms, the selected 3D structure was evaluated as a good quality model, some regions are still poorly modelled, and the residues involved can be easily identified by looking at the Local Quality Plot in Figure 4.2. According to this plot, most of the residues in the structure agree well with the estimated native structure, except for the residues belonging to the linker region (a disorganized coil region of about 75 residues in length). These residues were modelled with less confidence because their individual QMEAN scores are below the threshold of 0.6. One of the reasons why this region was modelled with less reliability might be that the linker region between the two homologous halves of the protein has not yet been resolved in any of the available P-gp crystal structures; therefore, it is not present in any of the templates used.



(a) (b) (c)

Figure 4.1: Three-dimensional structures of the selected *h*P-gp models. The models are shown in colours based on QMEAN values to allow instant visualisation of the well (blue) or poorly modelled(orange) regions: (a) Model generated with the tool SWISS-MODEL; (b) model generated with the tool I-TASSER; (c) model generated with the tool Discovery Studio 4.1/Modeler 9.12.

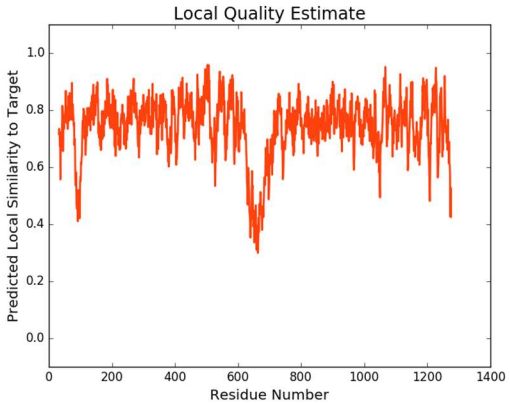


Figure 4.2: Local quality plot of Model 1, generated with the tool SWISS-MODEL.

Table 4.1: QMEAN and GMQE scores of the models generated with the tool SWISS-MODEL.

	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**
QMEAN	-2.71	-2.80	-2.72	-3.13	-3.28
GMQE	0.87	0.86	0.81	0.80	0.79

4.2.4.1.2 I-TASSER Five models were generated with the tool I-TASSER using the crystal structures of *m*P-gp (PDB IDs: 4M1M, 5KO2, 5KOY, 3G61, 3G5U) as template. The resulting models were evaluated using the scoring functions C-score, TM-score (Y. Zhang *et al.*, 2004a), and root mean square deviation (RMSD). The C-scores for the five generated models are given in Table 4.2, while the TM-score and RMSD are only given

for Model 1. The correlation between C-score and TM-score is weak for models with lower rank; therefore, they are not calculated. However, the C-score, Number of decoys and Cluster density for all models are reported for reference. Model 1 shown in Figure 4.1.b was selected for further evaluation and validation because it has the best quality factors in the set of generated models; a positive C-score value of 0.49, a high TM-score value of 0.78, and the largest cluster size. From the Estimated Local Accuracy Plot, shown in Figure 4.3, it can be seen that the residues belonging to the linker region have a larger distance deviation (in Angstroms) between the residue positions in the model and the predicted native structure. Figure 4.1.b also shows the poorly predicted regions of model 1; these regions match those evaluated in the previous model (Figure 4.1.a). As mentioned earlier, one reason why the linker region is modelled less reliably could be the lack of crystallographic information about it in the available P-gp crystal structures.

Table 4.2: C-score, TM-Score, and root mean square deviation (RMSD) of the models generated with the tool I-TASSER.

	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**
C-Score	0.49	-1.35	-1.24	0.29	-2.30
TM-score	0.78±0.10				
RMSD	8.3±4.5				
Number of decoys	2702	527	512	1993	0.82
Cluster density	0.1620	0.0256	0.0286	0.1325	0.0099

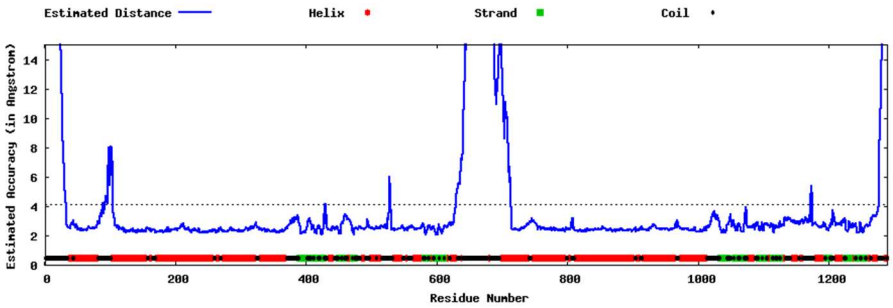


Figure 4.3: Local structure error profile of Model 1 generated with the tool I-TASSER.

4.2.4.1.3 Discovery Studio 4.1/Modeler A total of 20 initial models were generated with Discovery Studio 4.1/Modeler 9.12 using as template the crystal structures of *m*P-gp (PDB IDs: 6FN4, 4M1M, 5KPI, 4M2S, 5KO2, 3G60) and the crystal structure of *C. Elegans* P-gp (PDB ID: 4F4C). The resulting models were evaluated using the scoring functions Discrete Optimized Protein Energy (DOPE) and Probability Density Function (PDF) Total Energy (Table 4.3). Since all models have similar PDF Total Energy, the DOPE score was used for selecting the top-ranked model. Model 16 (Figure 4.1.c) was selected for further evaluation via the Verify Protein protocol in Discovery Studio 4.1.

According to the scores presented in Table 4.4, model 16 is of good quality as the Verify score obtained is higher than the Verify Expected Low Score value and very close to the Verify Expected High Score value. The regions of the structure with large violations of the homology restraints are shown in Figure 4.4.a using the PDF Total Energy plot. The Verify score per amino acid, indicating whether or not a residue is in the desired 3D environment, is shown in Figure 4.4.b. The regions of the protein where

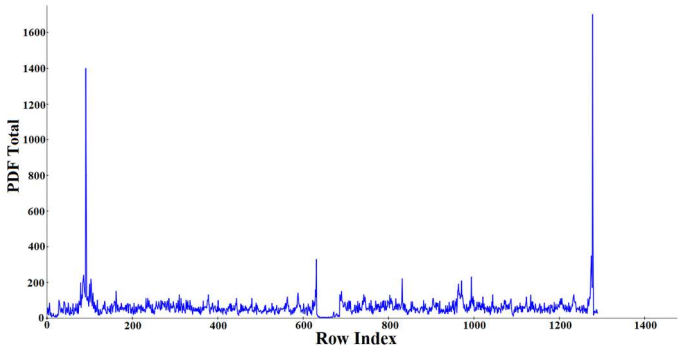
the score approaches zero or becomes negative are likely misfolded and should be carefully examined. Moreover, in this case, the less reliable regions of the model correspond to those found in the previous models (Figure 4.1. (a) and (b)).

Table 4.3: Probability Density Function (PDF) Total Energy and DOPE Score of the models generated with the tool Discovery Studio 4.1/Modeler 9.12.

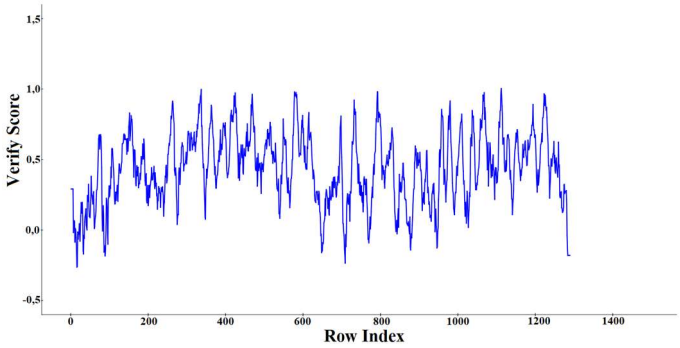
	PDF Total Energy	**DOPE Score**
Model 1	80.529,9	-151.500
Model 2	81.619,5	-151.832
Model 3	79.999,4	-153.656
Model 4	79.720,3	-153.494
Model 5	80.422,7	-153.242
Model 6	80.065,3	-153.370
Model 7	81.341,2	-151.484
Model 8	79.800,9	-153.459
Model 9	80.116,5	-153.266
Model 10	80.489,5	-153.900
Model 11	81.588,3	-152.353
Model 12	80.908,7	-152.955
Model13	80.406,8	-153.855
Model 14	80.046,3	-153.624
Model 15	80.453,5	-153.758
Model 16	**79.976,6**	**-154.280**
Model 17	80.841,3	-153.293
Model 18	82.785,2	-151.571
Model 19	79.876,5	-153.904
Model 20	80.607,2	-152.350

Table 4.4: Verify Scores of Model 16 generated with the tool Discovery Studio 4.1/Modeler 9.12.

	Verify Score	**Verify Expected High Score**	**Verify Expected Low Score**
Model 16	513,011	593,427	267,042



(a)



(b)

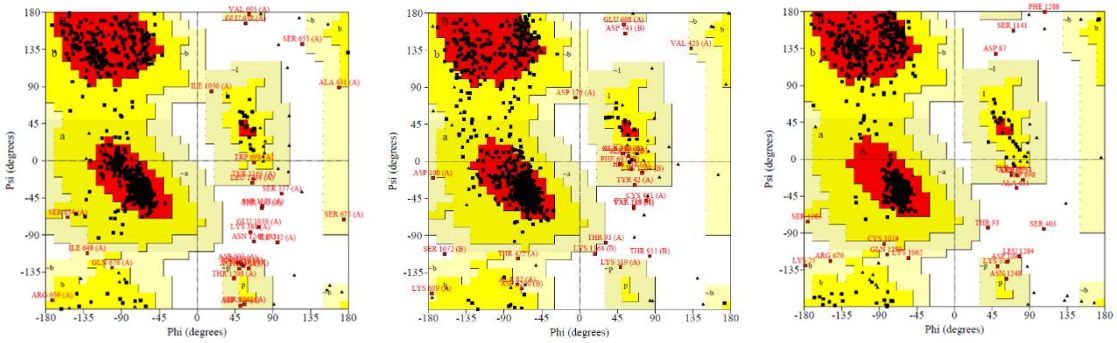
Figure 4.4: (a) PDF Total Energy Plot; (b) Verify Score Plot of Model 16 generated with the tool Discovery Studio 4.1/Modeler 9.12.

4.2.4.2 Models Validation

The quality of the models was assessed to verify that they were reliable and suitable for performing further molecular docking simulations. Validation of the selected models from each modelling tool was performed using available online structural quality assessment tools, such as PROCHECK (Laskowski *et al.*, 1993), Verify 3D (Bowie *et al.*, 1991), ERRAT (Colovos *et al.*, 1993), and PROVE (Pontius *et al.*, 1996).
The stereochemical properties of the *h*P-gp models were evaluated using the software PROCHECK via the Ramachandran Plot and the results were compared with those obtained from the crystal structure of *m*P-gp (PDB ID: 4M1M). The resulting Ramachandran Plots of the predicted *h*P-gp models are shown in Figure 4.5 and the corresponding statistics are listed in Table 4.5.

The plots revealed that the phi (φ) and psi (ψ) backbone dihedral angles in the *h*P-gp models are reasonably accurate, as the majority of the residues are within the allowed regions; less than 1.0% of the residues are in the disallowed regions in all the models evaluated. The residues in the disallowed regions are mainly located in the NBDs of the protein, except for one residue (Y710) in the I-TASSER model, which is located in the TM domain helix; however, none of the residues in the binding pocket are located in the disallowed regions. Considering the φ/ψ distribution of the amino acids in the modelled *h*P-gp structures, the results are consistent with those of the experimentally available

*m*P-gp structure (PDB ID: 4M1M) reported in Table 4.5. In summary, the stereochemical quality of the models is satisfactory, with similarities to that of the template used.



(a) (b) (c)

Figure 4.5: Ramachandran Plots for the modelled 3D structures of the *h*P-gp. The red, yellow, and white areas represent the favoured, allowed, and disallowed regions, respectively: (a) SWISS-MODEL model; (b) I-TASSER model; (c) Discovery Studio 4.1/Modeler 9.12 model.

Table 4.5: Ramachandran Plot Statistics of the *h*P-gp models and the crystal structure of *m*P-gp (PDB: 4M1M).

	4M1M	**SM** [1]	**IT** [2]	**DS** [3]
Residues in most favoured regions	91.5%	92.0%	87.7%	92.0%
Residues in additional allowed regions	7.0%	5.5%	10.0%	6.3%
Residues in generously allowed regions	1.3%	1.7%	1.4%	1.1%
Residues in disallowed regions	0.1%	0.8%	0.9%	0.6%

[1] SWISS-MODEL model. [2] I-TASSER model. [3] Discovery Studio model

To assess the overall folding of the models, a structural comparison of the developed *h*P-gp models was performed in Discovery Studio 4.1 by superimposing the *h*P-gp models over the crystal structure of the *m*P-gp (PDB ID: 4M1M) (Figure 4.6). The root mean square deviation (RMSD) of the main chain and the number of overlapping residues are shown in Table 4.6. The RMSD values in relation to the side chain, alpha carbons, and total protein were also calculated and are shown in Table 4.7.

Table 4.6: Alignment of the selected models with respect to the crystal structure of *m*P-gp (PDB: 4M1M). Main-chain RMSD (in Angstroms) are below the diagonal and Number of Overlapping Residues above the diagonal.

	4M1M	**SM** [1]	**IT** [2]	**DS** [3]
4M1M		1181	1167	517
Model SM [1]	0.2430		1167	517
Model IT [2]	0.7000	0.7010		517
Model DS [3]	1.8330	1.8350	1.8680	

[1] SWISS-MODEL model. [2] I-TASSER model. [3] Discovery Studio model

Table 4.7: RMSD (in Angstroms) of the selected models with respect to the crystal structure of *m*P-gp (PDB: 4M1M).

	Reference	**C-Alpha**	**Side-chain**	**All Protein**
SM [1]	4M1M	0.173	0.473	0.369
IT [2]	4M1M	0.528	2.074	1.510
DS [3]	4M1M	1.809	2.365	2.098

[1] SWISS-MODEL model. [2] I-TASSER model. [3] Discovery Studio model

The results showed a small RMSD of the main chain against the crystal structure of *m*P-gp (PDB ID: 4M1M) with values of 0.24 Å and 0.70 Å for SWISS-MODEL and I-TASSER models, respectively. The model generated with Discovery Studio 4.1/Modeler 9.12 yielded the largest RMSD value of 1.83 Å. When overlapping a model with the template, the generally accepted RMSD threshold is 2.0 Å; thus, the three *h*P-gp models evaluated are within the accepted limit. Superimposition between the *m*P-gp (PDB ID: 4M1M) and *h*P-gp models revealed that the developed models have significant 3D similarities and that the overall folding is correct.

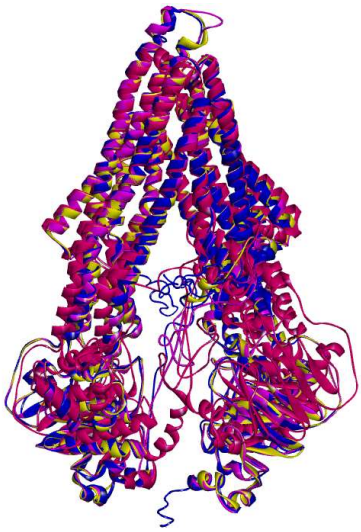


Figure 4.6: Superimposed protein structures of the *h*P-gp models generated and the crystal structure of *m*P-gp (PDB ID: 4M1M). The *m*P-gp is coloured yellow.

Additional analysis of the predicted *h*P-gp 3D structures was performed using the Verify 3D, ERRAT, and PROVE structural assessment tools (Table 4.8). The Verify 3D tool, which provides an analysis of the compatibility of the 3D models with their amino acid sequence (1D), yielded scores lower than 80% for the three models evaluated. This means that less than the 80% of the amino acids in the structures have a score ≥0.2 in the 3D/1D profile. The best score was obtained for the model I-TASSER with 63.41% of the residues having an averaged 3D/1D score ≥0.2; very close to the result obtained with the crystal structure of the *m*P-gp (PDB ID: 4M1M). SWISS-MODEL and Discovery Studio models both resulted with 45% of the residues with an averaged 3D/1D score ≥0.2. According to these results, none of the predicted 3D models of *h*P-gp passed the assessment; however, the quality indicator also performed poorly for the crystallized structure of *m*P-gp with only 65.20% of the residues within the scoring limit.

Table 4.8: Verify 3D, ERRAT and PROVE Scores of the selected models.

	Verify 3D	**ERRAT**	**PROVE**
SM [1]	44.69%	94.0120	6.8%
IT [2]	63.41%	96.0884	5.6%
DS [3]	45.22%	80.5934	7.2%
PDB ID: 4M1M	65.20%	86.5620	0.0%

[1] SWISS-MODEL model. [2] I-TASSER model. [3] Discovery Studio model

The overall quality factor for non-bonded atomic interactions between different atom types in the modelled structures was evaluated using the ERRAT program. The ERRAT score is expressed as the percentage of protein for which the calculated error value falls below the 95% rejection limit. The ERRAT score should be greater than 50% for considering a model to be of good quality. The overall quality factors for the SWISS - MODEL and I- TASSER models were approximately 95%, with the I- TASSER model achieving the highest value of 96.08%. The Discovery Studio model and the crystal structure of the *m*P-gp (PDB ID: 4M1M) were both with scores around 80% below the rejection limit. Based on the previous results, the currently evaluated 3D *h*P-gp models have good reliability.

The volume-based structure validation of the *h*P-gp models was done utilizing the program PROVE. The crystal structure of the *m*P-gp (PDB ID: 4M1M), which was used as a reference for validating developed models, yielded less than 1% buried outlier atoms, which is the threshold value for considering the test passed. Outliers are considered here as being those buried atoms for which the volume is more than 3.0 standard deviations away from the expected volume. The results for the three models evaluated showed that there are some errors in the structures, as the percentage of buried outlier atoms was greater than 5% in all cases.

The structure quality assessment using the online tools PROCHECK, Verify 3D, ERRAT, and PROVE revealed that the three developed *h*P-gp models are as good as the crystal structure of the *m*P-gp used as template (PDB ID: 4M1M). They are reliable and of suitable quality for further molecular docking simulations. Nevertheless, the I-TASSER *h*P-gp model had a better performance with respect to the other two models in a large part of the tests.

To check for a possible bias in the quality of the selected *h*P-gp model (I-TASSER) containing NBDs (corresponding to the complete primary sequence of the *h*P-gp protein used), an additional quality assessment of its truncated counterpart (without NBDs) was performed (see Appendix B, Table B.3, Figure B.4). As shown in Table B.3, there are no significant differences between the two models in terms of calculated quality scores, with a slight exception for the Verify 3D score (38.19% for the truncated *h*P-gp model compared to 63.41% for the selected full-length *h*P-gp model). The latter was somehow to be expected, since the Verify 3D scoring method is biased on the 1D sequence (primary sequence) of the model as well as its secondary-structure composition (Bowie *et al.*, 1991), which structural information is actually missing in the truncated *h*P-gp model. Nevertheless, the ERRAT and PROVE scores, as well as the PROCHECK assessment (Figure B.4), suggest that the truncated *h*P-gp model (without NBDs) is highly comparable to the selected model in terms of quality and reliability, allowing the use of the selected full-length *h*P-gp model I-TASSER for further structure-based (molecular docking) calculations.

4.2.4.3 Molecular Docking Calculations

4.2.4.3.1 Docking into Homology Model Ligand docking is a commonly used approach to identify ligand–protein interactions. However, in the case of P-gp, this could be challenging due to the high degree of flexibility and the large binding cavity consisting of multiple binding sites (Aller *et al.*, 2009; Tandon *et al.*, 2006). In addition, P-gp can bind more than one ligand simultaneously (Loo *et al.*, 2003a; Lugo *et al.*, 2005) and until 2019, when the cryoEM structure of human ABCB1 was resolved (PDB ID: 6QEX) (Alam *et al.*, 2019), a high-resolution crystal structure of *h*P-gp was lacking, necessitating the use of protein homology models, which added additional layers of uncertainty to the process. Nevertheless, a recent study reports the use of homology models in virtual screening applications with superior performance compared to crystal structures (Mordalski *et al.*, 2015). This fact is explained by the conformational flexibility offered by homology models, which allows a better accommodation of diverse ligands and thus better screening performance.

Two docking runs were performed utilizing two different algorithms, CDOCKER (G. Wu *et al.*, 2003) and GOLD (G. Jones *et al.*, 1995; G. Jones *et al.*, 1997). In order to analyse the binding pocket of the *h*P-gp, the simulations began with the docking of thirteen compounds, including eight well-known molecules that interact with P-gp as substrates, inhibitors or both: cyclosporine A (CsA), amiodarone (AM), doxorubicin (DOX), digoxine (DIG), loperamide (LPM), rifampin (RMP), verapamil (VER), carvedilol (CAR); and five non-interacting compounds with P-gp: valproic acid (VPA), busulfan (BU), gentamicin (GEN), pamidronate (APD), and paraquat (PQ).

The large binding pocket observed in the *m*P-gp crystal structure (PDB ID: 4M1M) binds the ligands at different sites with partially overlapping residues; some of them are identical to those involved in rhodamine and verapamil binding (Loo *et al.*, 2006; Loo *et al.*, 1997, 2001, 2002). Therefore, when defining the binding site for performing the docking simulations, the entire transmembrane (TM) region was considered. The binding region was delineated by the atoms within a 15 Å and 24.7 Å radius around the co-crystallized *m*P-gp ligand (PDB ID: 4XWK), for CDOCKER and GOLD calculations, respectively. Binding site selection and docking simulation settings were validated by a re-docking procedure (ligand reproduction). RMDS values for heavy atoms of 1.5697 Å, 1.6021 Å, 1.6427 Å for CDOCKER calculations and 0.5527 Å, 0.7988 Å, 0.8498 Å for GOLD calculations were obtained (see Appendix B, Table B.1, Figure B.1). The RMSD results are in agreement with the accepted threshold of 2 Å.

The resulting docking poses were subsequently rescored with fourteen additional scoring functions implemented in Discovery Studio 4.1. The main fitness function used during the docking runs and the rescoring functions calculated for the resulting poses are listed in Table 4.9.

Based on the results of the sum of ranking differences (SRD), the best ranking poses were selected using the consensus ranking scheme, fusing the six best performing scoring fitness functions, and using the geometric mean to calculate the fused ranking (Table 4.10).

Table 4.9: Docking runs performed and Scoring functions.

Docking run	Main scoring function	Rescoring functions
CDOCKER	-CDocker Energy	LigScore2_Dreiding, LigScore1_Dreiding, PLP1, PLP2, Jain, Ludi_1, PMF, PMF04, Goldscore, Chemscore, ChemASP ChemPLP, -Cdocker_IE[2], -Cdocker_Eopt[3], -Cdocker_IEOpt[4]
GOLD	GoldScore	LigScore2_Dreiding, LigScore1_Dreiding, PLP, PLP2, Jain, Ludi_1, PMF, PMF04, Chemscore, ChemASP, ChemPLP, -Cdocker_E[1], -Cdocker_IE[2], -Cdocker_Eopt[3], -Cdocker_IEOpt[4]

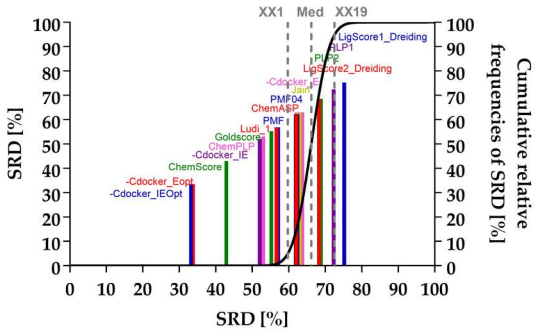
[1] -CDocker energy. [2] -CDocker interaction energy. [3] -CDocker energy optimized. [4] -CDocker interaction energy optimized

Table 4.10: Fusing ranking scheme.

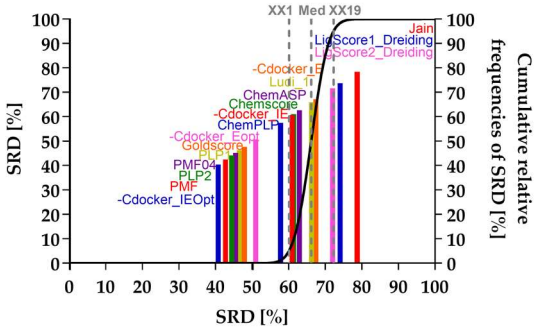
Docking run	Fused scoring functions ranks	Fusion Operator
CDOCKER	-Cdocker_IE[1], -Cdocker_Eopt[2], Chemscore, Goldscore ChemPLP, -Cdocker_IEOpt[3]	Geometric Mean
GOLD	-Cdocker_IEOpt[3], PMF, PMF04, PLP1, PLP2, Goldscore,	Geometric Mean

[1] -CDocker interaction energy. [2] -CDocker energy optimized. [3] -CDocker interaction energy optimized

According to the SRD results, the best scoring functions for the CDOCKER run had a very low probability that their performance was of random character, with values greater than 2.91E-15% and less than or equal to 0.92% (see Table 4.11); they performed better than random ranking, as they do not overlap with the cumulative relative frequency curve of a random ranking shown in Figure 4.7.a. For the results of GOLD, the best performing scoring functions had a probability greater than 2.64%-10% and less than or equal to 1.01%, as you can see in Table 4.12. In this case, they also do not overlap with the cumulative relative frequency curve of random ranking shown in Figure 4.7.b.



(a)



(b)

Figure 4.7: The sum of ranking differences (SRD) analysis of the 16 fitness functions calculated for each docking run: (a) CDOCKER; (b) GOLD. Normalized SRD values are plotted on the x and left y axes. The cumulative relative frequencies of SRD values for random ranking are plotted on the right y axis and shown as a black curve.

Table 4.11: SRD ranking of the 16 fitness functions used in the CDOCKER run.

Name	SRD	p% $x <$ SRD $\geq x$	
-Cdocker_Eopt	2360	2.91E-15	3.00E-15
-Cdocker_IEOpt	2378	4.57E-15	4.66E-15
ChemScore	3046	1.40E-07	1.47E-07
-Cdocker_IE	3696	1.80E-02	1.83E-02
ChemPLP	3764	4.41E-02	4.48E-02
Goldscore	3920	0.29	0.30
Ludi_1	4022	0.83	0.84
PMF	4034	0.91	0.92
XX1[1]	4228	4.99	5.03
ChemASP	4404	15.48	15.57
PMF04	4412	16.22	16.31
Jain	4455	20.21	20.31
-Cdocker_E	4466	21.33	21.45
Q1[2]	4497	24.99	25.10
Med[3]	4685	49.90	50.05
LigScore2_Dreiding	4850	72.29	72.41
PLP2	4862	73.73	73.85
Q3[4]	4871	74.94	75.06
PLP1	5134	94.66	94.70
XX19[5]	5142	94.97	95.01
LigScore1_Dreiding	5338	99.06	99.07

[1] First icosaile 5%. [2] First quartile. [3] Median. [4] Last quartile. [5] Last icosaile 95%

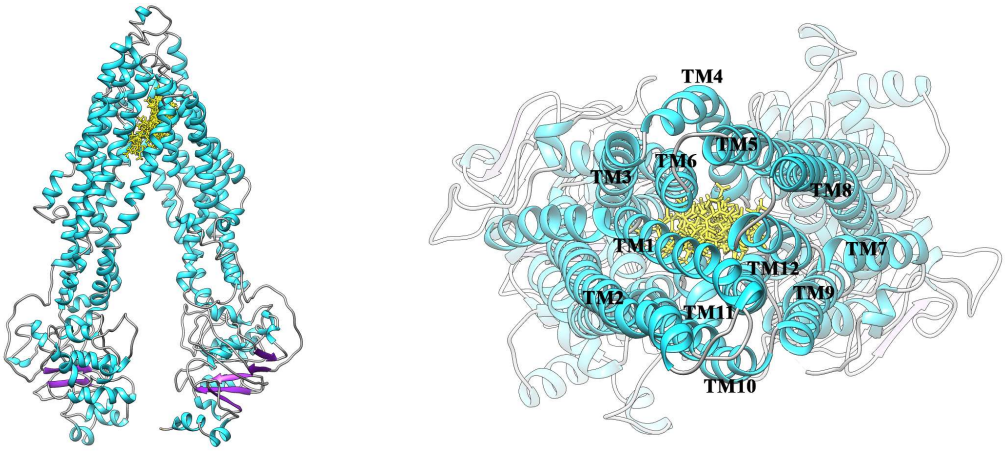
Table 4.12: SRD ranking of the 16 fitness functions used in the GOLD run.

Name	SRD	p% x < SRD ≥ x	
-Cdocker_IEOpt	3526	2.60E-10	2.64E-10
PMF	3698	9.53E-09	9.70E-09
PLP2	3850	1.71E-07	1.76E-07
PMF04	3942	9.50E-07	9.72E-07
PLP1	4028	4.23E-06	4.33E-06
Goldscore	4152	3.21E-05	3.25E-05
-Cdocker_Eopt	4422	1.66E-03	1.67E-03
ChemPLP	5010	1.00	1.01
XX1[1]	5234	4.97	5.01
-Cdocker_IE	5290	7.12	7.16
Chemscore	5320	8.36	8.41
ChemASP	5462	17.48	17.55
Q1[2]	5546	24.95	25.04
Ludi_1	5748	47.85	47.97
Med[3]	5765	49.93	50.06
-Cdocker_E	5853	60.66	60.77
Q3[4]	5983	74.98	75.08
LigScore2_Dreiding	6249	93.20	93.25
XX19[5]	6298	94.98	95.01
LigScore1_Dreiding	6428	97.96	97.97
Jain	6835	99.95	99.95

[1] First icosaile 5%. [2] First quartile. [3] Median. [4] Last quartile. [5] Last icosaile 95%

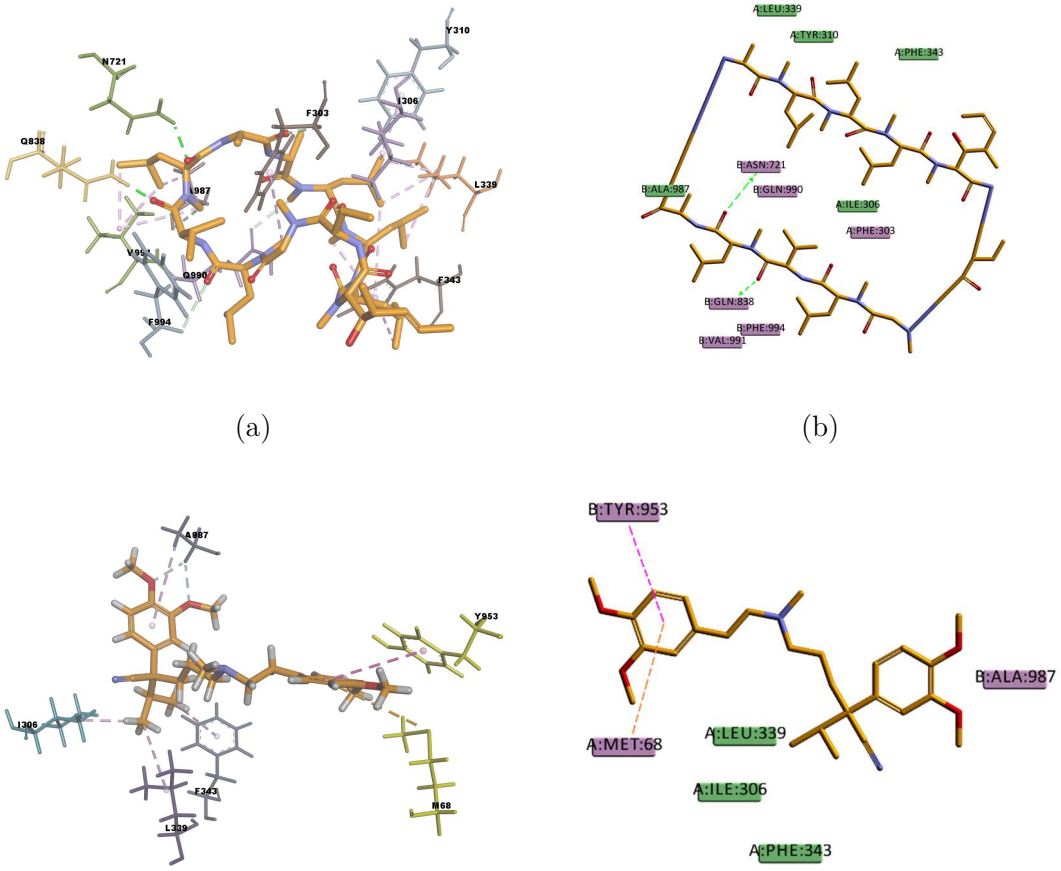
The resulting poses were distributed within the TM regions of P-gp (Figure 4.8) and showed interactions with protein residues of multiple TM helices located throughout the binding region. The interacting amino acids were identified using the Ligand Interactions tools in Discovery Studio 4.1. For the CDOCKER results, residues on TM helices 5, 6, 7, 8, 10 and 12 were mainly involved in binding, while for the GOLD results, additional residues in TM 1, 9 and 11 were involved in the binding of VER and CsA. The pose obtained for CsA shows a conventional hydrogen bond with Q838 (TM9) and the pose obtained for VER shows a π–sulphur interaction with M68 (TM1) and a π–π interaction with Y953 (TM11) (Figure 4.9).

The docking results obtained with the CDOCKER algorithm were comparable to the results obtained with the GOLD algorithm in terms of the calculated binding energies and the nature of the interactions between the docked compounds and *h*P-gp. The estimated binding energies for the docking set calculated using the Calculate Binding Energies protocol in Discovery Studio 4.1 and presented in Table 4.13, are in close agreement for the two methodologies employed. The nature of the interactions was essentially the same in both cases (Table 4.14; Appendix B, Table B.4); mainly hydrophobic π–sigma, π–alkyl, and π–π interactions with the presence of some hydrogen bonding interactions.



(a) (b)

Figure 4.8: Distribution of the selected ligand poses (yellow) in the homology model of *h*P-gp; (a) Frontal view (b) View from the extracellular side of the protein looking into the internal chamber. The colour representation is according to the secondary structure: helices are shown in cyan, beta sheets in purple, and coils in gray.



(a) (b)

(c) (d)

Figure 4.9: Cyclosporine A (CsA) and verapamil (VER) top-ranked poses obtained with GOLD algorithm. (a) 3D view of CsA interactions in the binding pocket. Residue Q838 (TM 9) is highlighted in yellow; (b) 2D interaction diagram of CsA with *h*P-gp interacting residues; (c) 3D view of VER interactions in the binding pocket. Residues M68 (TM1) and Y953 (TM11) are highlighted in yellow; (d) 2D interaction diagram of VER with *h*P-gp interacting residues. The green dotted lines represent conventional hydrogen bonds, light-green dotted lines represent carbon hydrogen bonds, light-rose dotted lines represent hydrophobic interactions, the pink dotted line represents π–π stacking interaction, and the orange dotted line represents π–sulphur interaction.

Table 4.13: Free energies of binding. Estimate of the overall binding free energies of the compounds under study, using the homology model.

Name	**CDOCKER Binding Energy (kcal/mol)**	**GOLD Binding Energy (kcal/mol)**
CsA[1]	-268.516	-299.839
AM[2]	-97.4066	-103.782
DOX[3]	-196.862	-199.953
DIG[4]	-211.755	-226.553
LPM[5]	-106.862	-120.356
RMP[6]	-214.872	-208.965
VER[7]	-101.617	-116.922
CAR[8]	-115.509	-120.018
VPA[9]	-45.9908	-54.5415
BU[10]	-59.7798	-69.8159
GEN[11]	-180.391	-189.158
APD[12]	-93.428	-99.6753
PQ[13]	-193.24	-206.54

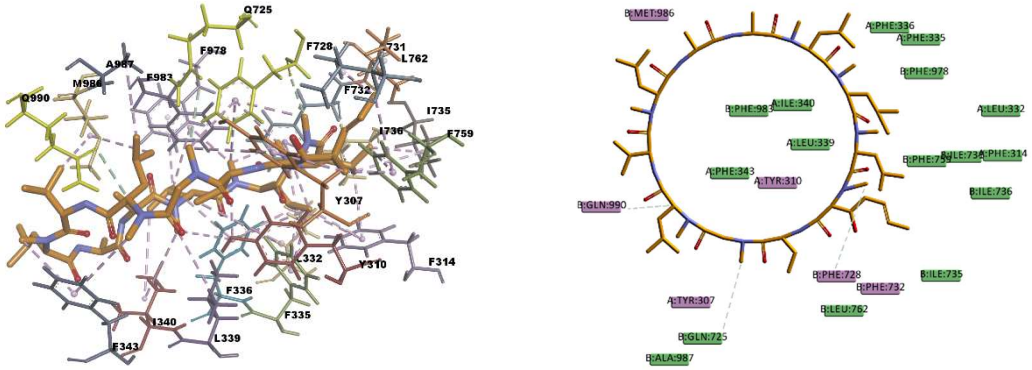
[1] Cyclosporine A; [2] amiodarone; [3] doxorubicin; [4] digoxin; [5] loperamide; [6] rifampin; [7] verapamil; [8] carvedilol; [9] valproic acid; [10] busulfan; [11] gentamicin; [12] pamidronate; [13] paraquat.

Table 4.14: Ligand–P-gp interactions. Nature of the interactions with P-gp and amino acid residues involved in binding of the compounds under study, using the homology model and CDOCKER protocol. Numbers in parenthesis indicate the number of interactions involving the residue.

Name	H–Bond	Alkyl	π–Sigma	π–Alkyl	π–π	π–Sulphur	Others
CsA[1]	Q990, Q725, F728	A987, M986 (2), L339, I340 (2), L332, I731, L762, I735 (2), I736	F335	Y307 (2), Y310, F728 (4), F732 (5), F314 (3), F335, F336 (5), F343 (2), F759 (3), F978, F983 (2)	-	-	-
AM[2]	I731	I731, I735 (2), I736	Y310	Y307, F728 (2), F314 (2) F759 (3)	F983 (2), F732, F728	-	-
DOX[3]	Y310, Y307, F732, F759, Q990	I731, l762	-	F759	F728 (3)	M986	-
DIG[4]	F728 (2), Q725, Q990	A987	Y310	Y307, Y310, F728 (3), F732, F336 (2), F343, F983 (2)	-	-	-
LPM[5]	F728	L762	F732	Y307, Y310, F728 (2), F759, L339, I340	F314	-	-
RMP[6]	F732, F728, F759	I340, M986	Y310, F732	Y307, Y310 (2), F335, F336, F343, F759 (2), F983 (2) F728, F732	F983	-	-
VER[7]	I731	-	F314	-	Y310, F728	-	F732*
CAR[8]	-	I306	-	-	F314, F759	M986	-
VPA[9]	-	-	-	Y310, F336, F728, F759	-	-	-
BU[10]	F732	-	-	-	-	F336, F728, F314, F759	-
GEN[11]	Y310, I731	I736	-	F314, F732	-	-	-
APD[12]	Y310	-	-	-	-	-	-
PQ[13]	-	-	-	-	Y310, F314, F728, F732	-	F314**, Y310**

[1] Cyclosporine A; [2] amiodarone; [3] doxorubicin; [4] digoxin; [5] loperamide; [6] rifampin; [7] verapamil; [8] carvedilol; [9] valproic acid; [10] busulfan; [11] gentamicin; [12] pamidronate; [13] paraquat; * amide···π stacking interaction; ** cation–π interaction.

Based on visual inspection of the selected docking poses, the interactions between the ligands and the residues in the binding pocket are mainly hydrophobic. In the case of CsA, at least 20 residues are involved in hydrophobic interactions (Y307, Y310, F314, L332, F335, F336, L339, I340, I343, F728, I731, F732, I735, I736, F759, L762, F978, F983, M986, and A987), which are listed in Table 13 and can be seen in Figure 10. Although CsA is a big molecule, it was found to have the lowest binding energy (-268.516 kcal/mol) in the set of docked molecules. The stability of CsA in the binding site could be attributed to the large number of π interactions present, such as π–alkyl interactions, and the simultaneous presence of hydrogen bonds in the binding pose. Even if the three hydrogen bonds present in the docked pose are carbon hydrogen bonds of the type C–H . . . O, i.e., weak interactions with a larger dispersive component, they may also play a role in stabilizing the ligand–protein complex. CsA is known to be a substrate with high affinity for P-gp (Litman *et al.*, 1997; Saeki *et al.*, 1993), therefore the docking results obtained are consistent with the available literature and with the experimental transport assay results shown in Figure 4.11, where CsA leads to a lower rate of rhodamine 123 (R123) excretion from Caco-2 cells. Nevertheless, it should be noted that Caco-2 cells also express other ABC transporters such as ABCG2 and ABCC1, which are involved in the active efflux of R123 from the cells and could be inhibited by CsA, as these transporters share many similarities with P-gp.



(a) (b)

Figure 4.10: Cyclosporine A (CsA) top-ranked pose obtained with the CDOCKER algorithm. (a) 3D view of CsA interactions in the binding pocket. Residues Q990, Q725, and F728 involved in the hydrogen bonds are highlighted in yellow; (b) 2D diagram of CsA with *h*P-gp interacting residues. Light-green dotted lines represent carbon hydrogen bonds and light-rose dotted lines represent hydrophobic interactions.

Similar to CsA, RMP is also a large molecule, with a very favourable estimated binding energy of -214.872 kcal/mol. The docking pose shown in Figure 4.12 shows hydrophobic interactions with 11 residues in the binding pocket (I340, M986, Y310, F732, Y307, F335, F336, F343, F759, F983, and F728. RMP exhibits a high number of π interactions in its binding mode, including π–sigma, π–alkyl, and π–π interactions. Hydrogen bonding interactions of a weak character are also present in the binding pose, such as weak carbon hydrogen bonds, and a π–donor hydrogen bond between the hydroxyl group (donor) in RMP and the π electron cloud over the aromatic ring in F759 (acceptor). The sum of these interactions undoubtedly creates a strong cohesive environment and thereby stabilises the complex formed. The docking results obtained are

in agreement with the experimental transport assay results reported in reference (Sevin *et al.*, 2019) and with the available literature on RMP and P-gp interactions. RMP is known to be a substrate (Collett *et al.*, 2004) and inducer (Geick *et al.*, 2001) of P-gp.

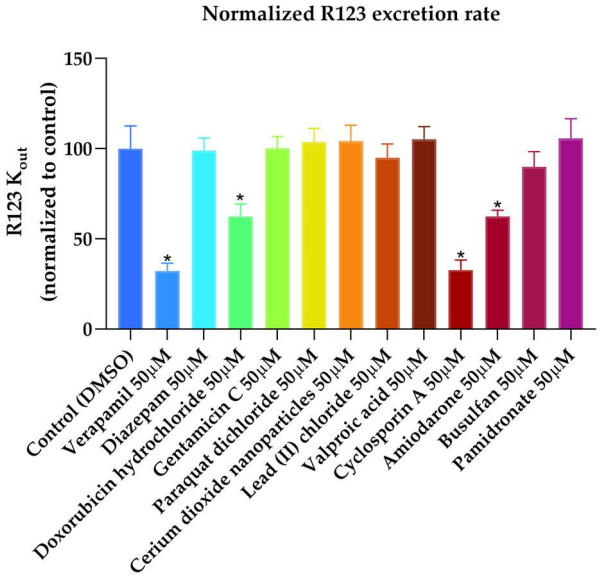
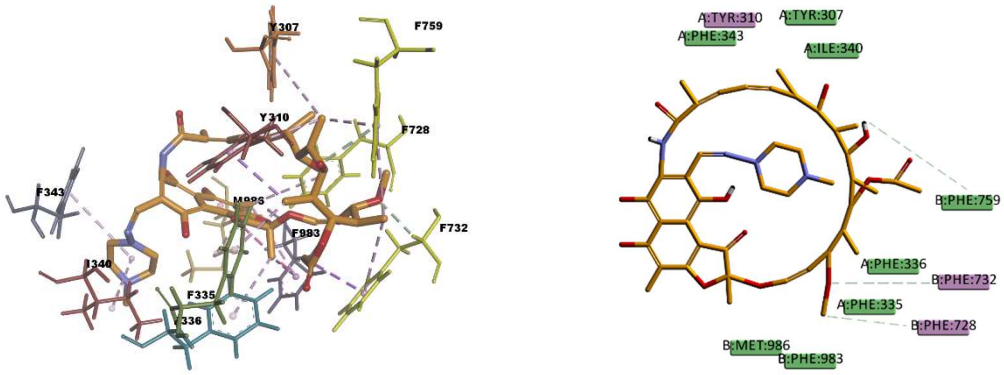


Figure 4.11: Effects of drugs on the excretion rate of rhodamine 123 (R123) from Caco-2cells. (n = 8, mean ± SD, * $p < 0{,}0001$). Results are expressed as percentages compared to the excretion rate of R123 in the absence of drug (i.e control DMSO). Error bars: SD; verapamil: positive control; diazepam: negative control.

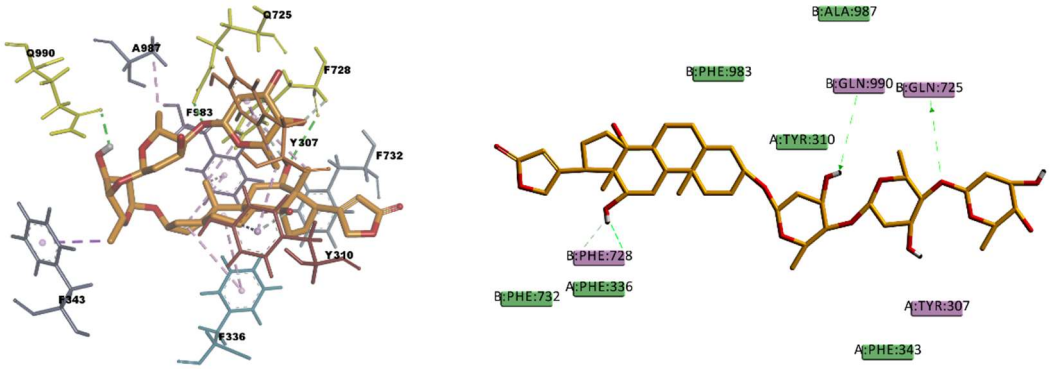


(a) (b)

Figure 4.12: Rifampin (RMP) top-ranked pose obtained with the CDOCKER algorithm. (a) 3D view of RMP interactions in the binding pocket. Residues F732, F759, and F728 involved in the hydrogen bonds are highlighted in yellow; (b) 2D interaction diagram of RMP with *h*P-gp interacting residues. Light-green dotted lines represent weak hydrogen bonds and light-rose dotted lines represent hydrophobic interactions.

Figure 4.13 shows the binding pose of DIG. The binding mode of DIG involves hydrophobic interactions with eight residues in the binding pocket (A987, F336, F343,

F728, F732, F983, Y307, and Y310) and four hydrogen bonds. Three of these are conventional, electrostatic type (N–H . . . O or O–H . . . O) with a strong character, and one is a π–donor hydrogen bond with a weaker character. The sum of these interactions contributes to the stability of the complex, as reflected in the value of the estimated binding energy of -211.755 kcal/mol. These results are in agreement with the available literature reporting DIG as a high affinity P-gp substrate (Taipalensuu *et al.*, 2004).

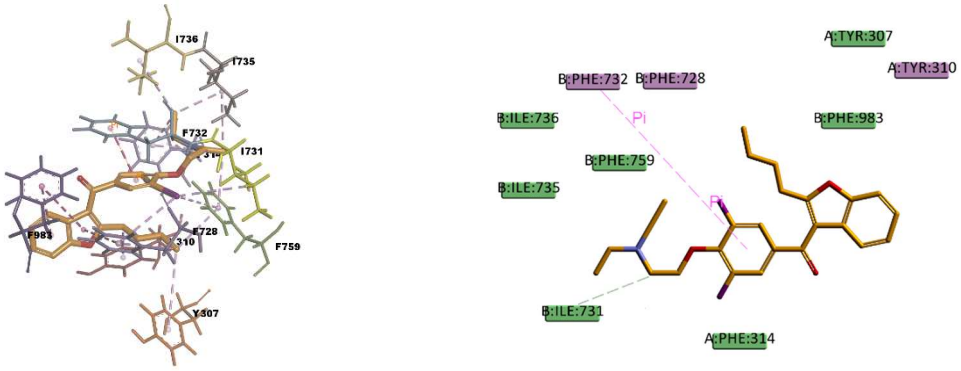


(a) (b)

Figure 4.13: Digoxine (DIG) top-ranked pose obtained with the CDOCKER algorithm. (a) 3D view of DIG interactions in the binding pocket. Residues Q725, Q990, and F728 involved in the hydrogen bonds are colored yellow; (b) 2D interaction diagram of DIG with *h*P-gp interacting residues. Green dotted lines represent conventional hydrogen bonds, the light-green dotted line represents π–donor hydrogen bonds, and light-rose dotted lines represent hydrophobic interactions.

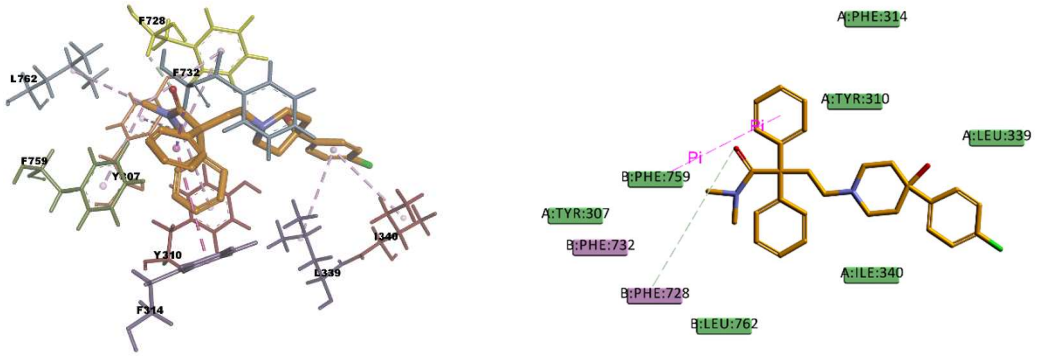
The docked pose of AM, shown in Figure 4.14, reveals many hydrophobic interactions involving ten residues in the binding pocket (Y307, Y310, F314, F728, I731, F732, I735, I736, F759, and F983), including many π–alkyl type of interactions, but also π–sigma and π–π interactions. The estimated binding energy is favourable but of lower order compared to CsA or RMP values. The difference in binding energies could be explained by the smaller number of hydrogen bonds in the binding mode; there is only one weak carbon hydrogen bond (C–H . . . O) with residue I731.

This is also the case of LPM, which shows hydrophobic interactions with nine residues in the binding pocket (L762, F732, Y307, Y310, F728, F759, L339, I340, and F314), but only one weak carbon hydrogen bond with residue F728 (Figure 4.15). The calculated binding energy of LPM is also favourable but smaller than the energy values of CsA or RMP, indicating lower binding affinity and stability compared to them. The obtained results are in agreement with the available literature and with the experimental transport assay results reported in Figure 4.11 and in reference (Litman *et al.*, 1997) for AM and in reference (Sevin *et al.*, 2019) for LPM. AM and LPM are known substrates of P-gp (Jouan *et al.*, 2016; Litman *et al.*, 1997; Wandel *et al.*, 2002).



(a) (b)

Figure 4.14: Amiodarone (AM) top-ranked pose obtained with the CDOCKER algorithm. (a) 3D view of AM interactions in the binding pocket. Residue I731 involved in the carbon hydrogen bond is highlighted in yellow; (b) 2D interaction diagram of AM with *h*P-gp interacting residues. Light-green dotted lines represent carbon hydrogen bonds, light-rose dotted lines represent hydrophobic interactions, and the pink dotted line represents π–π stacking interactions.

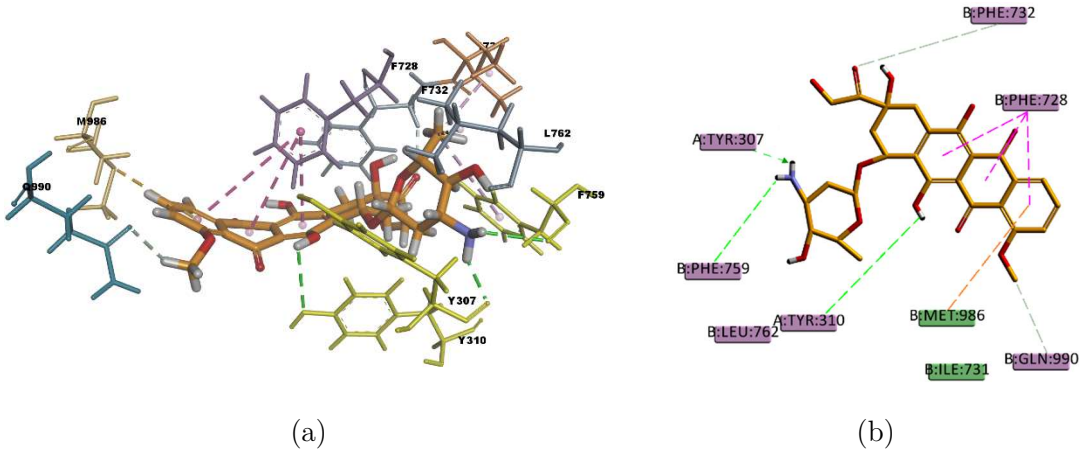


(a) (b)

Figure 4.15: Loperamide (LMP) top-ranked pose obtained with the CDOCKER algorithm. (a) 3D view of LMP interactions in the binding pocket. Residue F728 involved in the carbon hydrogen bond is highlighted in yellow; (b) 2D interaction diagram of LMP with *h*P-gp interacting residues. Light-green dotted lines represent carbon hydrogen bonds, light-rose dotted lines represent hydrophobic interactions, and the pink dotted line represents π–π interactions.

On the other hand, DOX, which forms hydrophobic interactions with only five residues in the binding pocket (I731, l762, F759, F728, and M986), has a very favourable binding energy of -196.862 kcal/mol. The stability in the binding pocket could be attributed to the presence of hydrogen bonds involving five different residues in the docked pose (Figure 4.16), three of which are strong hydrogen bonds of conventional type and two are weak hydrogen bonds of carbon type. In addition, DOX can be stabilised by

the π–sulphur interaction that exists between the π electron cloud of one of the aromatic rings in the structure and the lone electron pair cloud of the sulphur atom in M986. DOX is known to be a substrate of P-gp (Gao *et al.*, 2001; Takara *et al.*, 1999), so the docking results obtained are consistent with the available literature and the experimental transport assay results shown in Figure 4.11.

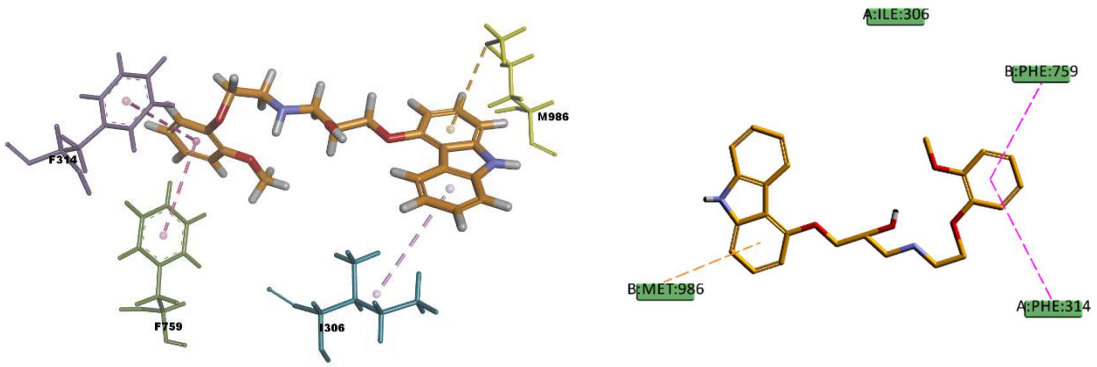


(a) (b)

Figure 4.16: Doxorubicin (DOX) top-ranked pose obtained with the CDOCKER algorithm. (a) 3D view of DOX interactions in the binding pocket. Residues F759, Y307, and Y310 involved in the conventional hydrogen bonds are highlighted in yellow; (b) 2D interaction diagram of DOX with *h*P-gp interacting residues. Green dotted lines represent conventional hydrogen bonds, light-green dotted lines represent carbon hydrogen bonds, light-rose dotted lines represent hydrophobic interactions, pink dotted lines represent π–π T-shaped interactions, and the orange dotted line represents the π–sulphur interaction.

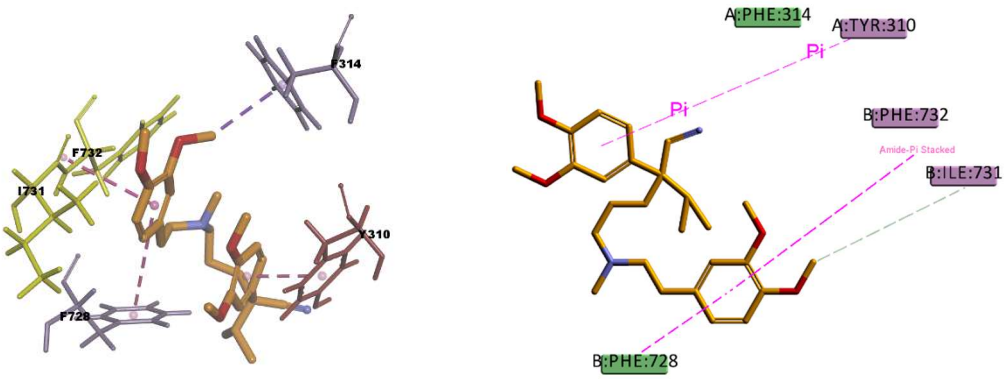
CAR, another well-known substrate of P-gp (Jouan *et al.*, 2016), forms hydrophobic interactions with three residues (I306, F314, and F759) in the binding pocket; these interactions are mainly of π character, two of which are of π–π type and one of π–alkyl type. The docking pose also forms a π–sulphur interaction (Figure 4.17) between the π electron cloud of one of the aromatic rings in the carbazole structure and the lone electron pair cloud of the sulphur atom in M986; it is well known thatπ–sulphur interactions play an important role in chemical and biological recognition as well as in drug development (Beno *et al.*, 2015; Motherwell *et al.*, 2018), thus they can contribute greatly to the stabilization of the molecule in the receptor binding site.

The binding mode of the VER pose (Figure 4.18) shows hydrophobic interactions with four residues (Y310, F314, F732, and F728) in the binding pocket, all π interactions, one π–sigma interaction, two π–π interactions, and one amide $\cdots$ π stacking interaction in which the π–surface of the amide bond between residues I731 and F732 stacks against the π–surface of the aromatic ring in VER. Amide $\cdots$ π stacking interactions are common and significant in protein structures (Harder *et al.*, 2013) and sometimes play an important role in ligand binding (Giroud *et al.*, 2016; Giroud *et al.*, 2017). VER binding mode also involves a weak carbon hydrogen bond with residue I731, which contributes to the stabilization of the ligand–protein complex. The estimated binding energies of CAR and VER, shown in Table 14, indicate that the complexes are stable and exhibit good binding affinity. The docking results are in agreement with the available literature as CAR and VER are known substrates of P-gp(Jouan *et al.*, 2016; Litman *et al.*, 1997).



(a) (b)

Figure 4.17: Carvedilol (CAR) top-ranked pose obtained with the CDOCKER algorithm. (a) 3D view of CAR interactions in the binding pocket. Residue M986 involved in π–sulphur interaction is highlighted in yellow; (b) 2D interaction diagram of CAR with *h*P-gp interacting residues. Light-rose dotted lines represent hydrophobic interactions, pink dotted lines represent π–π T-shaped interactions, and the orange dotted line represents the π–sulphur interaction.

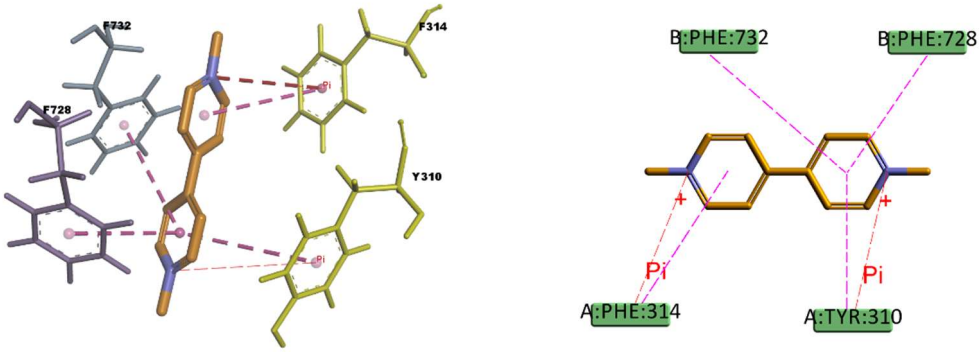


(a) (b)

Figure 4.18: Verapamil (VER) top-ranked pose obtained with the CDOCKER algorithm. (a) 3D view of VER interactions in the binding pocket. Residues I731 and F732 involved in the Amide · · · π stacking are highlighted in yellow; (b) 2D interaction diagram of VER with *h*P-gp interacting residues. The light-green dotted line represents carbon hydrogen bonds, pink dotted lines represent π interactions, and the purple dotted line represents a π–sigma interaction.

Regarding the compounds PQ and GEN, although the experimental transport assay results in Figure 4.11 show that these compounds at 50 µM do not interfere with the P-gp mediated efflux of R123 from Caco-2 cells, the estimated binding energies have very favourable values. The available literature shows mixed results regarding the interaction between PQ and P-gp (e.g., some authors state that PQ is transported by P-gp (B. Wu *et al.*, 2019), while others state that it is not a P-gp substrate (Lacher *et al.*, 2014)). In

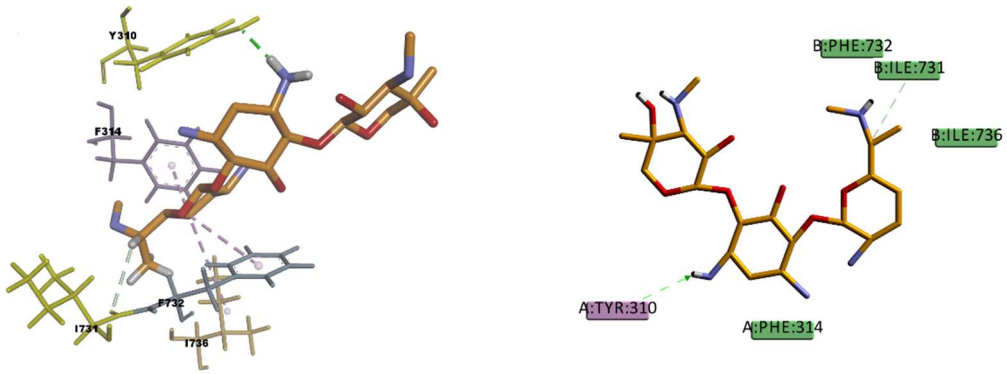
the resulting binding pose (Figure 4.19), PQ forms hydrophobic interactions involving four residues in the binding pocket (Y310, F314, F728, and 732), all π–π type. There is also a cation–π interaction between the positively charged nitrogen of PQ and the polarizable π electron cloud of the aromatic ring in residues F314 and Y310. These are essentially electrostatic interactions due to the negatively charged electron cloud of π systems, which are involved in many drug–receptor interactions, demonstrating that they play an important role in ligand–binding affinity (Salonen *et al.*, 2009).



(a) (b)

Figure 4.19: Paraquat (PQ) top-ranked pose obtained with the CDOCKER algorithm. (a) 3D view of PQ interactions in the binding pocket. Residues F314 and Y310 involved in the cation–π interactions are highlighted in yellow; (b) 2D interaction diagram of PQ with *h*P-gp interacting residues. The pink dotted lines represent π–π T-shaped interactions, and the red dotted lines represent cation–π interactions.

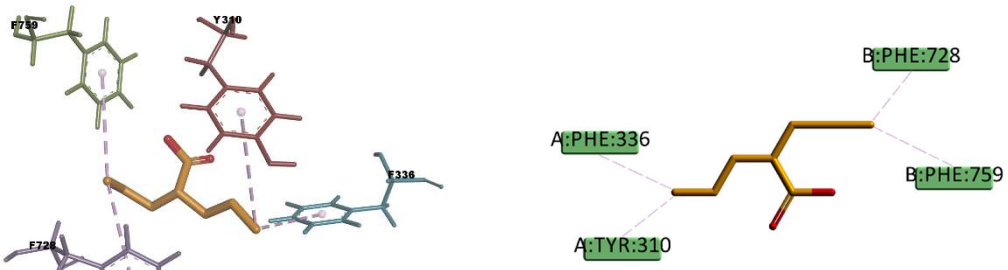
On the other hand, GEN binding pose (Figure 4.20) forms hydrophobic interactions with three residues in the binding pocket (I736, F314, and F732), two π–Alkyl type and one Alkyl type of interaction, besides, it also forms a strong conventional hydrogen bond with residue Y310 and one weak carbon hydrogen bond with residue I731, interactions that could explain the stability in the binding pocket reflected in the favourable binding energy. Despite the binding energies reflecting some stability in the binding site, both PQ and GEN are hydrophilic compounds, a property that may affect the ability of both compounds to reach the binding pocket due to its highly hydrophobic environment.



(a) (b)

Figure 4.20: Gentamicin (GEN) top-ranked pose obtained with the CDOCKER algorithm. (a) 3D view of GEN interactions in the binding pocket. Residues Y310 and I731 involved in the hydrogen bonds are highlighted in yellow; (b) 2D interaction diagram of GEN with *h*P-gp interacting residues. Green dotted line represents conventional hydrogen bonds, the light-green dotted line represents carbon hydrogen bonds, and light-rose dotted lines represent hydrophobic interactions.

Compounds VPA and BU had the highest estimated binding energies in the group of docked compounds and were not transported by P-gp in the experimental transport assay results (Figure 4.11). However, looking at the selected docking pose, VPA forms π–Alkyl interactions with four residues in the binding pocket (Y310, F336, F728, and F759) (Figure 4.21), i.e., interactions that could explain why VPA is described in some literature articles as a P-gp inducer (Eyal *et al.*, 2006) or as an inhibitor with weak affinity (Weiss *et al.*, 2003).

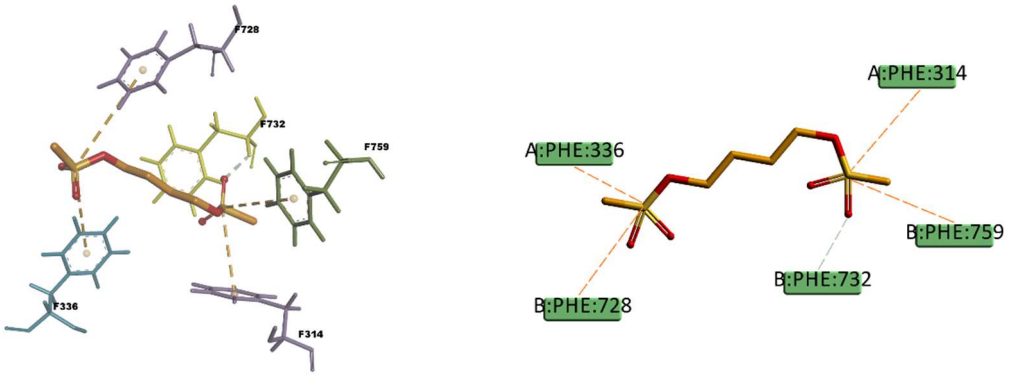


(a) (b)

Figure 4.21: Valproic acid (VPA) top-ranked pose obtained with the CDOCKER algorithm. (a) 3D view of VPA interactions in the binding pocket; (b) 2D interaction diagram of VPA with *h*P-gp interacting residues. Light rose dotted lines represent hydrophobic interactions.

The resulting docked pose of BU (Figure 4.22) involves four π–sulphur interactions with residues F336, F728, F314, and F759, as well as a weak carbon hydrogen bond with residue F732. Due to the slightly hydrophilic nature of BU, under experimental conditions it may have difficulty reaching the binding site, which, as previously mentioned, is located in a highly hydrophobic environment.

The APD compound instead exhibits only one π–donor hydrogen bond between the π electron cloud of the aromatic ring in residue Y310 and the hydrogen atom of the amine group in APD (Figure 4.23). This single interaction seems to confer certain stability to the complex according to the calculated binding energy (Table 4.14), although, the hydrophilic nature of APD can certainly interfere in reaching the binding place, as the results of the experimental transport assay showed no interaction between APD and P-gp.



(a) (b)

Figure 4.22: Busulfan (BU) top-ranked pose obtained with the CDOCKER algorithm. (a) 3D view of BU interactions in the binding pocket. Residue F732 involved in a carbon hydrogen bond interaction is highlighted in yellow; (b) 2D interaction diagram of BU with *h*P-gp interacting residues. The light-green dotted line represents a carbon hydrogen bond interaction, and the orange dotted lines represent π–sulphur interactions.

(a) (b)

Figure 4.23: Pamidronate (APD) top-ranked pose obtained with the CDOCKER algorithm. (a) 3D view of VPA interactions in the binding pocket; (b) 2D interaction diagram of APD with *h*P-gp interacting residues. The light-green dotted line represents π–donor hydrogen bond interactions.

Interestingly, six of the eight known active compounds of P-gp in the docking set (CsA, AM, DIG, DOX, LPM and RMP) showed simultaneous interactions with residues Y307, Y310, F728, and F732 in the binding mode, suggesting that these residues may play a crucial role in ligand recognition and binding. These four residues also interacted with the inhibitor PBDE (polybrominated diphenyl ether)-100 in the co-crystallized structure of *m*P-gp (PDB ID: 4XWK), demonstrating their relevance in ligand binding and the consistency between the amino acids predicted by molecular docking and the available experimental data.

Some compounds, such as PQ and GEN may not interact with the P-gp, as shown in our study, despite the calculated binding energies reflecting some stability in the binding site. Although this could be seen as evidence of the lack of predictability of this approach to identify compounds that interact with P-gp, it should also be noted that these results highlight the importance of the physicochemical properties of the compounds (particularly their lipophilicity), which may prevent them from reaching the binding pocket of P-gp. In addition, it should be noted that the predictive value of the applied model is still very good for compounds that have been found not to bind to P-gp, since for these compounds the ability to reach the binding pocket of P-gp would make no difference.

4.2.4.3.2 Docking into the *h*P-gp cryoEM Structure The set of thirteen compounds was also docked into the experimentally solved cryo–electron microscopy structure of *h*P-gp (PDB ID: 6QEX). The binding region was delineated by the atoms within a radius of 9.5 Å for CDOCKER and GOLD calculations, using the experimental coordinates of the cryoEM *h*P-gp ligand (PDB ID: 6QEX). The selection of the binding site and the settings for the docking calculations were validated via the re-docking procedure (ligand reproduction), obtaining heavy-atom RMDS values of 1.2723 Å, 1.3208 Å, 1.4630 Å for CDOCKER calculations and 1.0283 Å, 1.1974 Å, 1.2669 Å for GOLD calculations. The RMSD results are in agreement with the accepted threshold of 2 Å (see Appendix B, Table B.2, Figure B.2).

The resulting poses are distributed within the TM regions of P-gp (Figure B.5) and show interactions with protein residues of several TM helices located throughout the binding region (3D diagrams of the obtained binding poses can be found in Appendix B, Figure B.6). The interactions of the ligands in the binding pocket in the homology model compared to those in the experimentally determined structure (Table 4.15) are in good agreement. The type of interactions for most compounds are equivalent in both docking studies and share many of the interacting amino acid residues, e.g., CsA shows mainly interactions of hydrophobic character with residues in the binding pocket, and a large number of π interactions, such as π–alkyl interactions in the presence of hydrogen bonds, giving the complex a high stability.

The stability of the complex is also reflected in the calculated values of the binding energy, which are given in Table 4.16. The only compound in the set that had no interacting amino acids in common with the homology model was PQ. Nevertheless, the nature of the interactions agrees well with the results in the homology model, π–cation and the π–π interactions are involved in the stabilization of the complex. The calculated binding energy values are also comparable in terms of the binding stability of the complexes, although the absolute values are about half of the energy values reported in the homology model, e.g., the most stable complex in the set of docked compounds is CsA in both the homology model and the cryoEM structure of *h*P-gp; the less stable complexes are also VPA and BU in both docking systems.

Table 4.15: Ligand–P-gp interactions. Nature of the interactions with P-gp and amino acid residues involved in binding of the compounds under study, using the experimentally solved cryoEM structure of *h*P-gp (PDB ID: 6QEX). Numbers in parenthesis indicate the number of interactions involving the residue.

Name	H-Bond	Alkyl	π–Sigma	π–Alkyl	π–π	π–Sulphur	Others
CsA[1]	**Q725**, **Q990**(2), **A987**, Q347	**A987**, M876,**I340**(2), **M986**(2), M69(2), V991, I306(2)	W232, **F336**	H61, W232(2), F303, **Y307**, **Y310**(2), **F336**(3), **F343**(2), **F728**(2), **F732**, **F983**(2)	-	-	-
AM[2]	-	M986	-	W232, **Y307**, **Y310**, F343, **F728**, M986, L65,	-	M986, M949	M986*, Q990*
DOX[3]	**Y310**, **Q990**(2), **M986**, Q347, A871(2)	L339	-	-	-	**M986**(2)	**M986****
DIG[4]	S344, **Q990**, I340, A871, G872, **F983**	A871, L65, M986	-	**F336**, **F728**, **F983**	-	-	**F732**[§§]
LPM[5]	F303	**L339**, **I340**	-	**I340**, M986	F303	-	-
RMP[6]	**Y307**, **Y310**, Q725, A987, **M986**, W232	**M986**, **I340**(2)	**F728**	W232, F303	**F728**, **F983**	**M986**	**M986****
VER[7]	**Y310**, Q990, Y307	M986	**F728**, Y310, **F732**	-	F336	-	-
CAR[8]	Y310, **F759**(2), F732	-	-	I731, L762	F983(2), F728	**M986**	-
VPA[9]	Q990	M986	-	**F728**, F983	-	-	-
BU[10]	Y310	-	-	-	-	**F728**(2), Y310	**F759**
GEN[11]	**Y310**, 875(3)	-	-	F728, F983	-	-	-
APD[12]	**Y310**, Q724	-	-	-	-	-	-
PQ[13]	E875	-	-	M986(2)	F983	-	F983[§]

[1] Cyclosporine A; [2] amiodarone; [3] doxorubicin; [4] digoxin; [5] loperamide; [6] rifampin; [7] verapamil; [8] carvedilol; [9] valproic acid; [10] busulfan; [11] gentamicin; [12] pamidronate; [13] paraquat. Residues in bold are the shared residues in both docking systems. * Halogen interaction; ** sulphur-X interaction; [§] cation–π interaction; [§§] π–lone pair interaction. The bold printed amino acid residues identify those which are also involved in binding interactions according to the homology model

Table 4.16: Free energies of binding. Estimate of the overall binding free energies of the compounds under study, using the experimentally solved cryo–electron microscopy structure of *h*P-gp (PDB ID: 6QEX).

Name	**CDOCKER Binding Energy (kcal/mol)**	**GOLD Binding Energy (kcal/mol)**
CsA[1]	-133.400	-149.461
AM[2]	-67.974	-64.093
DOX[3]	-133.886	-149.007
DIG[4]	-89.074	-111.314
LPM[5]	-88.284	-73.676
RMP[6]	-104.424	-142.062
VER[7]	-80.413	-70.402
CAR[8]	-70.329	-68.845
VPA[9]	-29.039	-30.858
BU[10]	-43.311	-37.354
GEN[11]	-101.718	-99.725
APD[12]	-69.750	-89.721
PQ[13]	-112.572	-100.914

[1] Cyclosporine A; [2] amiodarone; [3] doxorubicin; [4] digoxin; [5] loperamide; [6] rifampin; [7] verapamil; [8] carvedilol; [9] valproic acid; [10] busulfan; [11] gentamicin; [12] pamidronate; [13] paraquat.

4.2.5 Conclusions

The quality assessment of the developed *h*P-gp models suggests that the overall folding of the 3D structure is as good as the available crystal structure of the *m*P-gp (PDB ID: 4M1M) and is therefore reliable and suitable for further *in silico* structure-based studies. The employed method was capable to generate a *h*P-gp model similar to the near-native *m*P-gp. The characterization of the binding pocket of our homology model revealed a large hydrophobic surface area. These results are consistent with biochemical investigations which concluded that the drug-binding pocket is assembled mostly of hydrophobic residues, creating a lipophilic environment. *In silico* QSAR models that classify between active and non-active compounds generally include physicochemical descriptors that measure the lipophilicity of the molecules, such as logP or log D (pH 7.4), which correlates well with our observations.

The analysis of the binding poses suggests that the large number of π interactions together with the simultaneous presence of hydrogen bond interactions contribute to the stability of the ligand–protein complex in the binding site; the hydrogen bond interactions present are mostly of weak character (carbon hydrogen bonds type C–H . . . O) with a larger dispersive component. Identical interacting amino acid residues observed in the *m*P-gp crystal structure (PDB ID: 4XWK) and the *h*P-gp cryoEM structure (PDB ID: 6QEX) contribute to drug binding in the *h*P-gp homology model, e.g., Q725, Y307, F983, which occur in both the *m*P-gp crystal structure and the *h*P-gp cryoEM structure.

Some amino acid residues that contribute to drug binding in our *h*P-gp homology model are also present in one or the other of the two experimentally solved structures, e.g. amino acid residues Q990, A987, I340, M986, F343 are also interacting residues in

the *h*P-gp cryoEM structure, while residues Y310, F314, F728, F732, F759 are interacting residues in the *m*P-gp crystal structure, which demonstrates the consistency between the interacting amino acid residues predicted by molecular docking calculations and the co-crystallized data available.

The study of the ligand–*h*P-gp complexes provides considerable insight into the drug binding mode for the set of investigated ligands. Different modes of interaction for different classes of compounds could be uncovered; therefore, molecular docking studies in this area should not be underestimated. In some cases, experimental data are highly controversial; therefore, the combination of computational methods and experimental data from efflux pump transport assays is essential in the complex field of P-glycoprotein.

Chapter 5

Structure-Based Modelling Approach

5.1 Introduction

Of the structure-based methods that can be applied in the field of computer-aided drug design (CADD), molecular dynamics simulations (MD), first developed in the late 1970s (McCammon *et al.*, 1977), is a methodology that can provide insight into protein dynamics beyond the information available from the static models generated by nuclear magnetic resonance (NMR), X-ray crystallography, cryo-electron microscopy (cryo-EM), and homology modelling. Since molecular recognition and drug binding are highly dynamic processes, molecular dynamics can be used to explore the associated conformational space and is often the method of choice for studying large molecules such as proteins.

In general, this approach replaces the static model with a dynamic model in which the system is set in motion. The simulation of the motion is performed by solving Newton's dynamic equations. First, a computational model of the molecular system is created from NMR, crystallographic, cryo-EM or homology modelling data. Then, the forces acting on each atom by all other atoms in the system are estimated and inserted into Newtonian equations of motion to predict the spatial position of each atom as a function of time. These equations are solved iteratively for each particle in the system to calculate the forces on each atom and then use these forces to update the position and velocity of each atom. The result is a 3D trajectory that describes the atomic-level configuration of the system at each point during the simulation time (Durrant *et al.*, 2011).

The forces in a MD simulation are calculated using a molecular mechanics force field, which is a model that is parametrized to fit the results of quantum mechanical calculations and experimental data (Cornell *et al.*, 1995), e.g., a force field contains terms describing electrostatic (Coulomb) interactions between atoms, spring-like terms modelling the length of covalent bonds, and terms capturing various other types of interatomic interactions, such as proper van der Waals atomic radii. Comparison of simulations with a variety of experimental data shows that the force fields have improved considerably over the last decade (Lindorff-Larsen *et al.*, 2012), nonetheless, the uncertainty introduced by these approximations should still be taken into account when analysing simulation results. The time steps in a MD simulation are typically on the order of 1 or 2 femtoseconds. Most biochemical events of interest occur on time scales of nanoseconds, microseconds, or even longer, so a typical simulation involves millions or billions of time steps. This fact, combined with the millions of interatomic interactions evaluated during a single time step, makes MD simulations computationally demanding.

Because MD simulations can predict how each atom in a protein or other molecular system will move over time (Karplus *et al.*, 2002), they can capture important

biomolecular processes with temporal resolution that is not available experimentally (Zhao *et al.*, 2015), including conformational changes, ligand binding, and protein folding. These simulations can also predict how biomolecules respond at the atomic level to perturbations such as mutation, phosphorylation, protonation, or the addition or removal of a ligand.

The application of MD in drug discovery is emerging as an important tool for understanding the physical basis of the structure and function of biological macromolecules. The previous view of proteins as relatively rigid structures has been replaced by a dynamic model in which internal motions and the resulting conformational changes play an essential role in their function. Therefore, in this chapter, we describe the application of the molecular dynamics technique to the study of P-gp using the human P-gp (*h*P-gp) 3D structure in an explicit membrane and water environment with the aim of better understanding the driving forces of ligand recognition and transport mechanism at the atomistic level.

5.2 Structure-Function Relationships in ABCB1: Insights from Molecular Dynamics Simulations

5.2.1 Abstract

P-glycoprotein (P-gp) is a transmembrane protein belonging to the ATP binding cassette superfamily of transporters and it is a xenobiotic efflux pump that limits intracellular drug accumulation by pumping compounds out of cells. P-gp contributes to a reduction in toxicity and has broad substrate specificity. It is involved in the failure of many cancer and antiviral chemotherapies due to the phenomenon of multidrug resistance (MDR), in which the membrane transporter removes chemotherapeutic drugs from target cells. Understanding the details of the ligand–P-gp interaction is therefore critical for the development of drugs that could overcome the MDR phenomenon, for early identification of P-gp substrates that will help to obtain more effective prediction of toxicity, or for subsequent outdesign of substrate properties if needed.

In this work, a series of molecular dynamics (MD) simulations of human P-gp (*h*P-gp) in an explicit membrane and water environment were performed to investigate the effects of binding different compounds on the conformational dynamics of P-gp. The results showed significant differences in the behaviour of P-gp in the presence of active and non-active compounds within the binding pocket. Different patterns of movement were identified which could be correlated with the conformational changes leading to the activation of the translocation mechanism. The results of this work also suggest an asymmetry in the motion patterns of the nucleotide binding domains (NBDs), since the changes that trigger the translocation mechanism only start at NBD1. A hypothesis linking the ability to generate conformational changes that regulate the volume of the binding pocket (induce fit model) to the ability to be transported by P-gp is also proposed. The estimated binding free energies of the studied complexes and the predicted ligand–P-gp interactions are in good agreement with the available experimental data, demonstrating the validity of the results derived from the MD simulations.

5.2.2 Introduction

P-glycoprotein (P-gp) is one of the most studied membrane transporters that belong to the ATP-binding cassette (ABC) superfamily, probably because of the role it plays in

multidrug resistance (MDR), a phenomenon in which there is cellular resistance to a variety of structurally and functionally unrelated chemotherapeutic agents. The identification of P-gp more than three decades ago enabled the discovery that reduced intracellular accumulation of anticancer drugs can lead to significant levels of drug resistance; (Lockhart *et al.*, 2003; Lum *et al.*, 1993). Over the years, P-gp has become the prototype MDR transporter, with studies concluding that drug resistance associated with P-gp leads to major failures of chemotherapy in human cancers (Juliano *et al.*, 1976).

P-gp is a 1,280-residue single polypeptide with a molecular weight of 170 kDa. It contains two transmembrane domains (TMDs) and two cytosolic ATP-binding regions called nucleotide binding domains (NBDs). These two symmetrical halves are connected by a highly charged "linker region" of ∼75 amino acids in length (Higgins *et al.*, 1997). Each TMD consists of six highly hydrophobic α-helices embedded in the membrane bilayer and extending into the cytosol to form the intracellular loops (ICLs). The TMDs form the pathway by which drug molecules cross the membrane, switching between inward and outward facing conformations (Zhou, 2008); although structurally similar across the transporter family, they have a large proportion of non-conserved amino acids. In contrast, the NBDs contain three highly conserved sequence motifs at which ATP is hydrolysed: the Walker A and Walker B motifs, and the ABC signature motif. The actual nucleotide binding site is shared by the NBDs and therefore arises only when they dimerize, consisting of the Walker A motif of one NBD and the ABC motif of the other NBD. The vast majority of recent studies assume an alternate cycle for the hydrolysis of ATP, implying that it occurs at alternate sites in two distinct steps (George *et al.*, 2012; P. M. Jones *et al.*, 2009), showing a possible conformational asymmetry at the NBDs.

Although it has often been assumed that ATP hydrolysis drives the transport process, conformational changes leading to substrate transport occur after ATP binding and not after ATP hydrolysis. Similarly, the reduction in binding affinity of drugs to P-gp seems to be due to ATP binding, which induce conformational changes that expose the binding site to the extracellular medium, rather than hydrolysis (Martin *et al.*, 2000; Martin *et al.*, 2001; Rosenberg *et al.*, 2001). The ATP hydrolysis may simply "reset" the transporter to the initial conformation (Sauna *et al.*, 2007), allowing a new catalytic cycle to occur. The P-gp ATPase activity and transport cycle are tightly coupled to substrate binding and can be stimulated or modulated by substrates or inhibitors, respectively (Eckford *et al.*, 2009), but the molecular mechanisms of how this occurs have not been fully elucidated. From this perspective, structural and dynamic insights into P-gp transport are essential to better understand how this transporter functions and, in this way, facilitate the development of inhibitors relevant to the clinical practice that could overcome MDR, or the identification of P-gp ligands that could contribute to more effective prediction of toxicity at early stages of drug development.

The molecular understanding of the transport mechanism of P-gp depends on the available experimental structural information that provides snapshots of the conformational cycle related to substrate transport. The diversity of conformations in which P-gp has been crystallised demonstrates the flexibility of the transporter (Alam *et al.*, 2019; Y. Kim *et al.*, 2018), a property associated with its polyspecificity, i.e., the ability to bind a large number of chemically diverse compounds. Over time, numerous ligand-based models have been developed to predict the P-gp activity, such as, quantitative structure–activity relationship (QSAR) models (Broccatelli *et al.*, 2011; L. Chen *et al.*, 2011; D. Li *et al.*, 2014; Mora Lagares, Minovski, & Novič, 2019) as well as structure-based models based on molecular docking and the use of the mouse P-gp (*m*P-gp) 3D structure (Dolghih *et al.*, 2011; Ricardo J. Ferreira *et al.*, 2013b), or homology models of the human P-gp (*h*P-gp) (Mora Lagares *et al.*, 2020), in which the binding

modes of substrates and inhibitors have been well characterized. Nonetheless, the molecular details of the effects of these molecules on the conformational dynamics of P-gp are not well understood. Therefore, methods to study the dynamics at the atomistic level are needed to gain a better understanding of the conformational changes that P-gp undergoes during the translocation mechanism.

Under this framework, MD simulations have proven useful in biological and chemical studies because they can provide structural and dynamical information at the atomistic level. MD simulations have already been used to study P-gp drug interactions and other dynamic processes of P-gp (Ricardo J Ferreira *et al.*, 2012). Some studies have highlighted the potential of P-gp to accommodate multiple drug molecules simultaneously in the binding pocket, and others have described the importance of a lipid bilayer environment in the study of P-gp (O'Mara *et al.*, 2012; B. Zhang *et al.*, 2021) revealing important details for the study of this transporter. However, since the human crystal structure of P-gp was not available at that time, the reported studies were performed either on homology models or on the crystal structure of mouse P-gp. Recently, in 2019, the cryo-electron microscopy (cryoEM) structure of *h*P-gp in the inward facing conformation was solved (PDB ID: 6QEX) (Alam *et al.*, 2019), which raised much hope for better and more efficient development in the study of the P-gp.

In this work, we describe a series of MD simulations based on the human cryo-EM structure of P-gp, aimed at understanding the behaviour of different molecules (substrates, inhibitors, and non-active compounds) within the binding pocket and evaluating their effects on the dynamics and conformations of NBDs and TMDs at the atomistic level. Another aim of the study is to identify patterns of movement exhibited by the transporter in the context of the translocation process, which will allow a better understanding of the initial steps of the efflux mechanism.

5.2.3 Materials and Methods

Molecular dynamics simulations were performed on a series of ligand–P-gp complexes formed by nine different compounds, including four drugs known to interact with P-gp as substrates, inhibitors, or both: cyclosporine A (CSA), a high-affinity substrate (Saeki *et al.*, 1993) and inhibitor (Wigler, 1999) of P-gp, amiodarone (AMI) (Jouan *et al.*, 2016), doxorubicin (DOX) (Gao *et al.*, 2001; Takara *et al.*, 1999), and carvedilol (CAR) (Jouan *et al.*, 2016), well-known substrates of P-gp; and five compounds that do not interact with P-gp: pamidronate (APD), busulfan (BUS), gentamicin (GEN), paraquat (PQT) (Lacher *et al.*, 2014), and valproic acid (VPA). We used the available knowledge on the interaction of substrates and inhibitors with P-gp to obtain a detailed description of the dynamics of the *h*P-gp at the atomistic level, which was previously limited to studies based on the structure of *m*P-gp or homology models of *h*P-gp.

5.2.3.1 Preparation of initial structures

The initial structure of P-gp used for the simulations corresponds to the cryoEM structure of the *h*P-gp (PDB ID: 6QEX) (Alam *et al.*, 2019). The starting configuration for the MD simulations of the ligand–P-gp complexes was obtained from previous docking calculations performed using the CDOCKER algorithm (G. Wu *et al.*, 2003) within the Dock Ligands protocol in Discovery Studio 4.1 (Accelrys, 2017), reported in (Mora Lagares *et al.*, 2020). The starting complexes were parametrized using the CHARMM36 force field through the CHARMM-GUI web-based graphical user interface (Jo *et al.*, 2014; Jo *et al.*, 2008).

5.2.3.2 Systems construction

After parametrization of the ligand–P-gp complexes, the systems were partially solvated using the TIP3P explicit water model with the software DOWSER (L. Zhang *et al.*, 1996) and SOLVATE (Heller *et al.*, 1993). The water molecules located in the hydrophobic region of the protein were then removed. The partially solvated systems were embedded into a 1-palmitoyl-2-oleoyl-phosphatidylcholine (POPC) membrane patch constructed using the membrane plugin from VMD (Humphrey *et al.*, 1996) and parametrized using the CHARMM27 force field. The system was placed in the centre of the POPC lipid bilayer with its long axis perpendicular to the lipid surface. Finally, the entire system was solvated using VMD's Solvate plugin to generate the water simulation box. Potassium (K^+) and chloride (Cl^-) ions were then added at a concentration of 0.2 M to neutralize the system using VMD's Autoionize plugin.

5.2.3.3 Molecular dynamics simulations

Since the MD simulations were performed using Amber 2018 software (D.A. Case *et al.*, 2018), the tool CHAMBER (M. F. Crowley *et al.*, 2009) was employed to convert CHARMM files into AMBER format and enable the use of the CHARMM force field within the AMBER's MD engines.

5.2.3.3.1 Simulation parameters Periodic boundary conditions (PBC) were applied in all the simulations. The bonds involving hydrogen atoms were constrained using the SHAKE algorithm. The long-range electrostatic interactions were estimated using the Particle-Mesh Ewald (PME) method and the non-bonded cutoff radius for the van de Waals and electrostatic interactions was set to 10 Å. The temperature of the system was equilibrated at 303 K with a collision frequency of 1.0 ps^{-1} using the Langevin thermostat. For the NPT runs, the pressure was maintained at 1 bar using the Berendsen barostat. The system pressure was semi-isotropically coupled, i.e., the pressure was balanced by separately coupling the lateral (x and y) and normal (z) box directions. Cpptraj (Roe *et al.*, 2013) module of AmberTools (D.A. Case *et al.*, 2018), VMD (Humphrey *et al.*, 1996), UCSF Chimera (Pettersen *et al.*, 2004), MDtraj (McGibbon *et al.*, 2015), and POVME 3.0 (Wagner *et al.*, 2017) were used for trajectory analyses.

5.2.3.3.2 Energy minimization An energy minimization run of 6,000 cycles was performed in three successive steps of 2,000 cycles each using the steepest descent method: first, restraints were applied to the whole system with a force constant of 100 kcal/mol $Å^2$, with only the water molecules moving freely, followed by a second cycle with restrains applied only on the ligand and protein. In the third step, the entire system was minimized without the application of any restrains.

5.2.3.3.3 Heating After the minimization steps, the system was gradually heated to 303 K in the NVT ensemble for 1 ns, applying a restrain force of 100 kcal/mol $Å^2$ to the protein, ligand, and POPC lipid bilayer. The heating was performed in two steps: the first 500 ps the system was heated from 0 to 150 K, the last 500 ps from 150 K to 303 K.

5.2.3.3.4 Equilibration The equilibration phase was performed in the NPT ensemble with a restrain force of 100 kcal/mol $Å^2$ for a total of 30 ns, divided into five steps: first 20 ns with restrains on the protein, ligand and POPC lipid bilayer, then 2.5 ns with restrains on the protein and ligand. Finally, three consecutive 2.5 ns NPT runs were performed to progressively remove the restrains on the protein: first with restrains on the

protein except Cβ, then with restrains on the protein except Cα and Cβ, followed by a final 2.5 ns equilibration run without any restrains. This equilibrated system was the starting point for the 500 ns completely unrestrained NPT production run.

5.2.3.3.5 Production The production phase was performed in the NTP ensemble at a temperature of 303 K and a pressure of 1 bar for 500 ns. The time step was set to 2 fs and the trajectories were saved every 90,000 steps (180 ps).

5.2.3.4 Trajectory analysis

To analyse the average deviations in the atomic positions and the stability of the MD trajectories, the atomic root-mean-square deviations (RMSD) of each system were calculated using the Cpptraj module (Roe *et al.*, 2013) of the AmberTools package (D.A. Case *et al.*, 2018). For the RMSD calculation, each frame of the trajectory was aligned along the protein backbone relative to the initial structure.

5.2.3.4.1 Binding free energy Calculations The free energies of binding of the ligand–P-gp complexes studied were estimated using the Molecular Mechanics/Poisson-Boltzmann Surface Area (MM/PBSA) method (Srinivasan *et al.*, 1998) based on 100 frames extracted from each trajectory and selected by clustering analysis. Only the heavy atoms of the ligands were considered for clustering. The representative MD snapshots were selected as the centroids of 100 clusters resulting from grouping all the ligand conformations obtained from the production runs. The K-NN algorithm as implemented in cpptraj was employed for cluster analyses.

The MM/PBSA calculations were performed using the Python script MMPBSA.py (Miller III *et al.*, 2012) implemented in Amber 2018, and the parameters used were set as follows: the lipid membrane was treated implicitly, therefore POPC lipid bilayer, water molecules, and ions were removed from the trajectories. The solute dielectric constant was set to 4 and the implicit solvent dielectric constant was set to 80. A heterogeneous membrane dielectric constant varying from 1 in its centre to 80 in the edge of the membrane (memopt=3) was used. The ionic strength was set to 0.15 M.

The binding free energies are calculated by subtracting the free energies of the unbound receptor and the ligand from the free energy of the bound complex, as shown in Equation (5.1).

$$\Delta G_{binding} = \Delta G_{Complex} - (\Delta G_{Protein} + \Delta G_{Ligand}) \tag{5.1}$$

The free energy change associated with each term for the protein, ligand and complex is estimated according to Equation (5.2)

$$\Delta G = E_{Gas} + \Delta G_{Solvation} - TS_{Solute} \tag{5.2}$$

where the E_{Gas} term includes the sum of the bonded and non-bonded interaction energies (V^{vdw} and V^{ele}). The gas phase energies are often the molecular mechanical (MM) energies from the force field, while the solvation free energies ($\Delta G_{Solvation}$), which include G^{polar} and $G^{non\text{-}polar}$ contributions, are calculated using an implicit continuum solvent model, representing the change in free energy due to the conversion of a solute in vacuum to its solvated state. In MM/PBSA, G^{polar} is obtained by solving the Poisson–Boltzmann equation, while the $G^{non\text{-}polar}$ term is often approximated by a term of the solvent accessible surface area (SASA). T and S represent the temperature of the system and the entropy of the solute in vacuum. The TS terms, which enter Equation (5.1) via the individual ΔG terms, account for the change in conformational entropy of the

ligand–protein complex upon ligand binding and are estimated using known approximations, e.g., Normal Mode Analysis of the protein–ligand coordinates resulting from the simulation of the complex. Often the entropic term is neglected because there is more interest in the relative free energies of binding for comparison with similar systems than in the true free energy values.

The energies described in the equations above are single point energies of the system. However, in practice, these energies are calculated according to averages from an ensemble of representative structures. Expressing Equation (5.2) in terms of averages yields to Equation (5.3)

$$\begin{aligned} \Delta G \; &\cong \langle E_{Gas} \rangle \; + \langle \Delta G_{Solvation} \rangle - \langle TS_{Solute} \rangle \\ &= \; \frac{1}{N} \sum_{i=1}^{N} E_{i,Gas} + \frac{1}{N} \sum_{i=1}^{N} G_{i,Solvation} - \frac{T}{N} \sum_{i=1}^{N} S_{i,Solute} \end{aligned} \tag{5.3}$$

where i is the index of the frame and N is the total number of frames analysed.

5.2.3.4.2 Ligand–Protein Interactions The frequencies of ligand–protein interactions were explored on the 500 ns production trajectories of each system using the structureViz2 Cytoscape plugin (J. H. Morris *et al.*, 2007) and UCSF Chimera software (Pettersen *et al.*, 2004). The analysis of the relevant interactions in each system was performed based on the interactions observed in at least 50% of the snapshots.

5.2.3.4.3 Clustering analysis Clustering analysis was performed on the 500 ns production trajectories of each system using the *Cluster* analysis command of the Cpptraj module (Roe *et al.*, 2013) in the AmberTools package (D.A. Case *et al.*, 2018), with the goal of determining structural populations from the simulated trajectories. Clustering is a way of partitioning data such that data points within a cluster are more similar to each other than to points outside a cluster, i.e., similar conformations are grouped together. The similarity between members of a cluster is determined by a distance metric, usually the coordinates RMSD.

A number of clusters, ranging from two to twenty, were analysed for each system studied and the behaviour of the metrics DBI, pSF, and SSR/SST (Shao *et al.*, 2007) was examined to determine the optimal number of clusters for each system. The DBI and pSF values are metrics of clustering quality; low DBI and high pSF values indicate better results. On the other hand, the R-squared value (SSR/SST) represents the percentage of variance explained by the data. Theoretically, the optimal number of clusters is reached when DBI has a minimum, pSF shows a maximum, and the SSR/SST plot reaches a plateau. These conditions are difficult to satisfy simultaneously in a real clustering problem, so a balance between them was made to select the optimal number of clusters. The corresponding plots and the number of clusters selected can be found in Appendix C, Figure C.5.

5.2.3.4.4 Principal Component Analysis Principal component analysis (PCA) was performed to analyse the conformational changes and dominant modes of motion of P-gp during the 500 ns production run. PCA is a method used to transform a set of potentially coordinated observations into a set of orthogonal vectors: the principal components (PCs). The PCs explain the variance in the data, with the first PC carrying the largest variance, the second PC the second largest, and so on.

The input to the PCA was the covariance matrix calculated from the time series of the position coordinates, such that the PCs represent specific modes of motion of the

system, with the first PC representing the dominant motion. The entries of this matrix are the covariance values between the X, Y and Z components of each atom, so that the final matrix has a size of 3N x 3N, where N is the number of atoms. Since, we are only interested in the internal dynamics of the system, the rotational and translational movements were removed by performing a coordinate RMS fit to a reference structure: the *h*P-gp cryoEM structure. After removing all translations and rotations, the covariance matrix was constructed based on the Cα coordinates.

The covariance matrix was then diagonalized to obtain the eigenvectors (PCs) and the eigenvalues (weight of each PC) describing the principal modes of structural variation. The diagonalization process results in an orthogonal set of unitary vectors describing the directions of maximum variation in the observed conformational distribution. Each eigenvector is associated with an eigenvalue that determines the amplitude of the movements along each principal axis. The sum of all the eigenvalues is considered as the total conformational variance of the system and then the relative contribution of each eigenvector to the total system flexibility is calculated as the ratio between its associated eigenvalue and the total system flexibility. PCA is useful for gaining insight into the dynamics of a system, but it should be kept in mind that the actual motion of the system over the course of a simulation is almost always a combination of the individual PCs.

PCA was performed using the Cpptraj module of the AmberTools package and visualization of principal component data was done using the Normal Mode Wizard (NMwiz) plugin (Bakan *et al.*, 2011) for VMD.

5.2.3.4.5 Binding pocket volume Binding pocket volume calculations were performed using the tool Pocket Volume Measurer POVME 3.0 (Wagner *et al.*, 2017). The POVME algorithm calculates the pocket volume by subtracting the volume occupied by the protein atoms in each frame from the defined inclusion region, with this region defining the boundaries of the pocket. The analysis of the binding pocket volume for each system was performed on input PDB trajectories containing 462 frames extracted from each individual production trajectory. The pocket inclusion region was ligand-defined, i.e., using the ligand residue name, the pocket was defined in all grid points within 3 Å of the ligand atoms in the loaded PDB trajectory.

In order to find representative binding pocket conformations and to determine the average pocket shape in each group of systems, a pocket shape clustering procedure was also performed. For this purpose, the ligand was removed from the trajectories, and they were combined to obtain two new sets of trajectories: one with the combined trajectories of the active-bound systems and one with the combined trajectories of the non-active-bound systems. The trajectories were aligned, and the inclusion region was geometrically defined using x, y, z coordinates and a radius value to define the location of the binding pocket.

The clustering of pocket shape is completed in two steps: First, the similarity matrix of the binding pockets of all analysed frames is calculated, followed by the clustering of this similarity matrix. The similarity matrix was calculated using the Tanimoto overlap score of each pocket pair. The Tanimoto score of a pair of frames ranges from 0 (the two pockets have no volume in common) to 1 (the two pocket shapes are identical). The similarity matrix was clustered using the hierarchical clustering method and the resulting average pocket shape of each group was visualized in VMD.

5.2.3.4.6 Solvent accessible surface area (SASA) The total SASA was calculated from the 500 ns production trajectories of each system using the Surf action command of

the Cpptraj module (Roe *et al.*, 2013) in the AmberTools package (D.A. Case *et al.*, 2018), which calculates the surface area in $Å^2$ using the Linear Combinations of Pairwise Overlaps (LCPO) algorithm (Weiser *et al.*, 1999).

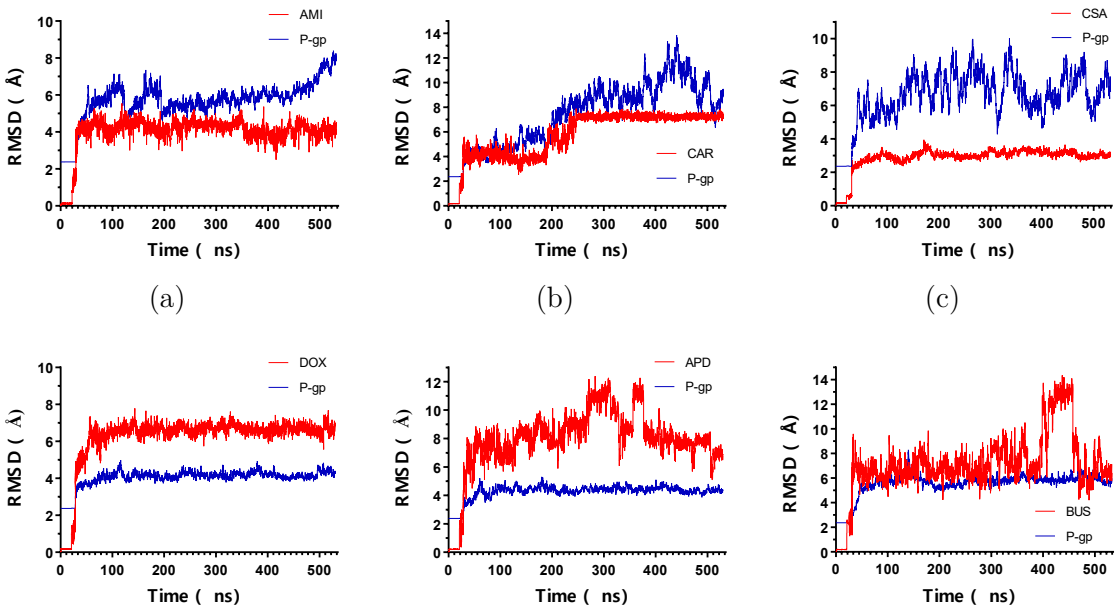
The per residue SASA instead was calculated using the Python library MDtraj (McGibbon *et al.*, 2015).

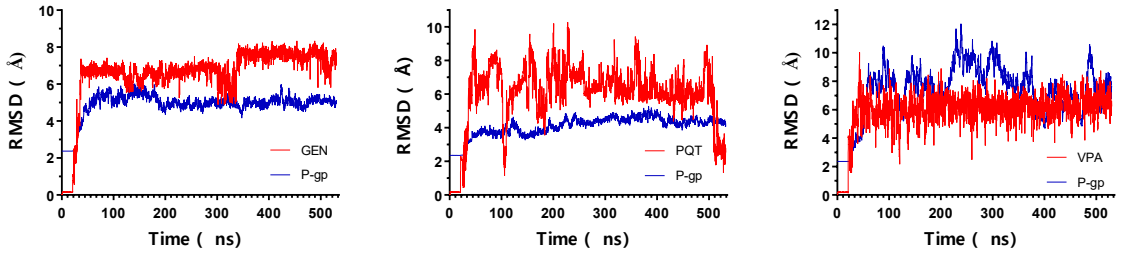
5.2.4 Results and Discussion

5.2.4.1 Overall systems dynamics

All the MD simulations were performed on systems that have reached a state of energetic equilibrium (see Appendix C, Figure C.6).

Figure 5.1 shows the backbone RMSD for all systems during the 530 ns time of simulation. As can be seen, the protein backbone RMSD shows a rapid increase in the first 30 ns, followed by a stable fluctuation throughout the course of all simulations, indicating that the conformational equilibrium has been reached. The trajectories are stable during the 500 ns production run, fluctuating within a range of 3 Å after the initial equilibration phases for most complexes. As for the ligands, they remained docked at their respective binding sites during the simulation time, resulting in very low RMSD values. For the P-gp active ligands, their RMSD values were indicative of the stability of the complexes, as the fluctuations were minimal and ranged within 1 Å, whereas the RMSD for the non-active compounds showed larger fluctuations along the entire simulation. In general, the low positional deviations of the active ligands suggest a stable binding. Interestingly, for CAR, there is an increase in RMSD values at about 180-250 ns, which then stabilizes again until the end of the simulation. Looking at the trajectory, this coincides with a change in the conformation of CAR within the binding pocket: the molecule rotates such that the carbazole moiety interacts with the residue Q990 and the methoxyphenyl ring interacts with the residue I340. This conformation is maintained until the end of the simulation. Previously, residue Q990 was interacting with the methoxyphenyl ring of CAR and its carbazole moiety with I340.

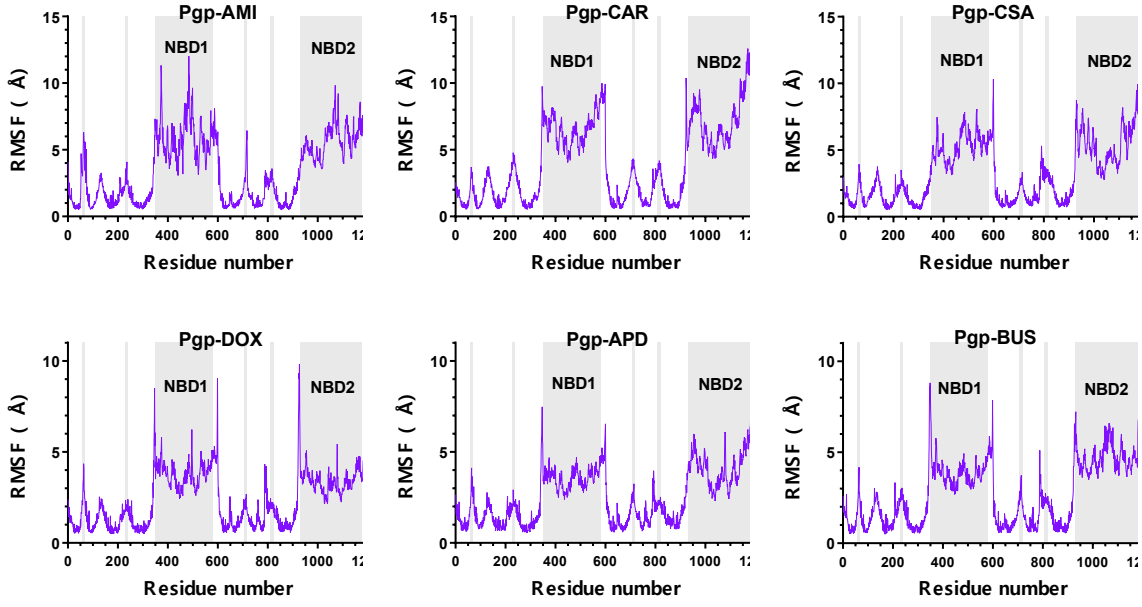




(g) (h) (i)

Figure 5.1: RMSD vs time of the simulated P-gp–ligand complexes: backbone atoms of protein chains (blue) and ligand (red). (a) P-gp–AMI; (b) P-gp–CAR; (c) P-gp–CSA; (d) P-gp–DOX; (e) P-gp–APD; (f) P-gp–BUS; (g) P-gp–GEN; (h) P-gp–PQT; (i) P-gp–VPA.

The high stability of the simulated complexes is further illustrated by the root-mean-square fluctuation (RMSF) plots (Figure 5.2), which show the fluctuations of each individual atom around its average position. It can be seen from the plots that the behaviour of the RMSF is similar for all systems. Most of the residues of the protein fluctuated less than 3 Å relative to the average structure and the RMSF peaks observed are associated with the same specific regions in all the complexes studied. Loop regions and the NBDs exhibited the highest fluctuations, whereas the helices of the transmembrane domain remained very stable throughout the entire simulation. In addition to the NBDs region, the highest values of atomic fluctuations were also located at one of the extracellular loops (approximately residues 90-100). The flexibility of these segments is important for the transport mechanism because it allows a rapid closure of the outward-facing conformation and in this way prevents re-entry of the substrate into the translocation pathway (Y. Kim *et al.*, 2018). Smaller peaks were also observed for some of the intracellular loops between TM4-TM5, TM8-TM9, TM10-TM11, and the helix breakers of TM4 and TM10. In the inward-facing conformation, TM4 and TM10 are interrupted by flexible loops that are thought to act as flexible hinges to open the drug-translocation pathway. Therefore, the observed coordinates fluctuations are in good agreement with the available literature (Jin *et al.*, 2012; Kodan *et al.*, 2014).



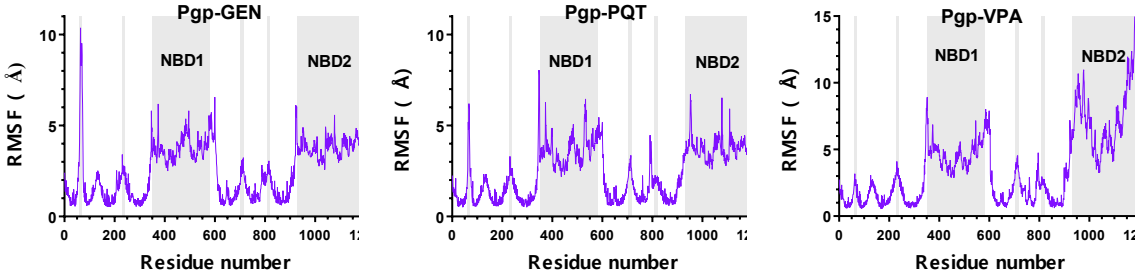


Figure 5.2: Per residue RMSF for the simulated ligand–P-gp complexes. The residue numbers do not correspond to the IDs in the PDB file of the cryo-EM structure, but are consecutive as required by the simulation software.

Figure 5.3 summarizes the flexibility of the protein backbone in each simulated complex. Although the behaviour of the RMSF of the protein backbone is similar in all systems, there are some differences in the magnitude of the flexibility of the NBDs. The complexes formed by P-gp and active compounds show a higher flexibility of NBDs compared to the non-active-bound complexes. This result is consistent with the hypothesis for the transport mechanism, as a higher movement of the NBDs is expected for the activation of the translocation pathway. It is noteworthy that two molecules in the studied group exhibited a different behaviour, namely P-gp–DOX, whose NBDs were less flexible compared to the other active-bound complexes, and P-gp–VPA, which displayed higher flexibility only in one of its NBDs.

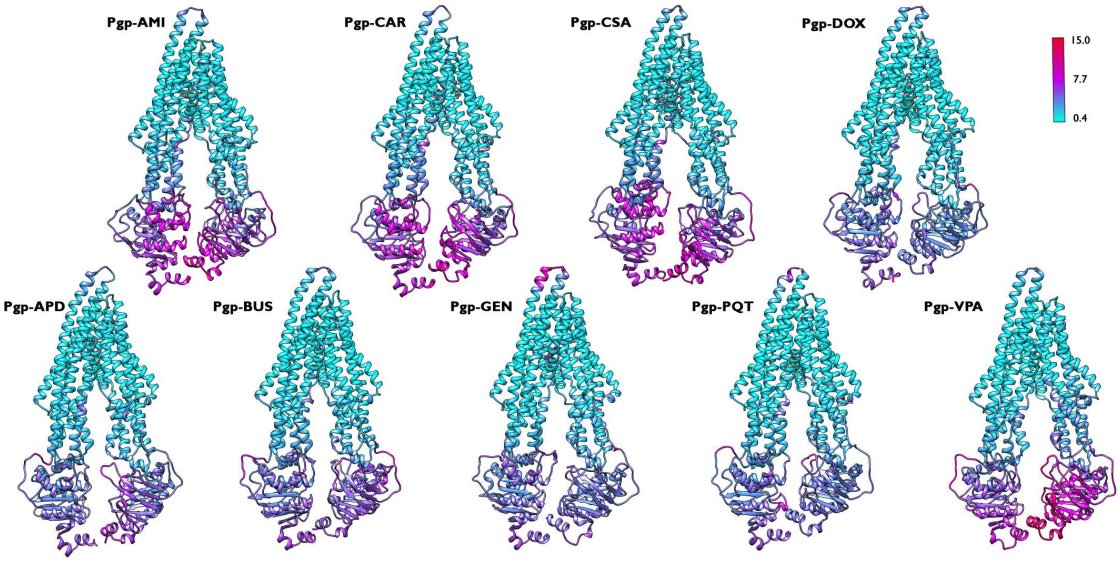


Figure 5.3: Backbone RMSF coloured representation for the simulated P-gp–ligand systems along the 500 ns production run. The flexibility scale goes from cyan (lower values) to red (higher values). The same regions are among the most flexible in all the studied systems; however, the flexibility is higher for the active complexes.

5.2.4.2 Ligand–Protein Interactions

Only the interactions observed in more than 50% of the production snapshots are considered in the next analysis. The ligand–P-gp interactions during the 500 ns production run mainly involved residues in TM1, TM4, TM5, TM6, TM12, and to a lesser extent in TM3, TM7, TM10, and TM11 (Figure 5.4). This result is consistent with our recent docking study showing interactions with the same TMDs (Mora Lagares *et al.*, 2020). Some regions that interact exclusively with the active compounds were identified during the molecular dynamics simulations, namely the residues near the breaking loops in TM4 (Ala229, Trp232, Leu236) and TM10 (Met876, Leu879). These could be associated with the transport mechanism, since the breaking loops in TM4 and TM10 are considered to have an important function in the drug-translocation pathway by acting like flexible hinges that open or close a gate region. Interaction with these residues could favour to keep the gate closed and prevent the ligands from re-entering the intracellular space. Also, residues in the second third of TM1 (Leu65, Met68), TM11 (Gln946, Met949, Tyr950, Tyr953) and TM12 (Ala987), and residues in the last third of TM5 (Tyr307), TM3 (Gln195), TM6 (Ser344) and TM7 (Phe728, Phe732) interact uniquely with the active compounds.

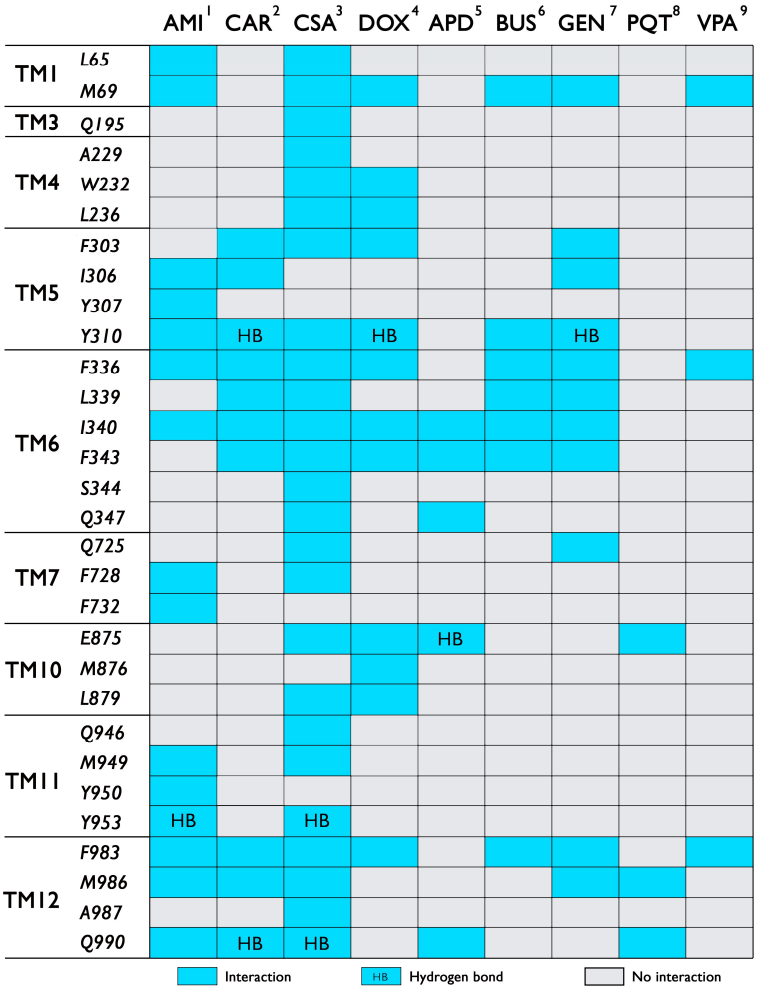


Figure 5.4: Ligand–P-gp interactions. Residues involved in non-bonded and hydrogen bond contacts. [1] Amiodarone; [2] carvedilol; [3] cyclosporine A; [4] doxorubicin; [5] pamidronate; [6] busulfan; [7] gentamicin; [8] paraquat; [9] valproic acid.

A significant difference was found in the number of interactions within the drug-binding site between active and non-active compounds (Table 5.1). The P-gp ligands formed a greater number of interactions within the binding pocket, along with at least one hydrogen bond contact, whereas the non-active compounds had fewer interactions and almost no presence of hydrogen bonds (see Appendix C, Figure C.1). The number of non-bonded interactions for the active compounds ranged from 13 to 27 and occurred mainly with aromatic and hydrophobic residues. Hydrogen bonding was also more frequent in the active compounds and 62,5% of all hydrogen bonds registered within the binding pocket occurred with tyrosine residues (Tyr310 and Tyr953). Hydrophobic and hydrogen bonding interactions are crucial in ligand–P-gp binding, as a large number of these interactions correlate with a high affinity for the protein (Sean Ekins *et al.*, 2002; Michael Wiese *et al.*, 2001).

It is worth mentioning that from the obtained results the size of the ligand could be related to the number of formed interactions, since the smaller molecules in the group of compounds studied have a lower number of interactions with the receptor. However, the formation of hydrogen bonds seems to be independent of the molecule size, e.g. AMI has only 4 hydrogen bond acceptors (HBA) and APD has 6 hydrogen bond donors (HBD) and 8 HBA, and both form only one hydrogen bond with P-gp.

Table 5.1: Number and type of P-gp residues involved in non-bonded and hydrogen bond contacts.

System	**Simulation contacts**		**Type of residue**		
	Non-bonded	Hydrogen bond	Aromatic	Aliphatic	Polar
AMI[1]	16	1	8	8	5
CAR[2]	11	2	6	5	2
CSA[3]	25	2	8	17	9
DOX[4]	12	1	6	6	2
APD[5]	5	1	1	4	3
BUS[6]	7	0	4	2	1
GEN[7]	11	1	5	6	2
PQT[8]	3	0	0	3	2
VPA[9]	3	0	1	1	0

[1] Amiodarone; [2] carvedilol; [3] cyclosporine A; [4] doxorubicin; [5] pamidronate; [6] busulfan; [7] gentamicin; [8] paraquat; [9] valproic acid.

From the analysis of the interactions, it was also detected that some residues seem to be essential for ligand binding. Simultaneous interactions with Tyr310, Phe336, Ile340, Phe983 were found in all the active compounds, suggesting that these residues may be crucial for the ligand–P-gp interaction. Furthermore, several other residues were identified to interact simultaneously with at least three active compounds, including Gln990 and Met986, interacting simultaneously in the P-gp–AMI, P-gp–CAR and P-gp–CSA systems, and Phe343, Phe303 and Trp232, interacting simultaneously in the P-gp–CAR, P-gp–CSA and P-gp–DOX systems (Figure 5.5). The above data suggest that aromatic and/or hydrophobic contacts may be the key feature that determines the binding affinity of substrates and inhibitors within the binding pocket. The results obtained seem to support the hypothesis that the main differences between active and non-active compounds lie in the number of interactions established within the binding

pocket, this being the major driving force that ultimately controls the efflux of the molecule.

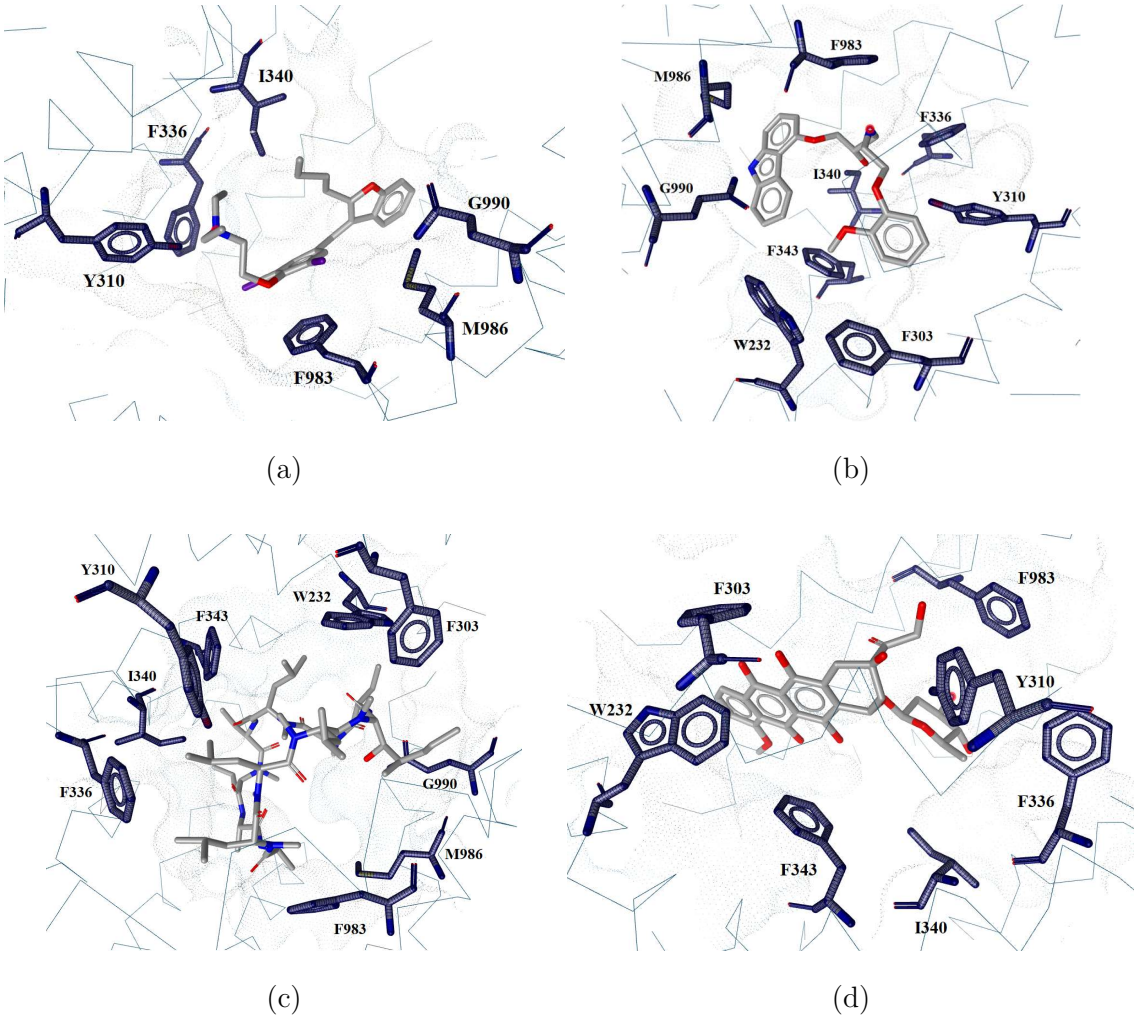


Figure 5.5: Simultaneous interactions detected in the active compounds. (a) amiodarone; (b) carvedilol; (c) cyclosporine A; (d) doxorubicin.

Looking at the available experimental data on residues involved in the binding of co-crystallized ligands, a high level of agreement is observed: 74,2% of the identified interacting residues correspond to residues experimentally found to be involved in substrate or inhibitor binding to P-gp (Alam *et al.*, 2019; Aller *et al.*, 2009; Nicklisch *et al.*, 2016). This demonstrates the consistency between the interacting residues predicted by the MD simulations and the available co-crystallized/cryoEM data.

Figures 5.6 and 5.7 show the interaction regions within the binding pocket for each studied system. The centroids of each cluster, obtained from cluster analysis of the production trajectories, were used to determine the position of the compounds in the binding pocket. When these positions were considered, it was found that within the group of molecules studied, all active compounds shared a common interaction region with a large overlap in the molecular surface (Figure 5.6.). The interactions were established in a central region of the binding pocket in all clusters, indicating that these molecules remained stable within the binding pocket during the simulation time.

Although the interacting residues are not the same in all the active-bound systems, the nature of the interactions is conserved.

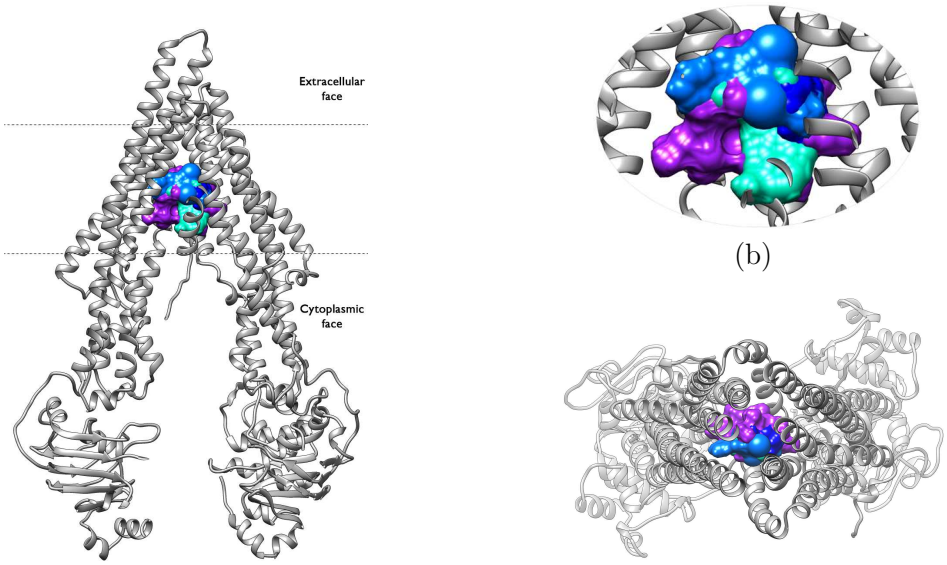


Figure 5.6: Distribution of the active compounds within the binding pocket; (a) frontal view; (b) zoomed view; (c) view from the extracellular side of the protein looking into the inner chamber. The ligands are shown in surface representation; AMI is shown in lighter blue, CAR in blue, CSA in purple and DOX in turquoise.

In contrast, the non-active compounds were all found to form interactions at different regions of the binding pocket (Figure 5.7), e.g. the most populated cluster of BUS shows interactions at a lower right site of the binding cavity, while the second most populated cluster displays interactions at a completely different site (upper left site), suggesting less stability in the binding mode. A similar behaviour was also observed for PQT, which shows interactions in a lower right site of the binding pocket in the most populated cluster, whereas in the second most populated cluster the interactions occur in a lower, more central region (see Appendix C, Figure C.2). For VPA and APD, on the other hand, the region of interaction was constant throughout the entire simulation run, however, different from the interacting region of the active compounds. The greater conformational variability of the inactive compounds and their ability to bind to different regions than the active ones could be related to their physicochemical properties such as molecular weight or volume.

The main contacts of GEN are registered in a central region of the binding pocket, exhibiting a behaviour similar to the active compounds. Experimentally, GEN has not been shown to be transported by or to inhibit the P-gp, so this small discrepancy between the results of the *in silico* experiment and the available *in vitro* data could be attributed to the physicochemical properties of the compound, which may affect its ability to reach the binding site under physiological conditions or in the *in vitro* assays. It should be noted that GEN is the inactive compound with the highest molecular weight in the group (Appendix C, Table C.5) which may be correlated with the region of the binding pocket in which it establishes interactions. Furthermore, the number of contacts formed by GEN and P-gp is similar to that observed for some active compounds (Table 5.1). However, it does not interact with any of the amino acids that are exclusive to active compounds, nor does it exhibit the simultaneous interactions observed for the active molecules. This compound is an example of how it is not enough to orient

"correctly" in the active site, but that specific interactions must also be present in order to be transported by the P-gp.

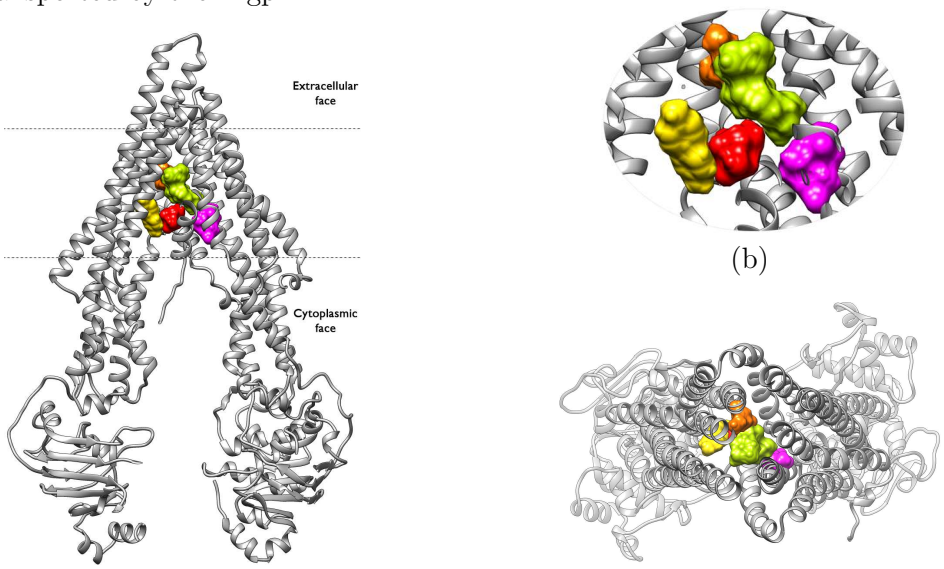


Figure 5.7: Distribution of the non-active compounds within the binding pocket; (a) frontal view; (b) zoomed view; (c) view from the extracellular side of the protein looking into the inner chamber. The ligands are shown in surface representation; APD is shown in red, BUS in magenta, GEN in light green, PQT in yellow and VPA in orange.

When studying the P-gp ligand activity, the physicochemical properties of the compounds under study are another factor that should be taken into account (see Appendix C, Table C.5), since they play an important role in determining the bioavailability under physiological conditions. In the group of studied compounds, there is a clear difference in the logP values between ligands and non-ligands of P-gp. Most of the non-active compounds in the group have negative logP values, meaning that they are more hydrophilic, a physicochemical property that may prevent them from reaching the binding pocket. Previous studies have highlighted the importance of physicochemical properties such as logP value for ligand–protein interactions. The high permeation rate of some compounds within the lipid bilayer is fundamental to the modulation of P-gp due to a competitive mechanism for the drug-binding site between inhibitors and substrates (Akiyama *et al.*, 1988), but is not limited to modulation, as many of the substrates also have high logP values that favour the permeation within the membrane.

5.2.4.3 Binding free energy calculations

Although the analysis of ligand–protein contacts provides good insight into the binding affinity of a potential ligand, the free energy of binding can help to provide a better overview and accurate ranking within a pool of potential drug candidates. Figure 5.8 shows the binding free energy estimate for each compound using MM/PBSA calculations over 100 frames sampled from the entire 500ns trajectory. The resulting low energy values indicate strong binding as well as stability of the simulated complexes. A complete description of the energy components can be found in Appendix C, Table C.9.

Across the simulations, all active compounds exhibited lower and consequently more favourable free energies of binding than the non-active ones, with CSA being the

compound with the lowest estimated energy in the group (-55.09 kcal/mol). The binding stability of CSA, reflected by the large number of hydrophobic interactions and the simultaneous presence of hydrogen bonds within the binding pocket, is further confirmed by the value of the estimated binding energy. This result is consistent with the available literature, since CSA is known to be a substrate with high affinity for P-gp (Litman *et al.*, 1997; Saeki *et al.*, 1993), and with the experimental transport assay results reported in (Mora Lagares *et al.*, 2020).

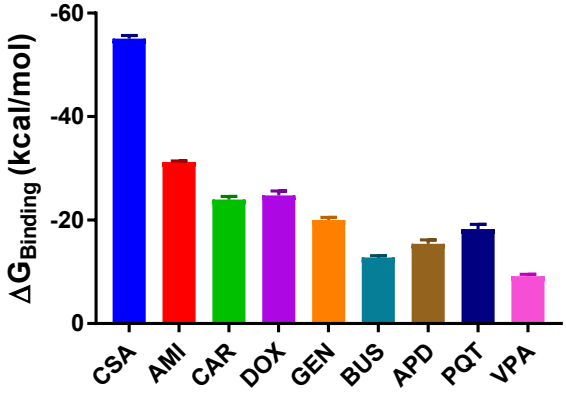


Figure 5.8: Bar graph of the estimated free energy of binding for the ligand–P-gp complexes studied using MM/PBSA calculations over 100 frames sampled from the entire 500ns trajectory. (n = 100, mean ± SEM). Error bars: SEM.

On the other hand, the highest and less favourable estimated free energy of binding (-9.15 kcal/mol) corresponds to the small molecule VPA. The binding energy estimate confirms the weak binding of the compound and the poor stability of the complex, which is also reflected by the low number of established interactions within the binding pocket. The result is consistent with the experimental transport assay results reported in (Mora Lagares *et al.*, 2020), where the compound was found not to be transported by P-gp. The hypotheses generated from the analysis of the ligand–P-gp interactions are further confirmed by the estimated free energies of binding of the compounds simulated in this study.

5.2.4.4 Structural analysis

To analyse whether significant changes in the protein structure could be detected during the simulation run and associated to any particular pattern, the centroids of the most populated cluster in each system, obtained from cluster analysis of the production trajectories were used. For this purpose, the Cα-distance between each centroid and a reference structure (cryoEM structure PDB ID: 6QEX) was measured with the aim of detecting local conformational changes in each system. As can be seen in Figure 5.9, the largest deviations in the structure of all systems occur in the region of the NBDs and to a lesser extent in the region of TM4 and TM5, and in the loop region between TM8 and TM9. Since the major deviations occur in the region of the NBDs, the distance between NBD1 and NBD2 was monitored during the simulation time. The separation of NBDs was measured as the distance between the N atom in the Lys residue of the Walker A motif in NBD1 and the Cα of the Ser residue in the signature motif of NBD2, conserved motifs important for ATP binding and hydrolysis (Domicevica *et al.*, 2015). As a result, significant changes in the NBDs distance were observed throughout the simulations,

indicating a large relative motion between NBDs in each simulated system (see Appendix C, Figure C.3).

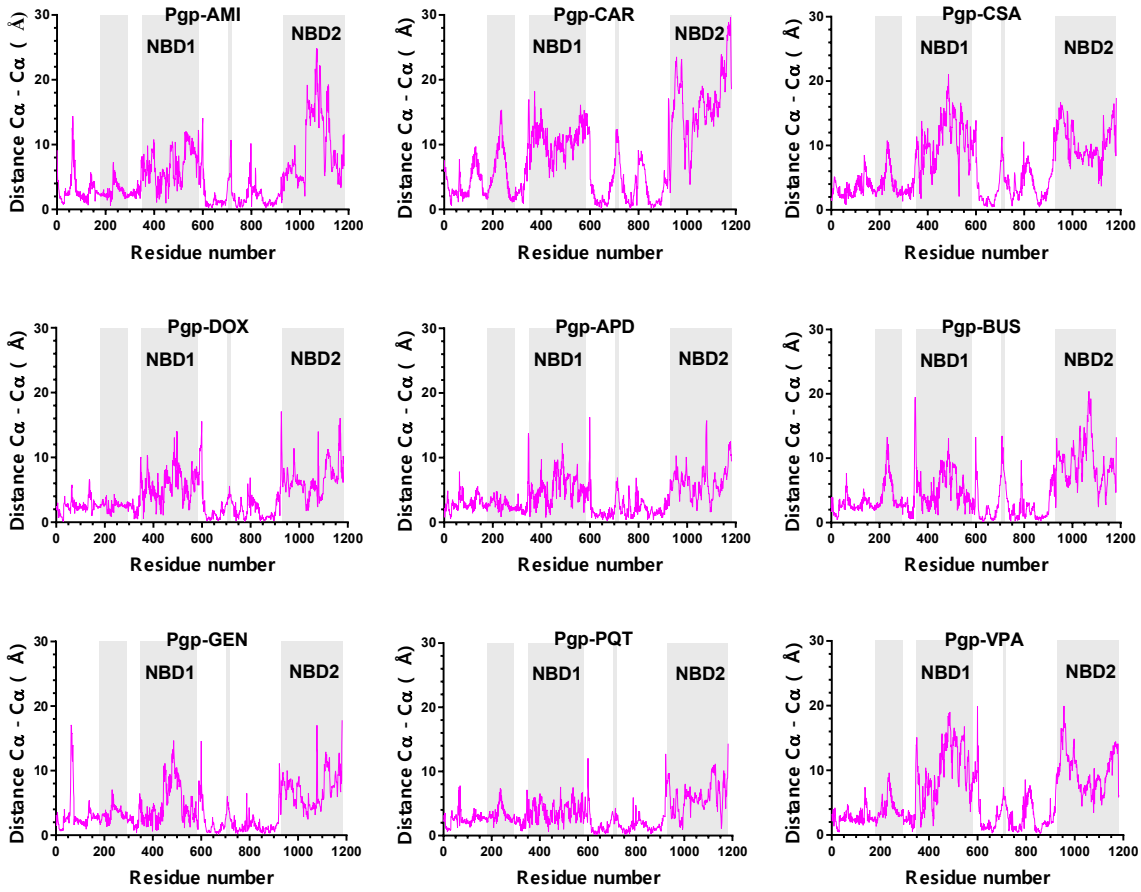


Figure 5.9: Cα-distance between the centroids of the most populated cluster in each system and the cryoEM structure of the *h*P-gp (PDB ID: 6QEX). The residue numbers do not correspond to the numbers in the PDB file of the cryo-EM structure, but are consecutive as required by the simulation software.

The shape of the distance distribution curves (Figure 5.10) reveals a fundamental conformational difference between active-bound and non-active-bound states, which display constant fluctuations between NBDs. The broad distance distributions in the active-bound complexes indicate that these systems are highly dynamic despite the absence of nucleotides. In contrast, the non-active bound systems show narrower distance distribution curves demonstrating more limited conformational flexibility. These findings suggest that the protein is in a less active state and tends to stabilise in a single conformation during the production runs of the non-active compounds. The observed fluctuations are consistent with the predominant movement of the protein; a movement of distancing and approaching between NBDs. It is noteworthy that although DOX is an active compound, the NBDs distance distribution is quite similar to that of BUS, an inactive compound, suggesting that there may be a difference in the transport mechanism of DOX compared to the other active compounds. Another exception in the distribution curves is given by VPA, whose NBDs distance distribution is more similar to the active compounds in the group.

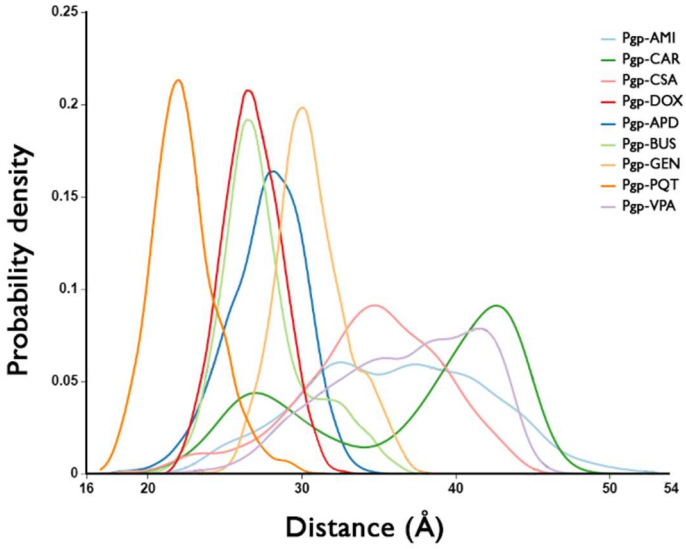
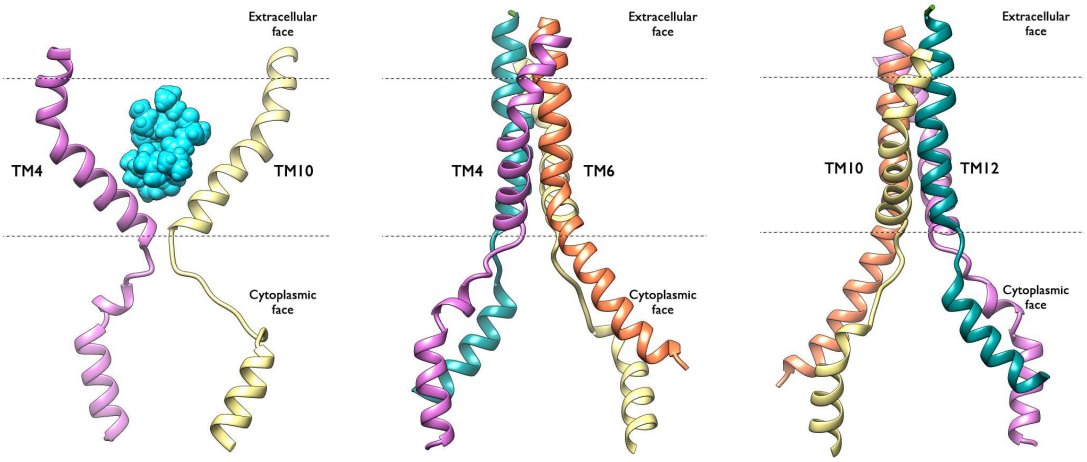


Figure 5.10: NBDs distance distribution curves over all trajectories for the studied systems.

Previous studies have found that TM4 and TM10 shape a gate region open to the cytoplasm. Interestingly, in all the simulated systems, TM4 and TM10 segments adopt a kinked conformation that closes the cytoplasmic gate to the drug-binding site forming an occluded cavity. During the simulation run, all the studied compounds were found in the centre of this cavity (Figure 5.11.a). Moreover, the two portals formed by TM4/TM6 and TM10/TM12 (Figure 5.11.b and c), which allow the access of hydrophobic molecules directly from the inner leaflet of the membrane, adopt the same conformation in all the complexes studied; the two portals are open wide enough to accommodate hydrophobic molecules and allow P-gp to scan the inner leaflet to select and bind specific hydrophobic drugs prior to transport.



(a) (b) (c)

Figure 5.11: Ribbon representation of TM segments important for the P-gp activity; (a) TM4 (pink) and TM10 (yellow) adopting a kinked conformation, with CSA located in the centre of the occluded cavity; (b) TM4 (pink) and TM6 (orange) portal; (c) TM10 (yellow) and TM12 (green) portal.

5.2.4.5 Concerted motions in P-glycoprotein

Principal Components Analysis (PCA) was used to better analyse the conformational changes and dominant modes of motion undergone by the P-gp during the simulation run. The approach is based on the diagonalization of the covariance matrix and the projection of the resulting eigenvectors into the structure of the protein. Thus, a set of orthogonal axes of motion (eigenvectors) is obtained, each representing one direction of the atomic fluctuations in the system, while the eigenvalues give the amplitude of the fluctuations along each principal component. The sum of all eigenvalues can be considered as the total conformational variance observed during the simulation.

The covariance matrix of the 1,182 alpha carbons in the protein was calculated and diagonalized to obtain the eigenvectors and eigenvalues describing the motion of the system. The first two principal axes, PC1 and PC2, together explain between 44.49% and 71.55% of the total system motion (Table 5.2). To gain a better insight into the modes of motion and their influence on the dynamics of the system, the trajectories were projected on the PC1 (the PC that captures the most structural variance along the trajectory). A cumulative projection was also performed on PCs explaining at least 85% of the total flexibility of the complexes.

Table 5.2: Summary of the PCA analysis for the 500 ns simulation run of P-gp ligand-bound systems. Only the results for the first 12 eigenvectors are presented

Eigenvector	Cumulative % of motion[a]								
	AMI[1]	**CAR[2]**	**CSA[3]**	**DOX[4]**	**APD[5]**	**BUS[6]**	**GEN[7]**	**PQT[8]**	**VPA[9]**
1	34.02	58.04	33.35	28.39	36.07	37.58	32.61	30.24	40.62
2	54.10	71.55	54.74	44.49	50.55	55.95	50.48	50.83	60.08
3	72.36	84.42	73.97	65.63	68.96	73.99	66.34	69.08	79.22
4	77.40	87.92	79.90	71.06	73.86	77.89	71.59	73.77	84.27
5	81.31	90.28	83.37	74.68	77.73	80.48	75.29	76.59	87.40
6	84.36	91.67	85.85	78.10	80.73	82.94	77.94	78.86	88.90
7	86.74	92.75	87.37	80.78	82.66	85.18	80.06	80.59	90.12
8	88.31	93.48	88.43	82.37	84.11	86.38	81.76	82.18	91.14
9	89.34	94.09	89.39	83.78	85.41	87.52	83.39	83.35	92.01
10	90.14	94.66	90.14	85.05	86.46	88.43	84.60	84.44	92.69
11	90.70	95.13	90.86	86.00	87.32	89.11	85.57	85.42	93.25
12	91.20	95.43	91.49	86.81	88.06	89.76	86.40	86.29	93.73

[1] P-gp–AMI; [2] P-gp–CAR; [3] P-gp–CSA; [4] P-gp–DOX; [5] P-gp–APD; [6] P-gp–BUS; [7] P-gp–GEN; [8] P-gp–PQT; [9] P-gp–VPA; [a] Flexibility explained by eigenvectors from 1 to n.

The results of mapping the motion of the systems along the first principal component (Figure 5.12) show a significant difference between the active- and non-active-bound systems. For the active-bound complexes, most of the conformational changes are associated with NBD1, with the exception of P-gp–CAR, where both NBDs carry most of the flexibility. In contrast, for the non-active bound systems, the conformational changes

are less wide and more evenly distributed between both NBDs. In terms of the motion pattern, the P-gp–VPA system exhibits a completely different behaviour, as only the NBD2 carries most of the flexibility in this system. When the atomic fluctuations along the PCs that explain at least 85% of the conformational changes are analysed, the results are similar to the RMSF results described previously in section 5.2.4.1, where the flexibility is concentrated in both NBDs for all systems, with a major difference in the amplitude of the movement between active- and non-active-bound complexes (see Appendix C, Figure C.4).

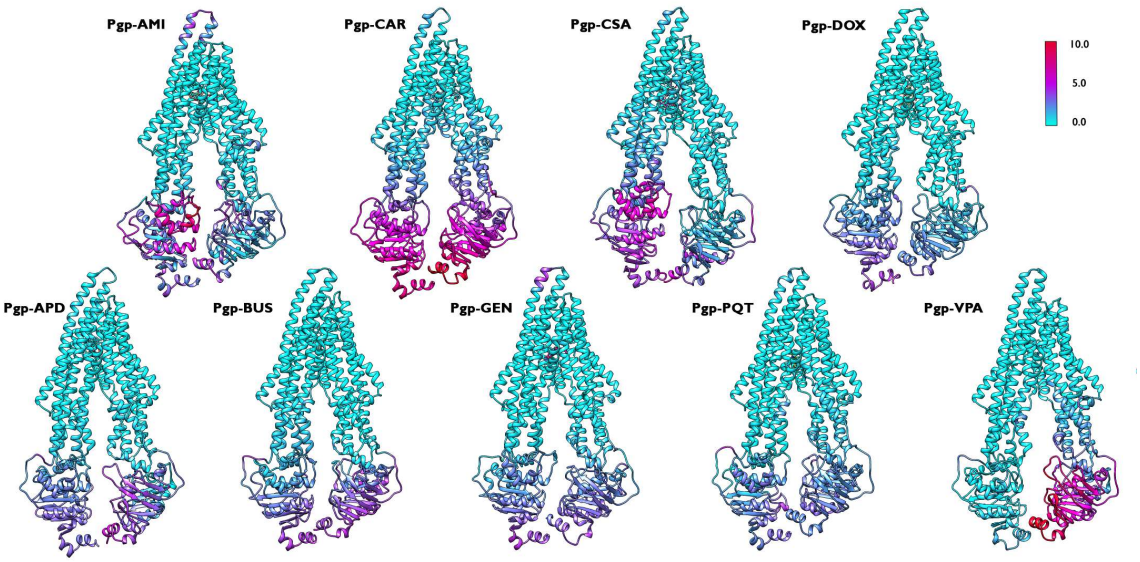
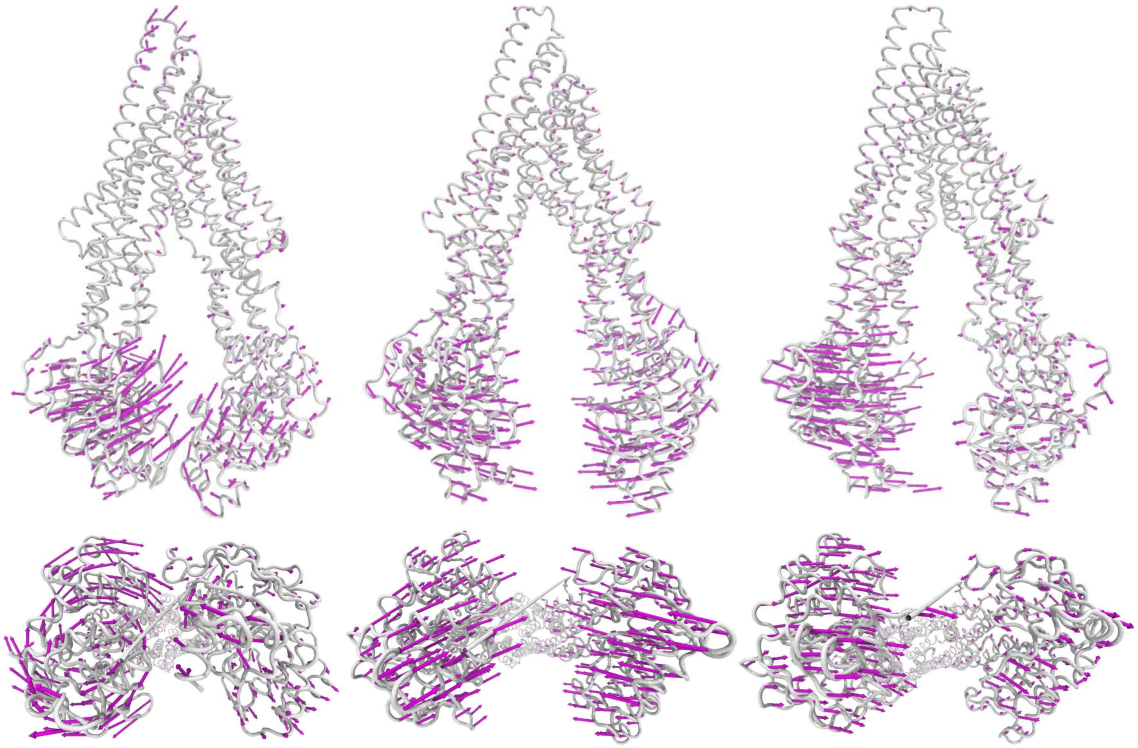


Figure 5.12: Cα-RMSF coloured representation for the simulated P-gp–ligand systems along the first principal component (PC1) calculated from the 500 ns production run. The flexibility scale goes from cyan (lower values) to red (higher values).

It is interesting to note that when analysing the single atomic fluctuation values of the individual residues along the first principal component, the residues forming the ABC motif in NBD1 show higher fluctuations in each active-bound system than in any of the other systems, with a flexibility increase ranging from 6.08% to 13.42% relative to the non-active complexes. This variation is not visible when analysing the PCs explaining at least 85% of the motion of the complexes. The fact that the residues of the ABC motif involved in ATP binding are more flexible in the active-bound systems and that this flexibility is observed only in the ABC motif of one NBD suggests some kind of asymmetry in the motion pattern of the NBDs when the protein is bound to an active compound.

The PCA results were also used to identify and visualize the main motion patterns experienced by these highly flexible regions. The analysis shows that the structural variance along PC1 for the transporter in the presence of an active compound is dominated by the motion between NBD1 and NBD2 getting closer, specifically for the complexes formed with AMI, CAR and CSA. Although the nature of the movement in each system is different (Figure 5.13), the ABC motif in NBD1 and the Walker A motif in NBD2 tend to approach as the NBDs move. This is important to highlight because the binding of the nucleotide molecule required to initiate the conformational changes characteristic of the transport cycle occurs between these two conserved motifs.



(a) (b) (c)

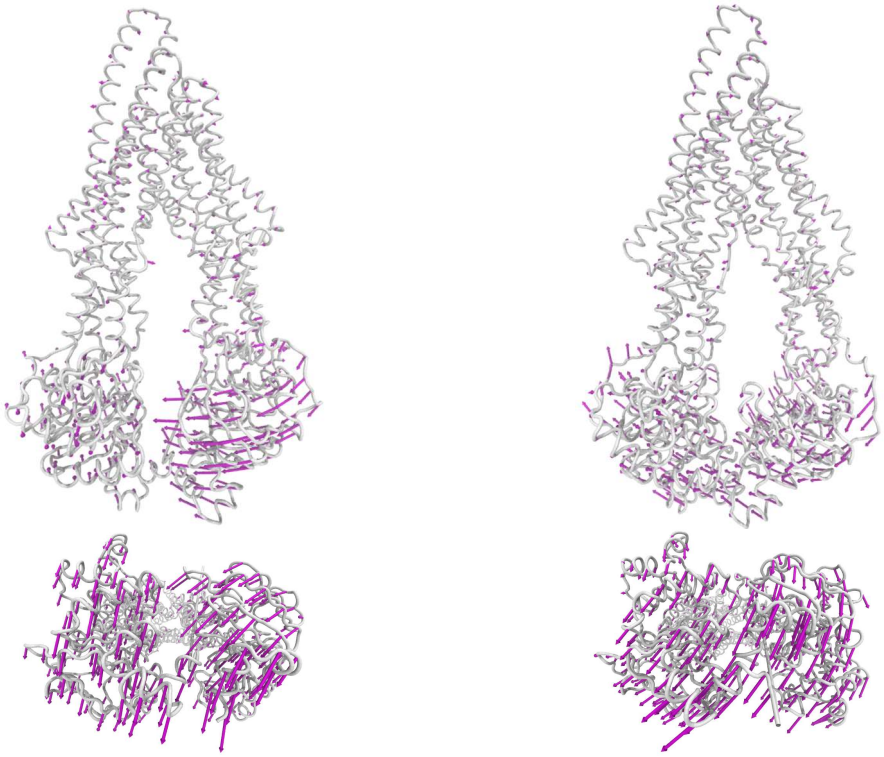
Figure 5.13: Front and Cytoplasmic view of P-gp motion patterns along PC1; (a) P-gp–AMI; (b) P-gp–CAR; (c) P-gp–CSA. The direction of the movement is represented by magenta arrows and the size of the arrows is proportional to the magnitude of the movement. For clarity, the reverse direction is not shown.

Looking at Figure 5.13, the main difference in the motion patterns of the individual active-bound complexes is that in P-gp–AMI and P-gp–CSA complexes the nature of the motion is asymmetric, with the largest amplitude carried by NBD1, whereas in the complex formed with CAR, the nature of the motion is symmetric, with a large amplitude of movement distributed equally between both NBDs. Moreover, the motion of the NBDs is coupled with the rotation of TM4, TM6, TM10, and TM12, which slightly open the binding pocket when the NBDs come together, and with the movement of the extracellular domain going from residues Asn81 to Phe104, whose direction of movement is opposite to that of NBD1. These observations suggest that these three active-bound complexes may have a common transport mechanism in which the predominant motion of the transporter is related to the approach of the NBDs and to a large amplitude of movement, indicating a very flexible and active system.

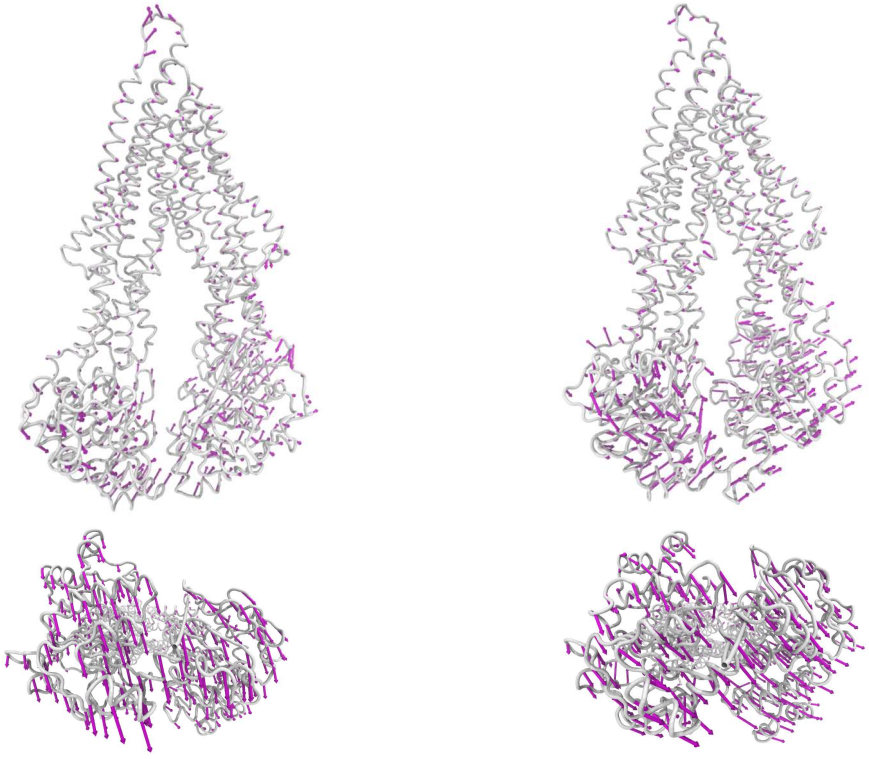
On the other hand, the structural variance of the transporter along PC1 for the group of non-active-bound complexes was also determined by the motion of the NBDs, but in a symmetric and concerted manner. The movement of the NBDs in these systems is characterised by a pendular motion perpendicular to the bilayer plane and narrower amplitudes, (Figure 5.14) with no tendency to approach to each other, indicating a less flexible and active system in which the conformational changes associated with the transport cycle do not seem to be initiated.

Interestingly, in the analysis of the structural variance along PC1, the P-gp–DOX- and P-gp–VPA systems do not follow the trend of the other complexes in the group to which they belong (Figure 5.15). In the case of the P-gp–DOX system, the direction and

amplitude of the NBDs movement more closely resemble the pattern followed by the non-active compounds, suggesting that the transport mechanism of this particular P-gp ligand may differ from that of the other active compounds in the group. It might undergo a slower transport process and therefore, longer simulation times would be required to observe the same concerted movements as in the other systems. In the P-gp–VPA system, on the other hand, although the NBDs did not tend to approach each other as in the rest of the non-active-bound systems, the structural variance occurred almost exclusively in NBD2, which followed a pendular motion perpendicular to the bilayer plane characterised by a very large amplitude. These results indicate a significant difference in the conformational distribution and dynamics of NBDs when active or non-active compounds are bound to P-gp.

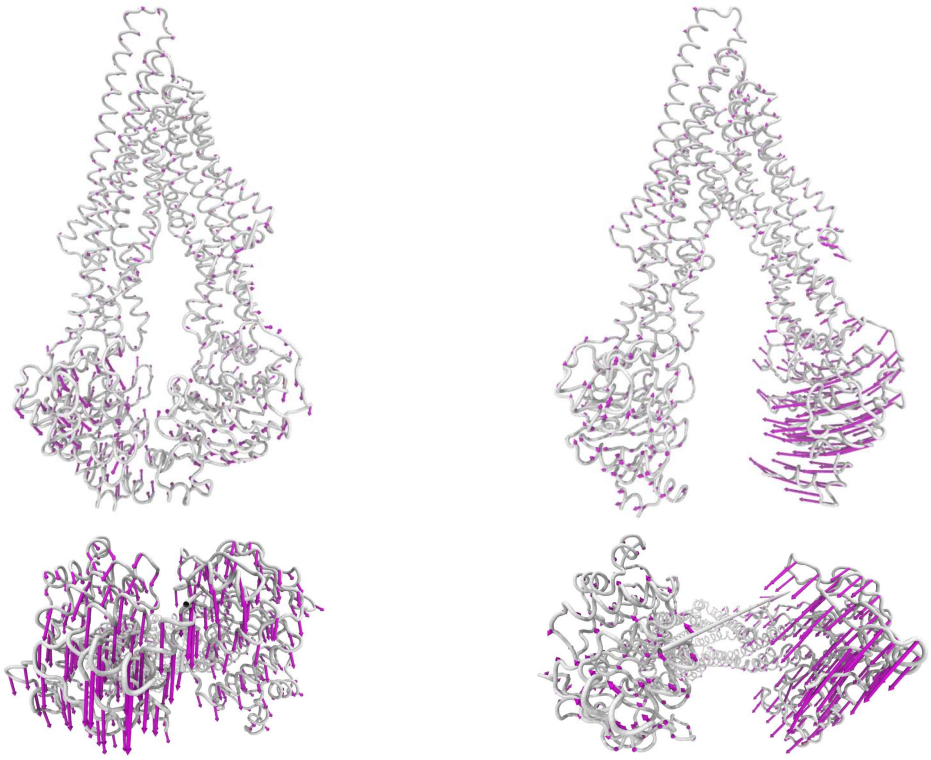


(a) (b)



(c) (d)

Figure 5.14: Front and Cytoplasmic view of P-gp motion patterns along PC1; (a) P-gp–APD; (b) P-gp–BUS; (c) P-gp–GEN; (d) P-gp–PQT. The direction of the movement is represented by magenta arrows and the size of the arrows is proportional to the magnitude of the movement. For clarity, the reverse direction is not shown.



(a) (b)

Figure 5.15: Front and Cytoplasmic view of P-gp motion patterns along PC1; (a) P-gp–DOX; (b) P-gp–VPA. The direction of the movement is represented by magenta arrows and the size of the arrows is proportional to the magnitude of the movement. For clarity, the reverse direction is not shown.

5.2.4.6 Binding Pocket

It is important to evaluate how conformational changes in the TM helices might affect the volume of the internal cavity. To this end, the volume of the binding pocket for each system was calculated during the MD simulations using POVME 3.0. The results show a consistent decrease in the volume of the binding pocket for all the studied systems compared to the starting point of the simulation, except for the P-gp–AMI system whose volume remained constant with a slight increase at the end of the simulation (Figure 5.16). The changes in the computed binding pocket volumes during the simulations can be correlated with the different motion patterns of P-gp, since the presence of other movements associated with less wide eigenvectors could be directly related to the translocation mechanism. Analysis of the first 10 eigenvectors of each system revealed that eigenvectors 4, 6, 8, 9, and 10 contribute to motions associated with variations in the volume of the internal cavity. The motion pattern associated with these eigenvectors consists of a change in the distance between the NBDs and a helix rotation about the major protein axis of TM helices 4, 6, 10, and 12, resulting in a motion pattern of contraction and expansion of the cavity. The decrease in the distance between NBDs is coordinated with the rotation of the TM helices, yielding to a small degree of expansion of the binding cavity. The correlation of the NBDs dimerization with TMD conformational changes is important for understanding the ligand affinity changes of the binding site.

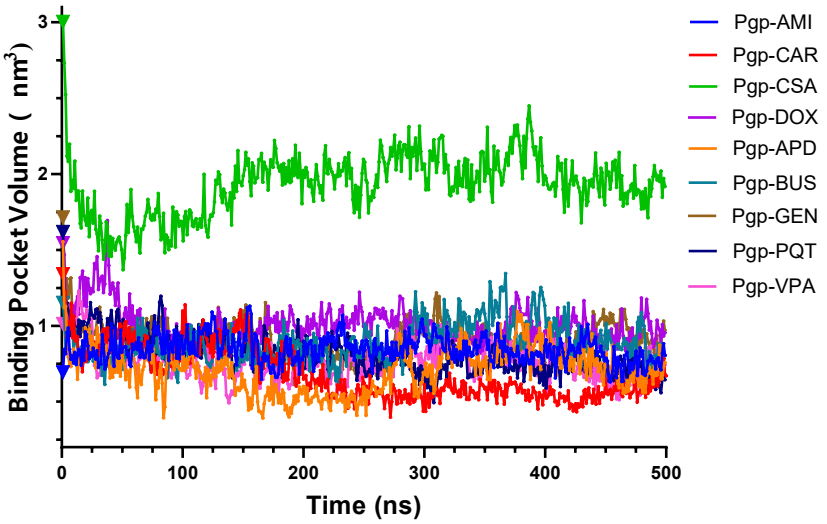
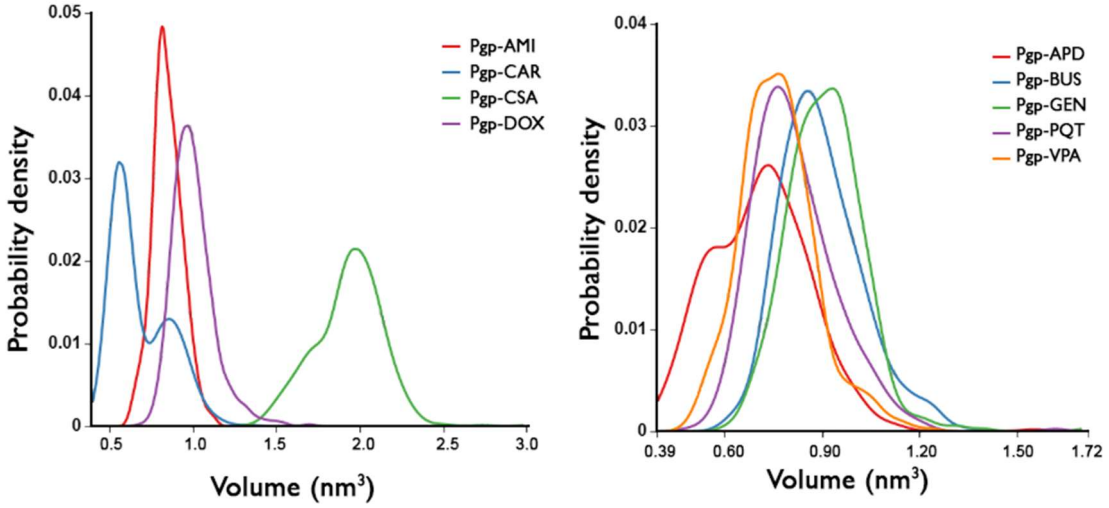


Figure 5.16: Variation of the internal cavity volume for each studied system as a function of time from the MD simulations.

Figure 5.17 shows how the binding pocket in all the studied systems exhibits a broad distribution of volumes along the simulation, with no significant differences or patterns found in the volume distribution between active- and non-active-bound systems. This

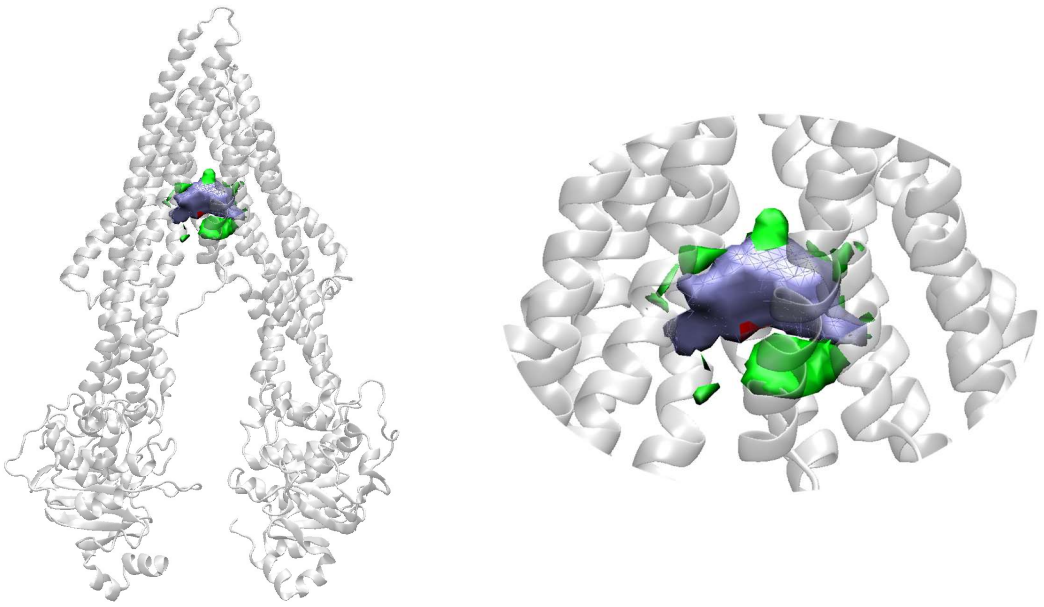
behaviour suggests some conformational flexibility in the TM domains independent of the bound ligand. As expected, the largest volumes of binding pocket were found in the system formed by P-gp and the large CSA molecule, demonstrating the ability of the binding pocket to adapt to the size of the ligand. A secondary peak was also observed when simulating P-gp–CAR and P-gp–APD systems; this secondary peak reflects a secondary population of P-gp conformations in which the cavity closes more for the APD-bound system but opens for the CAR-bound system.



(a) (b)

Figure 5.17: Distributions of the internal cavity volumes for; (a) active-bound systems; (b) non-active-bound systems.

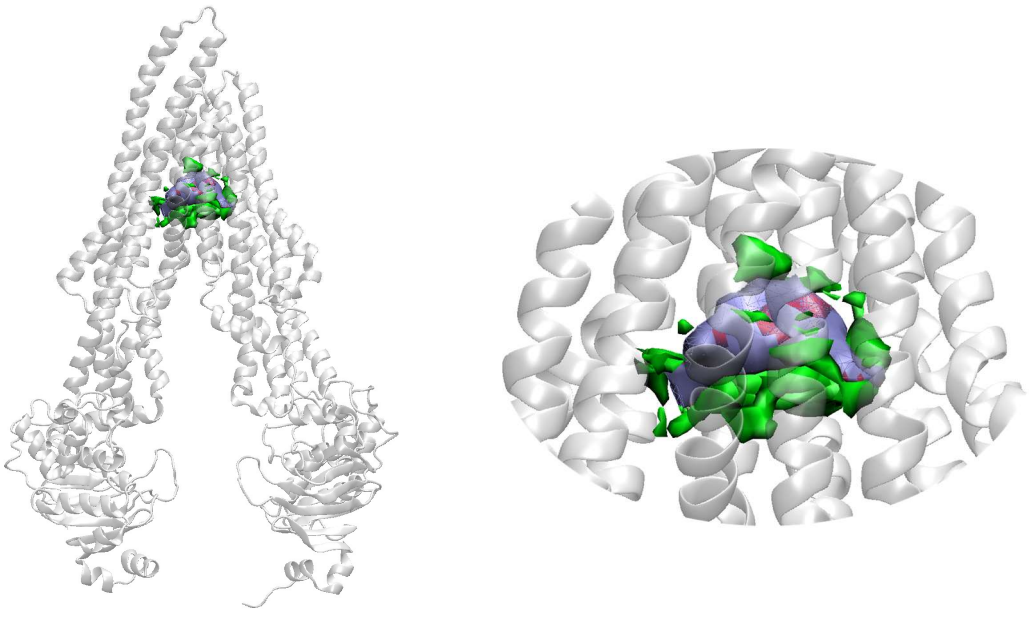
Figures 5.18 and 5.19 show the results from the clustering analysis of the binding pocket volumes. The clustering analysis was performed for the combined trajectories of the active-bound and non-active-bound systems, separately. Eight clusters were generated from the analysis of the combined trajectories of all the active-bound systems and twelve clusters were obtained from the non-active-bound systems. Each analysed frame was assigned to a single cluster and the resulting clusters represent frequently visited pocket shapes. From the clustering results of the active-bound systems, it was found that each of the first four most populated clusters contains members (snapshots) of only a single active-bound system, suggesting that the binding pocket shapes of these complexes differ from each other. Figure 5.18 shows the average pocket shape for the analysed frames, as well as the areas where each cluster opens or closes more than this average.



(a) (b)

Figure 5.18: Average pocket shape of the four most-populated cluster centroids for the active-bound systems shown as a blue surface; (a) front view; (b) zoomed view. Regions more open or closed than the average in each cluster are shown as green and red surfaces, respectively.

In the clustering results of the non-active-bound systems, it was instead observed that the most populated cluster contains members of multiple non-active-bound systems, indicating that the binding pocket shapes of these complexes are very similar. However, when looking at Figure 5.19, it can be seen that the cluster centroids are much more open than the average shape compared to the results of the active-bound systems, confirming the higher fluctuations of the non-active compounds within the internal cavity.



(a) (b)

Figure 5.19: Average pocket shape of the five most-populated cluster centroids for the non-active-bound systems shown as a blue surface; (a) front view; (b) zoomed view. Regions more open or closed than the average in each cluster are shown as green and red surfaces, respectively.

Interestingly, the non-active compounds, although small molecules, do not induce conformational changes that reduce the volume of the binding pocket. This lack of induced-fit capability probably due to the weak interactions with the receptor could be related to the inability of the protein to transport these compounds.

5.2.4.7 Exposure of surfaces to solvent

Changes in the solvent accessibility of the protein can be determined by calculating the solvent accessible surface area (SASA). The total surface area SASA of the different P-gp systems was determined using the CPPTRAJ module (Roe *et al.*, 2013) in AmberTools (D.A. Case *et al.*, 2018) and the per residue SASA using the Python library MDtraj (McGibbon *et al.*, 2015). The computed SASA of the whole protein showed a slight decreasing trend for almost all systems under study, only P-gp–AMI and P-gp–CSA systems exhibit a slight increase compared to the initial phases of the simulation. From the graph of the change in total SASA with time shown in Figure 5.20, it can also be seen that the P-gp–CSA system has higher SASA values compared to the other systems in the group. In general, the values of the total SASA fluctuate around a constant value along the entire simulation.

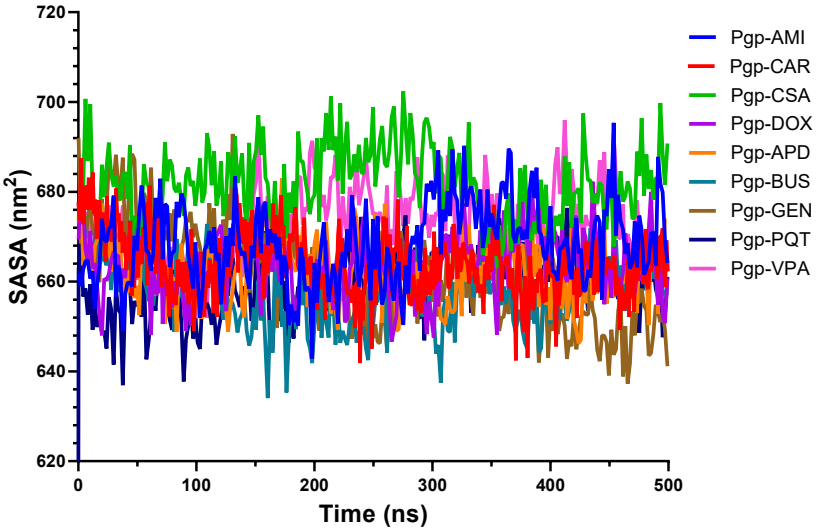


Figure 5.20: Total solvent accessible surface area (SASA) of the studied systems as a function of time from the MD simulations.

The probability density curves (Figure 5.21) give a clearer representation of the accessibility of the systems to the solvent. The non-active-bound systems have smaller peak values compared to the active-bound group, suggesting that in these systems the hydrophobic core is more protected from the external environment and that they may be shrinking, resulting in more compact and less flexible systems, with the exception of P-gp–VPA. The loss of flexibility of the non-active-bound systems observed in the RMSF

plots is further supported by a decrease in their SASA values. This reduced flexibility can have a major impact on the structural conformation of the protein and consequently affect its functional behaviour.

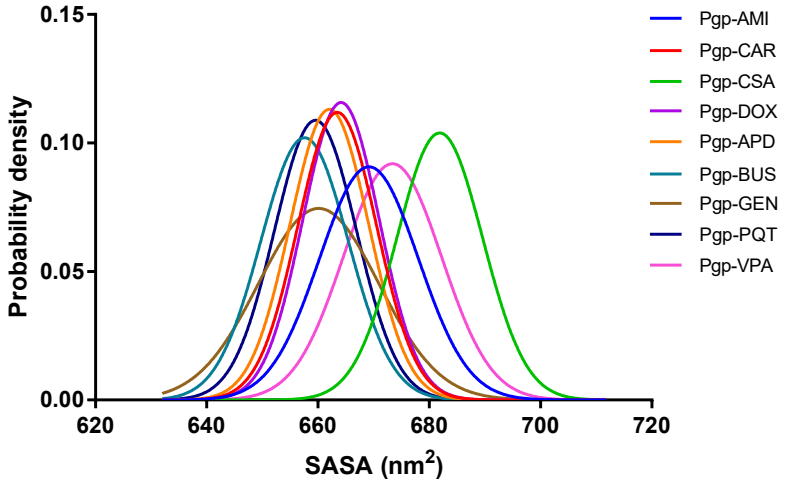
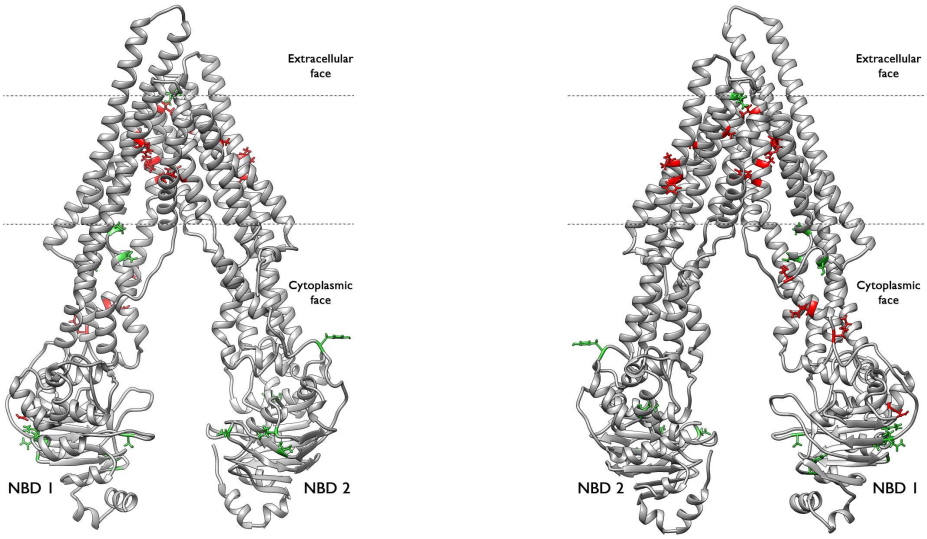


Figure 5.21: Distributions of the total solvent accessible surface area (SASA) for the studied systems.



(a) (b)

Figure 5.22: P-glycoprotein residues with significant variations in the solvent accessible surface area (SASA); (a) front view; (b) back view. Residues coloured red have smaller SASA in the systems formed by P-gp and an active compound. Residues coloured green have larger SASA in the systems formed by P-gp and an active compound.

Looking for changes in specific regions of the protein during the simulation, the SASA values of individual residues were analysed, revealing an interesting result (Figure 5.22): some residues located in the TM region have smaller average values for the SASA in each active-bound system than in any of the other systems. Interestingly, residues M69, I340, F343, Y953, and M986, which are involved in non-bonded contacts that stabilize the ligand–P-gp complexes, have the largest difference in terms of their maximum value in

the non-active-bound systems (see Appendix C, Table C.7). The results suggest that these residues are more buried and less flexible in the active-bound systems, probably due to the strong hydrophobic interactions that stabilize the ligands. On the other hand, there are some residues in the NBDs that have larger average SASA values in each active-bound system than in any of the other systems (see Appendix C, Table C.8). These residues are more exposed to the solvent, which confirms the higher flexibility of these domains when there is an active ligand in the binding pocket; they are less buried and can better explore the conformational space, which can lead to the motion patterns analysed in section 5.2.4.5.

5.2.5 Conclusions

The simulation results provide a broad overview in the task of understanding the different capabilities and limitations of P-gp in binding to different types of ligands. The study shows that P-gp can behave differently depending on the nature of the ligand placed in the binding cavity. The different behaviour is reflected in the motion patterns and structural flexibility variations that P-gp undergoes depending on the bound molecule. In general, the complexes formed by P-gp and an active compound show higher flexibility of the NBDs compared to the non-active-bound complexes, suggesting a significant difference in the conformational distribution and dynamics of NBDs when active or non-active compounds are bound to P-gp.

Through PCA analysis, it was possible to discriminate different motion patterns between active- and non-active-bound systems. It was found that the main motion pattern of P-gp when bound to an active compound corresponds to the variation of the NBDs distance by an asymmetric movement that tends to approach and separate both regions. This movement could be related to the NBDs dimerization, and it is coordinated with TM helix rotations that influence the volume of the binding cavity. This result is consistent with the hypothesis for the transport mechanism, as a higher movement of the NBDs is expected for the activation of the translocation pathway.

PCA analysis also showed that most conformational changes for the active-bound complexes along the first principal component are associated with NBD1. In contrast, for the non-active bound systems, the conformational changes have lower amplitude and are more evenly distributed between both NBDs. The asymmetry in the motion exhibited for the active-bound complexes was further confirmed by the higher flexibility of the ABC motif residues of NBD1, which may support the hypothesis that binding, and hydrolysis of ATP occur alternately at one site. The activity of the nucleotide binding site even in the absence of the ATP molecule suggests that what activates the ATP-binding domains is the presence of the ligand in the binding pocket, then the binding and hydrolysis of the ATP molecule would provide the energy required to initiate and complete the translocation mechanism.

The P-gp binding pocket forms a hydrophobic environment in which various ligands bind. Analysis of the volume of the binding pocket confirmed that the binding cavity is able to adapt to the size of the molecule and hence to accommodate ligands of different sizes. The volume of the P-gp cavity is directly related to the size of the bound molecule, supporting the induced fit model. However, the size of the binding pocket for the molecules that are not transported conserve higher dimensions than expected, leading to the formulation of the hypothesis that an important requirement for the transport of molecules by P-gp is their ability to induce changes in the binding pocket. Interestingly, all active compounds shared a common interaction region with a large overlap of the molecular surface area. The interactions were established in a central region of the binding pocket, while the non-active compounds fluctuated more within the internal

cavity, indicating less stability and strength of binding. The higher hydrophobicity of the binding pocket in the active-bound systems is reflected in the smaller average SASA values found in residues involved in non-bonded contacts. These residues are more buried and less flexible, probably due to the strong hydrophobic interactions that stabilize the ligand–P-gp complexes.

The binding free energy calculations confirmed that a large number of the interactions correlate with a high affinity for the protein and that the term contributing most to the energy is the hydrophobic term. The estimated binding energy values proved to be very useful in validating the P-gp interaction affinity, for example, when a few complexes showed some dissimilarities with respect to the behaviour of the other systems in the studied group, they could be used to weight the affinity level for the protein. The slight variations in the behaviour observed for the P-gp–VPA and P-gp–DOX complexes in some of the analyses performed could be attributed to different interaction mechanisms specific to these compounds, since the estimated free energies of binding confirmed their respective P-gp affinities.

A high degree of agreement was observed between the predicted and experimentally found interacting residues: 74,2% of the identified interacting residues correspond to residues experimentally found to be involved in substrate or inhibitor binding to P-gp (Alam *et al.*, 2019; Aller *et al.*, 2009; Nicklisch *et al.*, 2016), demonstrating the consistency between the interacting residues predicted by the molecular dynamics simulations and the available co-crystallized/cryoEM data.

Overall, this work provides clear evidence that the changes that initiate the translocation pathway are activated by the presence of a ligand in the binding pocket and that these changes start at only one NBD (NBD1). The results of this work support the hypothesis that aromatic and/or hydrophobic contacts may be the key feature that determines the binding affinity of substrates and inhibitors within the binding pocket, but also demonstrate the importance of the physicochemical properties of the molecule, which would determine the ability of a potential ligand to reach the binding site. However, considering that protein conformational changes can occur on at least the microsecond time scale (Sekhar *et al.*, 2013), it should be clear that our results are limited to the simulated time and to the simulation conditions, e.g., absence of ATP molecule. Therefore, further studies involving the nucleotide molecule with longer simulation times, as well as new experimental studies, are suggested to validate the findings of this work.

Chapter 6

Conclusions

Since the discovery of the P-glycoprotein (P-gp) and its link to the multidrug resistance (MDR) more than thirty years ago (Juliano *et al.*, 1976), this biologically important ATP-binding cassette (ABC) membrane transporter has been a target of extensive research. The interest in developing new strategies and therapies to reverse multidrug resistance mediated by P-gp has led to numerous studies attempting to understand how it works, but despite all the efforts, the exact mechanism of ligand recognition and transport is still unknown; the way so many molecules are recognized by P-gp to enter the binding pocket and how they are transported is still not clear.

As a result of all this research, new generations of P-gp inhibitors have been developed, but none of them have passed the phase III in clinical trials, mainly because of unexpected toxic effects. Therefore, improved *in silico* approaches to understand and predict the ligand-binding interactions of the human P-glycoprotein are needed. To contribute to this issue, we have focused our research on an integrated approach because the complex problem of computational drug design or virtual screening cannot be solved by either ligand-based or structure-based methods alone. Each approach has its own advantages and limitations that make them suitable for specific applications. However, hybrid systems, where two or more techniques are combined to overcome the limitations of each individual approach and create a synergy between them, provide valuable modelling tools where the ligand-based and structure-based techniques are integrated in a sequential or parallel manner.

As a first step, we employed a ligand-based approach and developed a model to predict the activity of P-gp on candidate compounds. This developed multiclass classifier is a useful tool for saving significant time in the drug development pipeline, as it provides a quick initial screening step for selecting predicted molecules that would interact with P-gp as inhibitors, if the goal is to develop potential drugs to overcome MDR, or as substrates if the interest goes more in evaluating the toxicity of a drug candidate or the potential drug-drug interactions. The model showed a good classification performance and was successfully implemented within the online platform VEGAHUB (Benfenati *et al.*, 2013), which is freely available to the public at https://www.vegahub.eu/portfolio-item/vega-qsar/.

The second step of our approach focused on a structure-based method where we analysed the binding modes of P-gp ligands through molecular docking studies and assessed that molecular docking could be used as an additional tool to evaluate and screen potential drug candidates and provide additional information about the mechanism of interaction with the transporter. A deeper look into binding modes and estimated binding energies proved useful, for example, to select the top compound within the battery of pre-selected molecules under investigation; the estimated energies proved

to be a good tool for ranking the compounds. The agreement found between the predicted interactions and the available experimental data indicated that molecular docking is a good structure-based approach that offers a good time-accuracy ratio, requiring less time than more complex methods such as molecular dynamics simulations, while still providing insight into the molecular interactions between ligands and P-gp.

In the final step, we focused on a more detailed and sophisticated structure-based method, such as molecular dynamics simulations. The in-depth study of the molecular interactions and dynamics of P-gp leading to the translocation mechanism, performed using the recently solved 3D structure of *h*P-gp, led to the formulation of the hypothesis that an important requirement for the transport of molecules by P-gp is their ability to induce changes in the binding pocket (induce fit model). The asymmetry in the dynamics of NBDs, as well as the influence of the ligand present in the binding pocket on the motion patterns of the NBDs was demonstrated. The estimated relative free energies of binding and the predicted interactions were in agreement with the available experimental data, proving that molecular dynamics is a sufficiently accurate method if we want to take into account the flexibility of the target protein, which in this case is a fundamental feature for the mechanism of action and the polyspecificity of the membrane transporter.

The whole integrated approach, starting with the ligand-based model followed by the structure-based approaches, can be used in a sequential manner, filtering out the inactive compounds using the classification model; this output then serves as input to the next structure-based step. The prioritized list of compounds would be further validated using very accurate free energy calculation methods, such as MD simulations. Although structure-based methods may not yet be able to fully classify P-gp ligands, the application of such methods is essential for understanding the ligand recognition and transport mechanism, which is fundamental for a successful implementation of the combined approach in identifying and designing new P-gp ligands.

Overall, this work contributes a piece to the puzzle of understanding and predicting the activity of P-gp on potential ligands, and we hope it will inspire future research, including longer simulation times as well as the presence of the ATP molecule. We also hope that in the future a complete structure of *h*P-gp, including the flexible linker region, will be available and in this way allow a better study of the entire transport cycle.

Since one of the biggest unanswered questions of P-gp is related to its mechanisms of polyspecificity, it would also be very helpful to develop similar models for other ABC transporters that would allow us to predict the affinity of compounds for a battery of transporters and thus compare their selectivity. This would be very useful to elucidate the polyspecificity of P-gp.

Appendix A

Chapter 3 Supporting Materials

A.1 Distribution of the 24 Overlapping Negative Compounds in the Response Map of the Non-Active Class

Figure A.1. shows the only absolutely certain non-active compounds. The position of the neurons where these compounds are located is coloured black and highlighted with a pink circle. They are well distributed throughout the area occupied by all non-active compounds, indicating that their structural similarity with the non-active molecules, tested with only one assay, is significant. Structural similarity is the only factor that influences the formation of clusters in the output layer of the Counter-Propagation Artificial Neural Network.

The only region of non-active compounds that does not contain any of the double-tested molecules, is around the neuron at position (13, 5) in the lower left part of the map, so we may consider the predictions associated with this area to be less reliable.

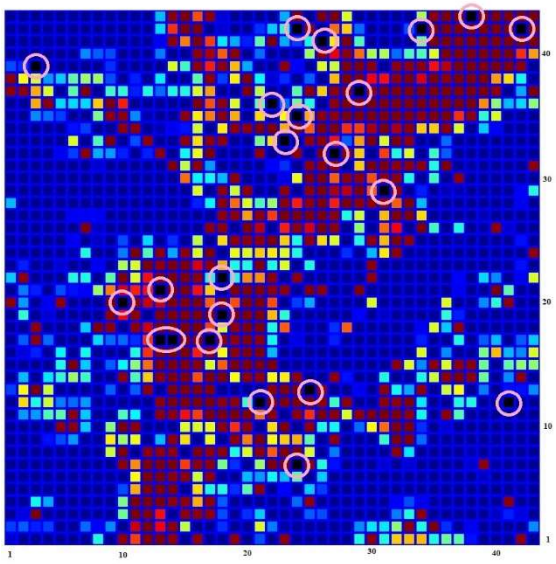
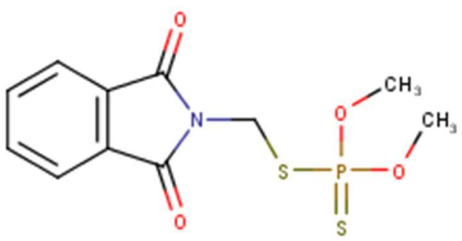
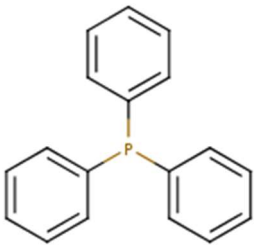


Figure A.1: Distribution of the 24 overlapping negative compounds (P-gp non-inhibitor and non-substrate) in the response map of the non-active class.

A.2 Structures of the Compounds with Largest ED

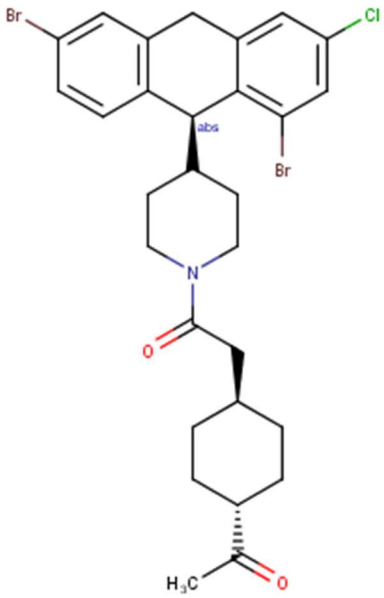


(a)

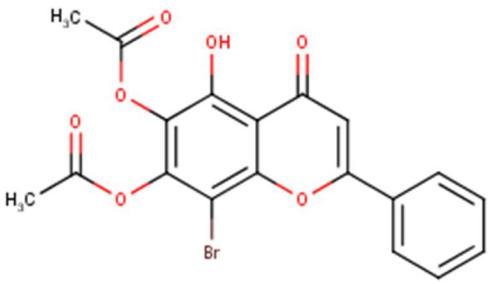


(b)

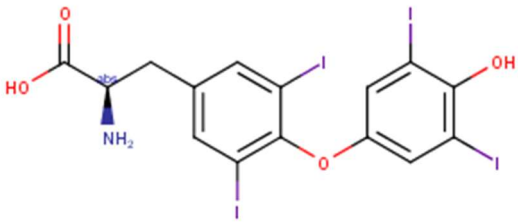
Figure A.2: Chemical structures of the compounds with largest ED to the central neuron in the TE set: (a) Phosmed; (b) Triphenylphosphane.



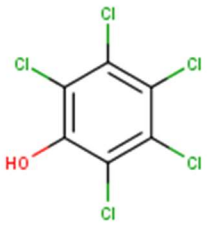
(a)



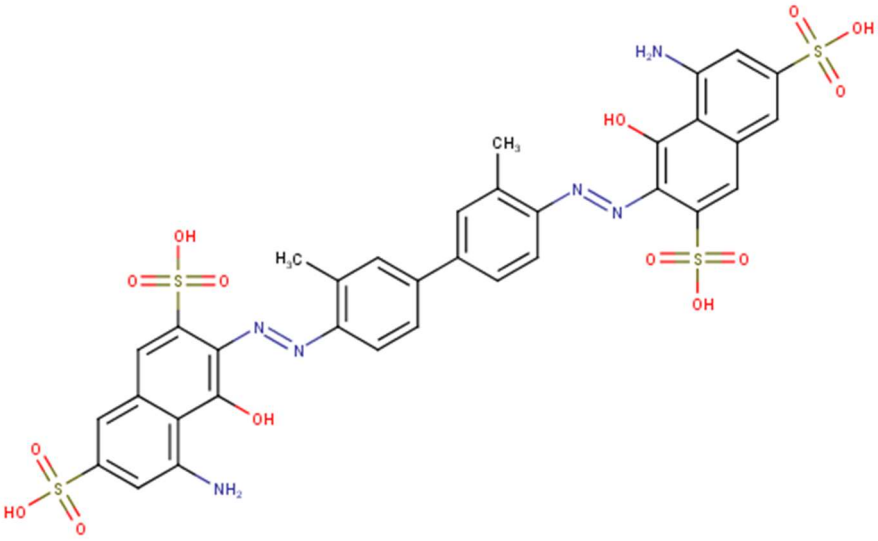
(b)



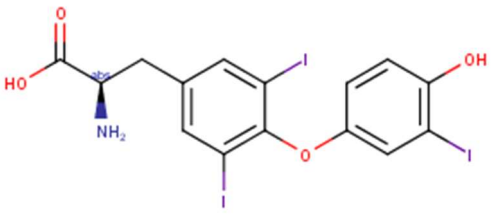
(c)



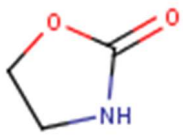
(d)



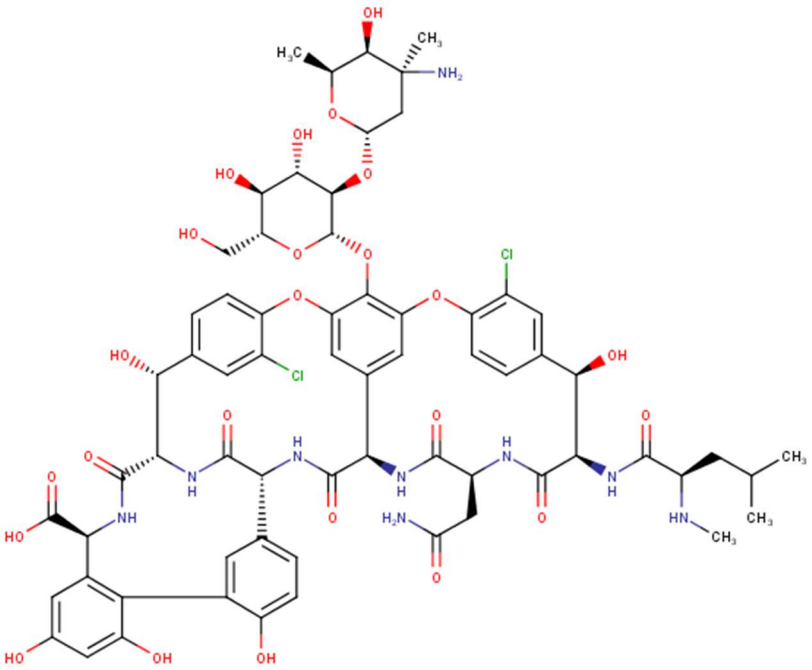
(e)



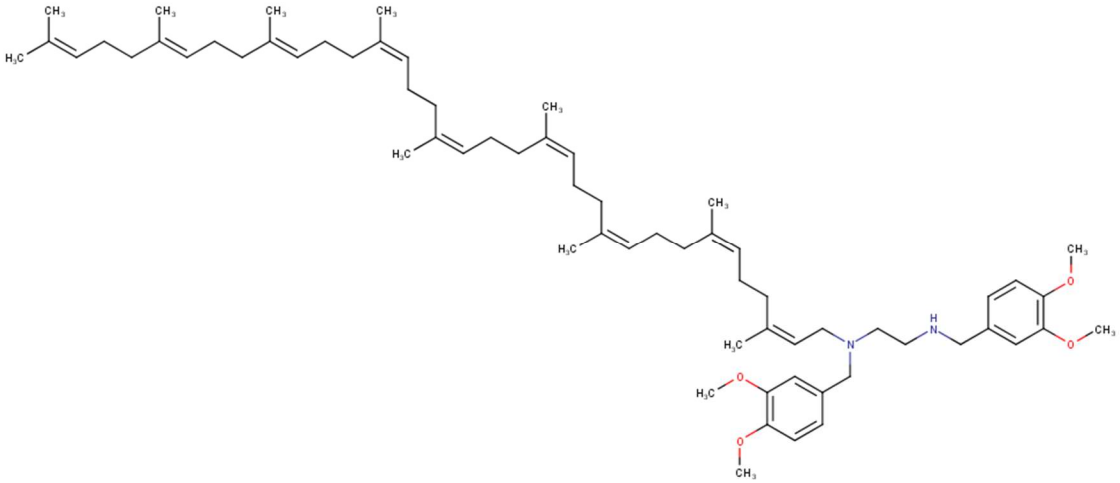
(f)



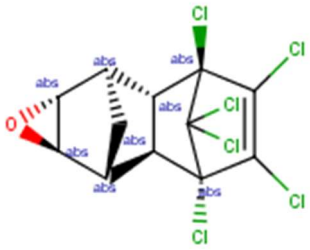
(g)



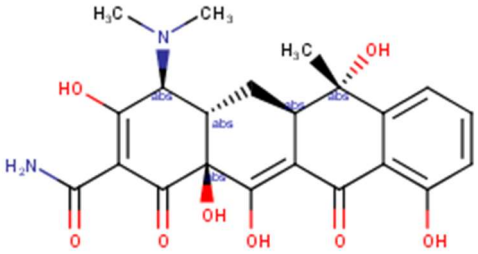
(h)



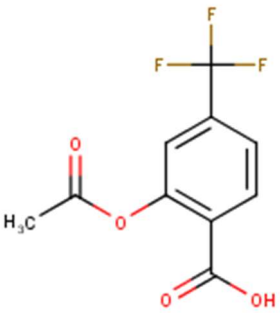
(i)



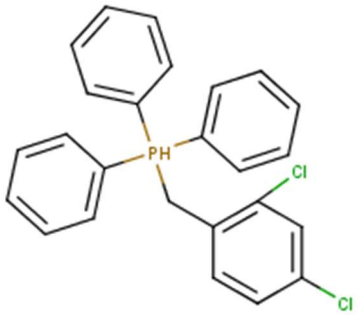
(j)



(k)



(l)



(m)

Figure A.3: Chemical structures of the compounds with largest ED to the central neuron in the V set: (a) 2-(4-acetylcyclohexyl)-1-[4-[(9S)-1,6-dibromo-3-chloro-9,10-dihydroanthracen-9-yl]piperidin-1-yl]ethanone; (b) (7-acetyloxy-8-bromo-5-hydroxy-4-oxo-2-phenylchromen-6-yl) acetate; (c) Thyroxine; (d) Pentachlorophenol; (e) Trypan blue; (f) Triiodothyronine; (g) 2-Oxazolidinone; (h) Vancomycin; (i) Sdb-ethylenediamine; (j) Dieldrin; (k) Tetracycline; (l) Triflusal; (m) (2,4-dichlorophenyl)methyl-triphenylphosphanium.

Appendix B

Chapter 4 Supporting Materials

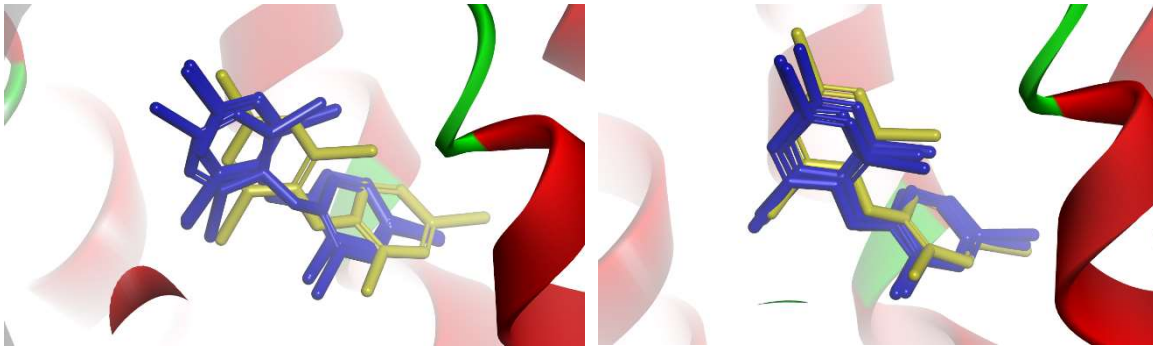
B.1 Results of the Re-Docking Validation Procedure

Table B.1: Root-mean-square deviation (RMSD) values in Å calculated by spatial comparison (heavy atoms) between the experimentally determined conformation of the co-crystallized ligand (PBDE-100) and its top-ranked docking poses, generated by the performed re-docking calculations using the homology model.

Model	**Homology Model**					
Ligand	PBDE-100					
Method	CDOCKER			GOLD		
Docked pose	Dock 1	Dock 2	Dock 3	Dock 1	Dock 2	Dock 3
RMSD (Å)	1.5697	1.6021	1.6427	0.5527	0.7988	0.8498

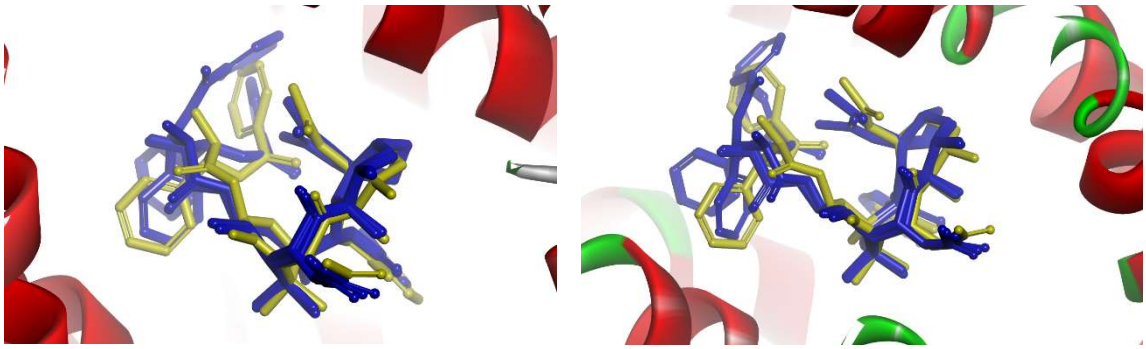
Table B.2: Root-mean-square deviation (RMSD) values in Å calculated by spatial comparison (heavy atoms) between the experimentally determined conformation of the cryoEM ligand (Taxol) and its top-ranked docking poses, generated by the performed re-docking calculations using the cryoEM structure of *h*P-gp.

Model	**cryoEM structure of *h*P-gp**					
Ligand	Taxol					
Method	CDOCKER			GOLD		
Docked pose	Dock 1	Dock 2	Dock 3	Dock 1	Dock 2	Dock 3
RMSD (Å)	1.2723	1.3208	1.4630	1.0283	1.1974	1.2669



(a) (b)

Figure B.1: Top-ranked binding poses obtained by re-docking calculations of the co-crystallized ligand PDBE-100 into its defined binding pocket in the homology model of *h*P-gp using: (a) CDOCKER algorithm; (b) GOLD algorithm. The experimental co-crystallized ligand is shown in solid yellow, while the calculated top-ranked ligand poses are represented in solid blue (Table B.1).



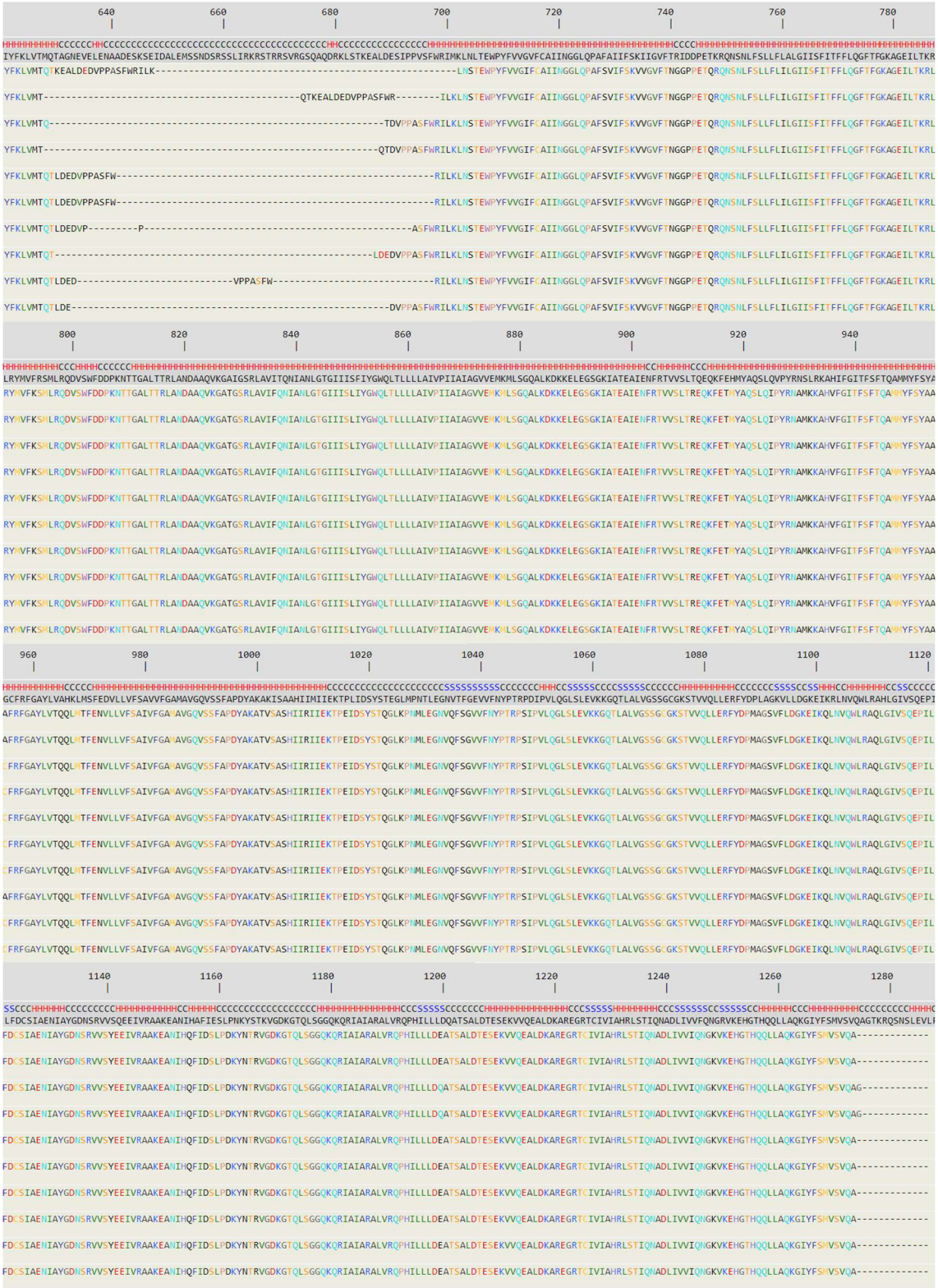
(a) (b)

Figure B.2: Top-ranked binding poses obtained by re-docking calculations of the cryoEM ligand (Taxol) into its defined binding pocket in the cryoEM structure *h*P-gp using: (a) CDOCKER algorithm; (b) GOLD algorithm. The experimental cryoEM ligand is shown in solid yellow, while the calculated top-ranked ligand poses are represented in solid blue (Table B.2).

B.2 Alignment of the *h*P-gp Sequence with the Different Templates

(a)


Rank PDB Hit Iden1 Iden2 Cov Norm. Z-score Download Align.
Sec.Str
Seq
1 user 0.88 0.82 0.92 10.00 Download
2 4m1mA 0.87 0.82 0.92 8.21 Download
3 5ko2A 0.89 0.82 0.92 8.29 Download
4 5kovA 0.89 0.82 0.92 8.21 Download
5 3g61A 0.88 0.82 0.92 6.68 Download
6 3g61A 0.88 0.82 0.92 6.99 Download
7 3g61 0.89 0.82 0.92 2.22 Download
8 3g5uA 0.89 0.82 0.92 13.72 Download
9 3g61A 0.88 0.82 0.92 7.48 Download
10 3g61 0.89 0.82 0.92 1.78 Download



(b)

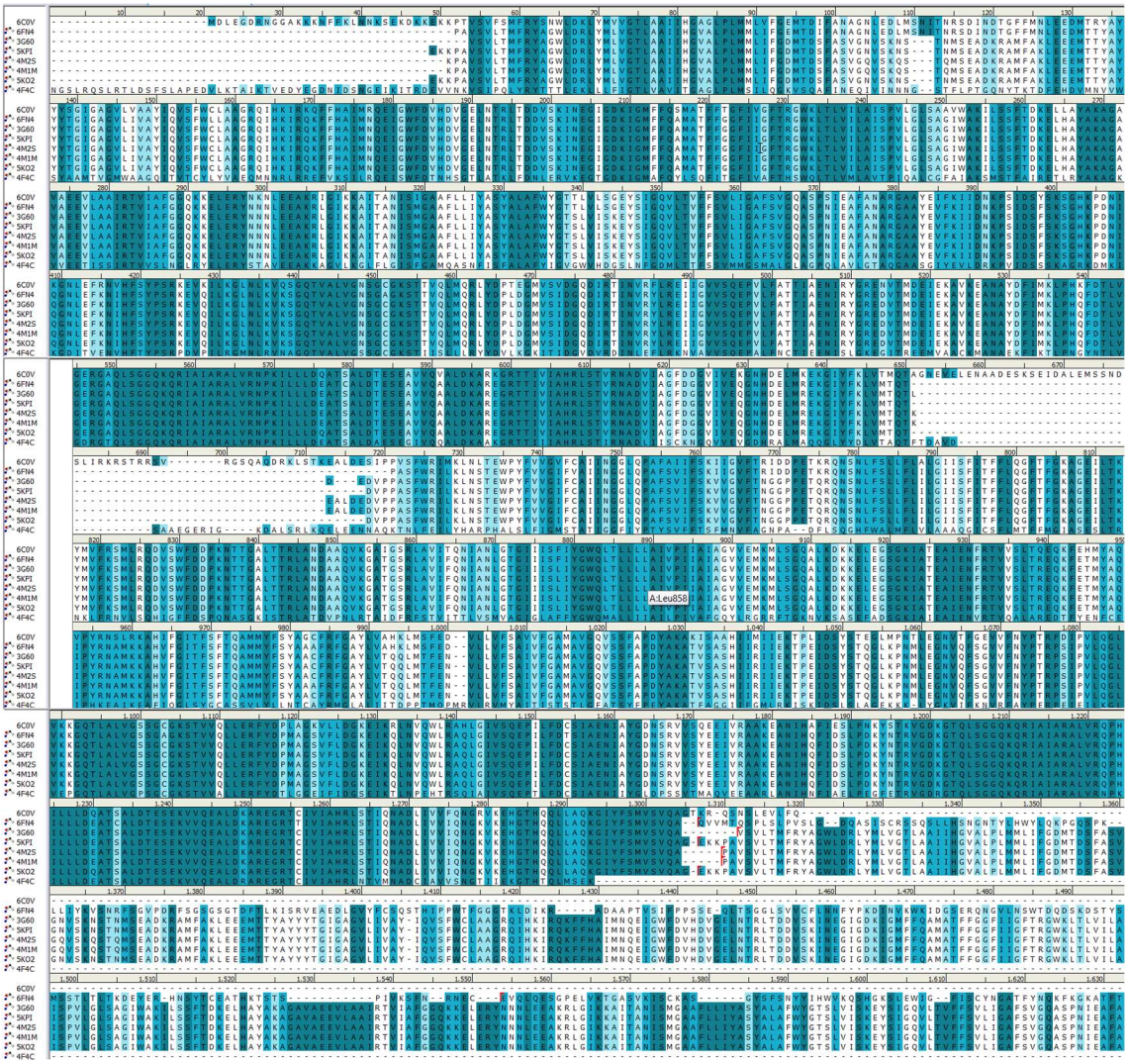


Figure B.3: Alignment of the *h*P-gp sequence with the different templates: (a) *m*P-gp (PDB ID: 4M1M); (b) *m*P-gp (PDB IDs: 4M1M, 5KO2, 5KOY, 3G61, 3G5U)); (c) *m*P-gp (PDB IDs: 6FN4, 4M1M, 5KPI, 4M2S, 5KO2, 3G60) and *C. elegans* P-gp (PDB ID: 4F4C).

B.3 Quality Assessment of the Truncated *h*P-gp Model (TMDs only)

Table B.3: Comparison of the Verify 3D, ERRAT and PROVE Scores of the truncated *h*P-gp model.

	Model TMD[1]	Model IT[2]	PDB ID: 4M1M[3]
Verify 3D	38.19%	63.41%	65.20%
ERRAT	95.1368	96.0884	86.5620
PROVE	5.0%	5.6%	0.0%

[1] Truncated *h*P-gp model (TMDs only). [2] Full-length I-TASSER *h*P-gp model. [3] Reference crystallographic structure of *m*P-gp

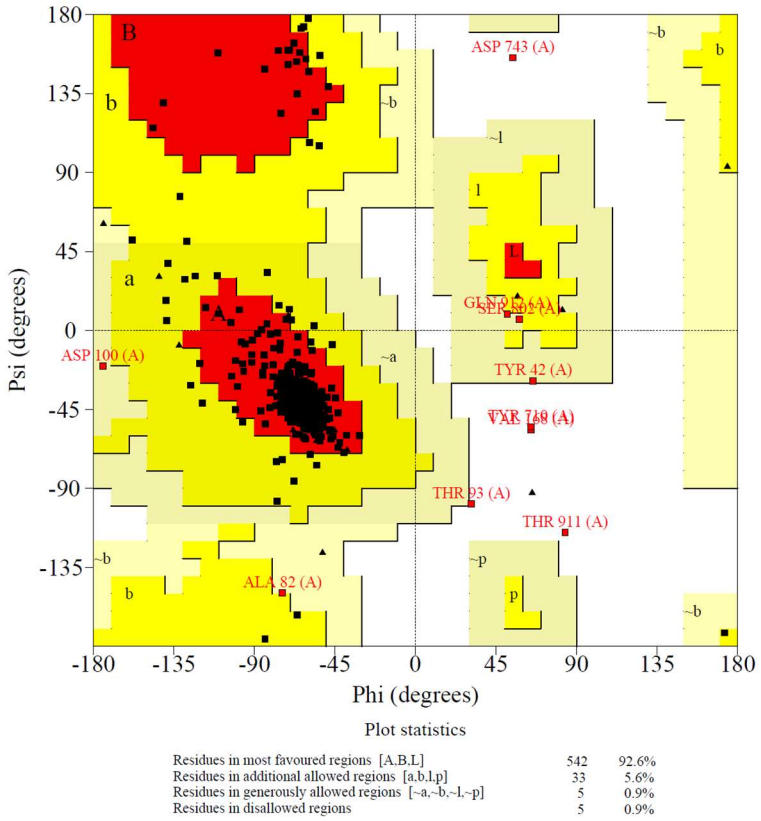


Figure B.4: Ramachandran plot of the truncated *h*P-gp model. The red, yellow, and white areas represent the favoured, allowed, and disallowed regions, respectively.

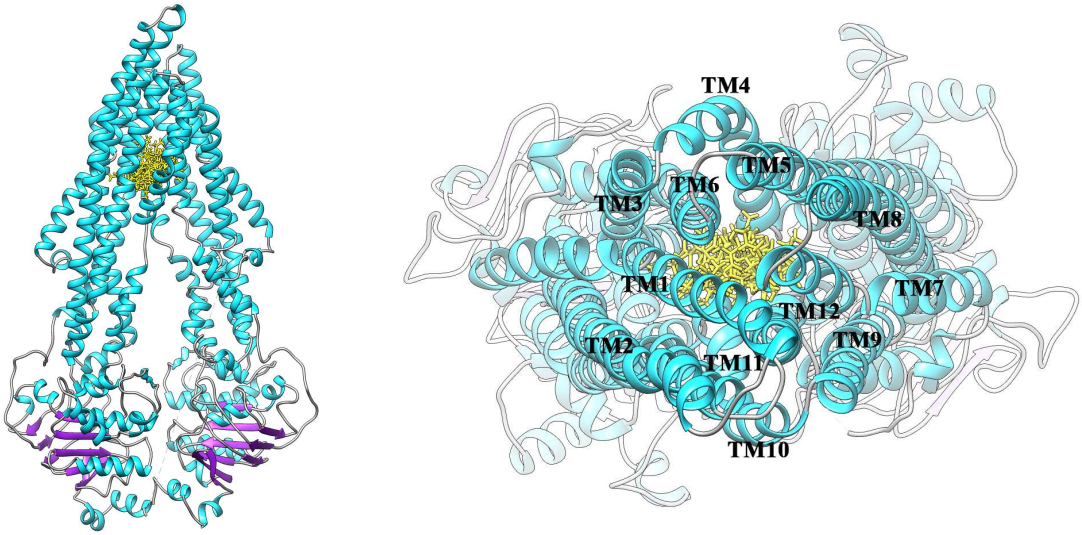
B.4 Nature of the Ligand–P-gp Interactions and Amino Acid Residues Involved in Binding

Table B.4: Ligand–P-gp interactions. Nature of the interactions with P-gp and amino acid residues involved in binding of the compounds under study, using the homology model and GOLD protocol. Numbers in parenthesis indicate the number of interactions involving the residue.

Name	H–Bond	Alkyl	π–Sigma	π–Alkyl	π–π	π–Sulphur	π–Lone Pair
CsA[1]	N721, Q838, F303, F994, A987, Q990	A987 (2), V991 (3), L339 (3), I306	-	F303, Y310, F343 (3)	-	-	-
AM[2]	-	A311, I340, M986	F759	Y307, Y310, F336, F759	F728, Y310	-	Y310
DOX[3]	A980, S979	A729	-	-	-	-	F983
DIG[4]	Q725	-	F983	Y310, F314, F336, F 343, F728 (2), F732, F883 (2)	-	-	-
LPM[5]	Y310	-	F728	F728, F983, L339, I340	F314, F732, F759	-	F983
RMP[6]	F728, S979, I306	-	-	Y307 (2), F336 (2), F343, F728 (2), F732(2), F983 (3), M986	-	-	-
VER[7]	A987 (2)	I306, L339	-	F343, A987	Y953	M68	-
CAR[8]	F983	-	F732	I735	F314, F336, F732, F759, F978	-	-
VPA[9]	F728	-	-	Y307, F314, F732, F759	-	-	-
BU[10]	Y307, F732, F759	-	-	-	-	F732, Y310	-
GEN[11]	-	I340	-	-	-	-	-
APD[12]	N721, Q725	-	-	-	-	-	F303
PQ[13]	S979	-	-	-	F728 (2), F732		

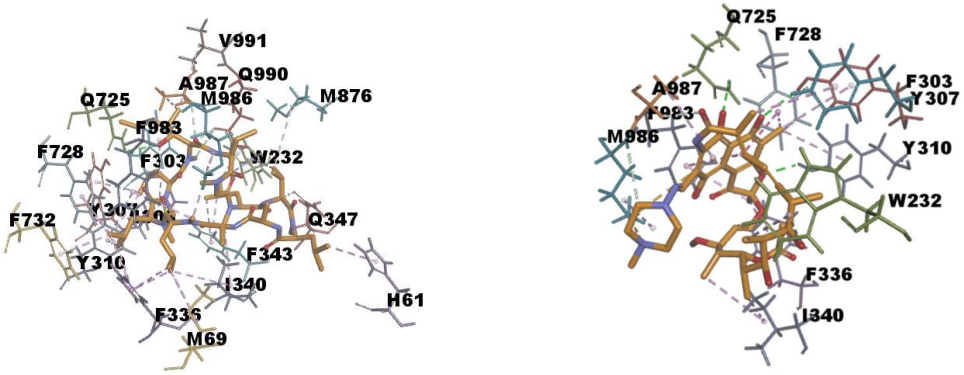
[1] Cyclosporine A; [2] amiodarone; [3] doxorubicin; [4] digoxin; [5] loperamide; [6] rifampin; [7] verapamil; [8] carvedilol; [9] valproic acid; [10] busulfan; [11] gentamicin; [12] pamidronate; [13] paraquat

B.5 Docking into the *h*P-gp cryoEM Structure

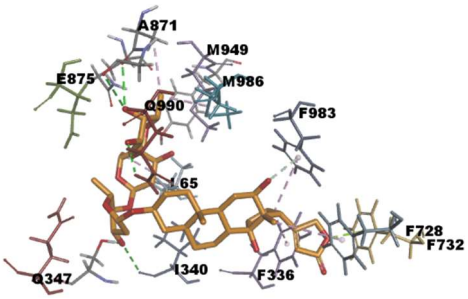


(a) (b)

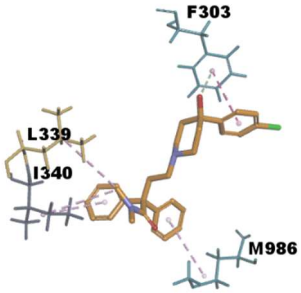
Figure B.5: Distribution of the selected ligand poses (yellow) in the experimentally solved cryoEM structure of *h*P-gp (PDB ID: 6QEX). (a) Frontal view; (b) View from the extracellular side of the protein looking into the internal chamber. The colour representation is according to the secondary structure: helices are shown in cyan, beta sheets in purple, and coils in grey.



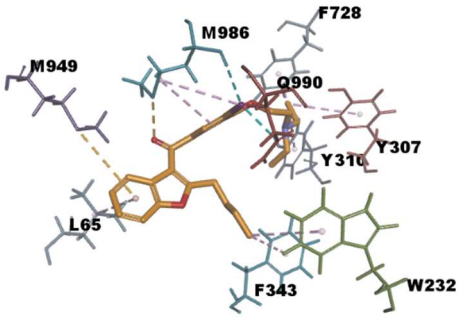
(a) (b)


A871
M949
E875
M986
Q990
F983
L65
F728
F732
Q347
I340
F336

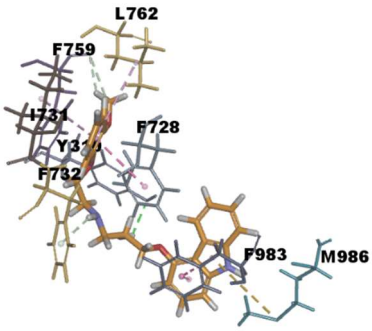
(c)


F303
L339
I340
M986

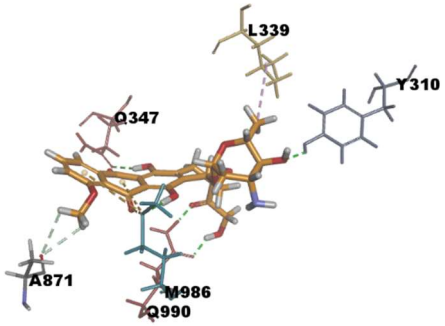
(d)


F728
M986
M949
Q990
Y307
Y310
L65
F343
W232

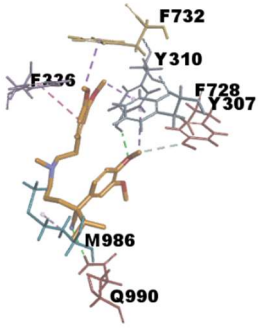
(e)


L762
F759
I731
F728
F732
F983
M986

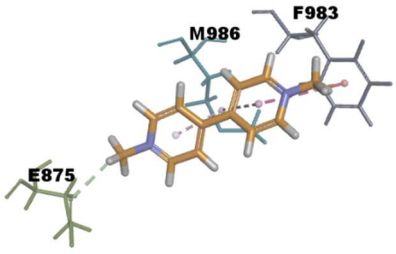
(f)


L339
Y310
Q347
A871
M986
Q990

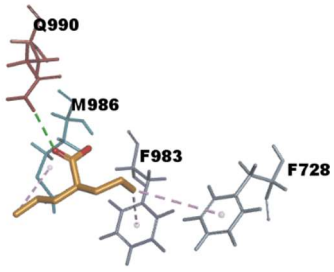
(g)


F732
Y310
F728
Y307
M986
Q990

(h)


F983
M986
E875

(i)


Q990
M986
F983
F728

(j)

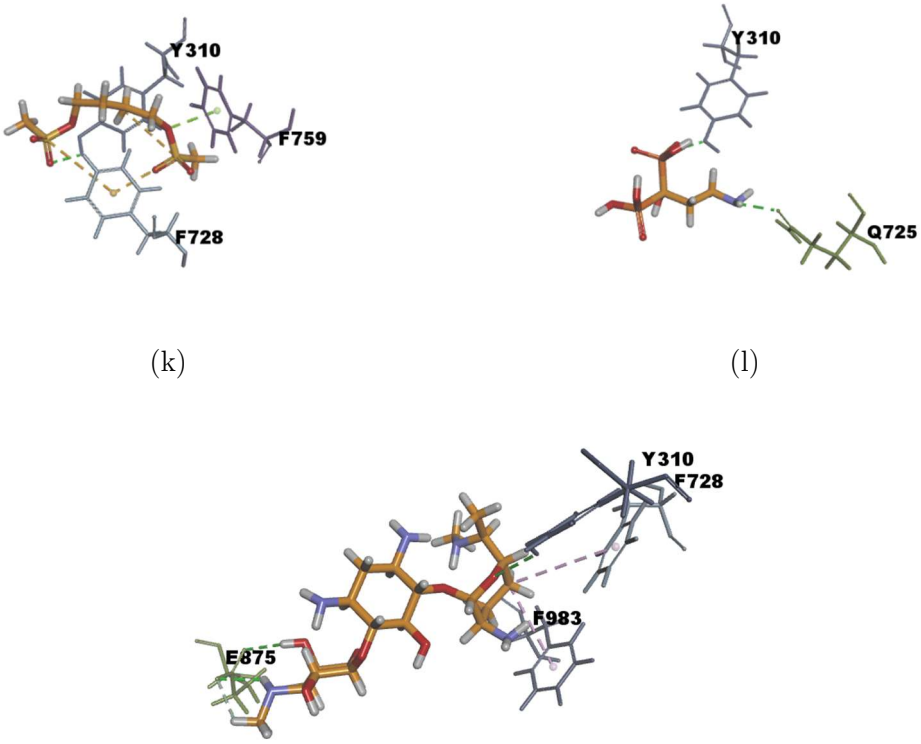
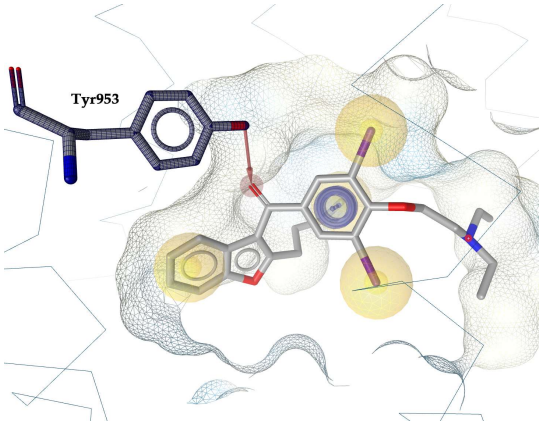


Figure B.6: 3D representation of the top-ranked poses and their interactions in the binding pocket obtained with the CDOCKER algorithm in the experimentally solved cryoEM structure of *h*P-gp (PDB ID: 6QEX). (a) Cyclosporine A; (b) Rifampin; (c) Digoxin; (d) Loperamide; (e) Amiodarone; (f) Carvedilol; (g) Doxorubicin; (h) Verapamil; (i) Paraquat; (j) Valproic Acid; (k) Busulfan; (l) Pamidronate; (m) Gentamicin. Green dotted lines represent conventional hydrogen bonds, light-green dotted lines represent carbon hydrogen bonds, light-rose dotted lines represent hydrophobic interactions, orange dotted lines represent π-sulphur interactions and sulphur-X interactions, cyan dotted line represents halogen interactions and fluorescent green represents π-lone pair interactions.

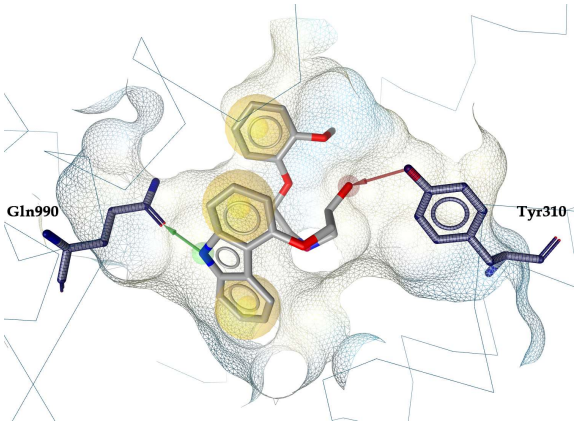
Appendix C

Chapter 5 Supporting Materials

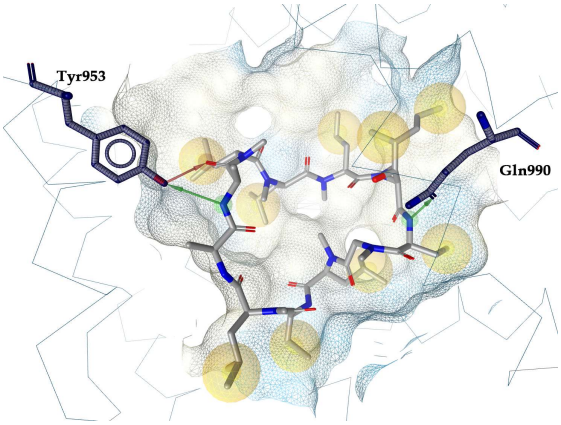
C.1 Ligand–P-gp Interactions



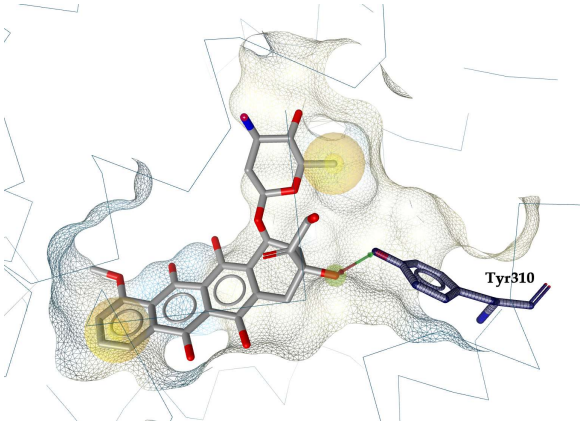
(a)



(b)



(c)



(d)

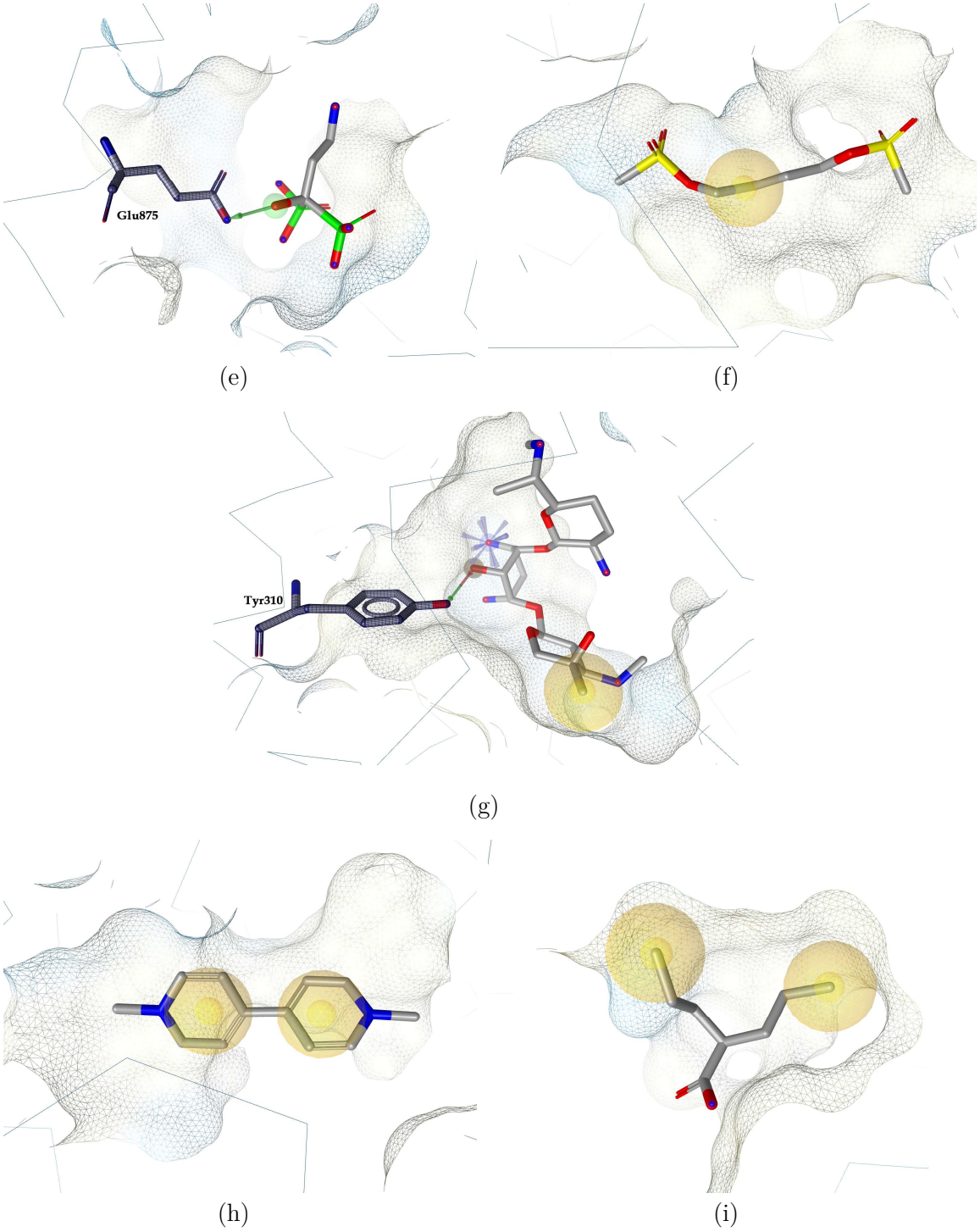
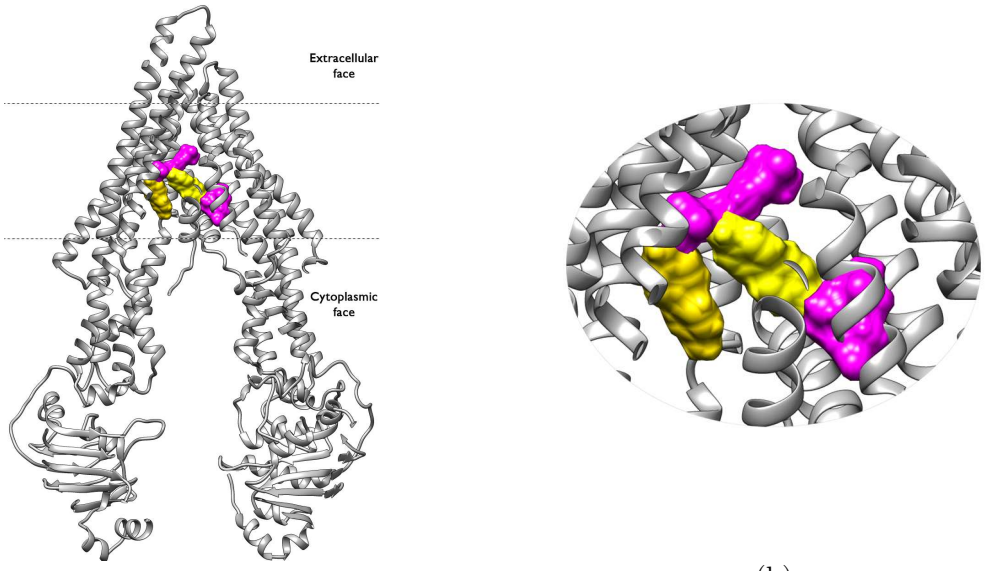


Figure C.1: 3D representation of the most relevant ligand–P-gp interactions within the binding pocket. (a) P-gp–AMI; (b) P-gp–CAR; (c) P-gp–CSA; (d) P-gp–DOX; (e) P-gp–APD; (f) P-gp–BUS; (g) P-gp–GEN; (h) P-gp–PQT; (i) P-gp–VPA. The binding pocket is shown in surface representation with a colour scheme corresponding to the hydrophobicity; non-polar regions are coloured yellow. Residues involved in hydrogen bonding are exposed and highlighted with a dark blue mesh. Red arrows indicate hydrogen bond acceptor relationships, green arrows indicate hydrogen bond donor relationships, yellow spheres indicate hydrophobic interactions, and the blue ring

indicates aromatic interactions.



(a) (b)

Figure C.2: BUS (magenta) and PQT (yellow) molecular surface representation of their different positions within the binding pocket during the 500 ns production run; (a) frontal view; (b) zoomed view; (c) view from the extracellular side of the protein looking into the inner chamber.

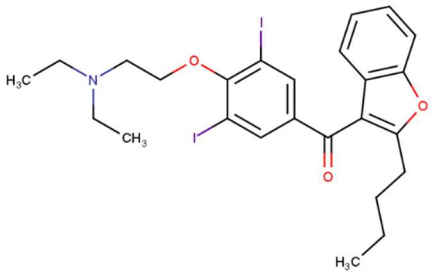
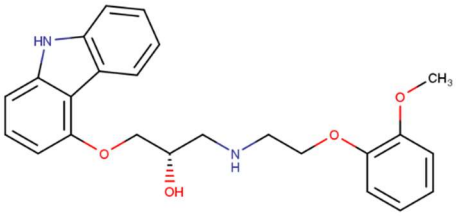
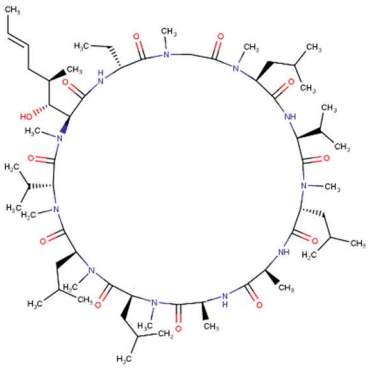
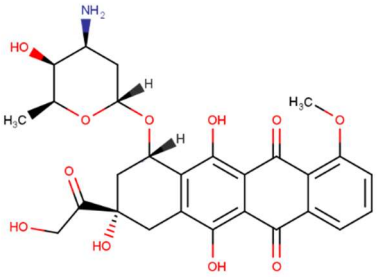
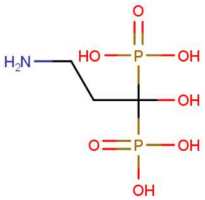
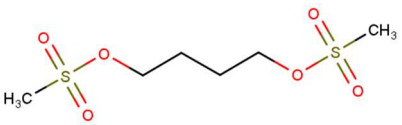
C.2 Compounds Properties

Table C.5: Physicochemical properties of the studied molecules.

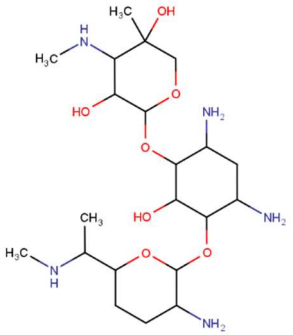
	LogP	**HBD**[10]	**HBA**[11]	**TPSA**[12] **(Å²)**	**Heavy atom count**	**Aromatic rings**
AMI[1]	7.57	0	4	42.7	31	3
CAR[2]	4.19	3	5	75.7	30	4
CSA[3]	2.92	5	12	279.0	85	0
DOX[4]	1.27	6	12	206.0	39	2
APD[5]	-4.70	6	8	161.0	13	0
BUS[6]	-0.52	0	6	104.0	14	0
GEN[7]	-3.10	8	12	200.0	33	0
PQT[8]	-4.22	0	0	7.8	14	2
VPA[9]	2.75	1	2	37.3	10	0

[1] Amiodarone; [2] carvedilol; [3] cyclosporine A; [4] doxorubicin; [5] pamidronate; [6] busulfan; [7] gentamicin; [8] paraquat; [9] valproic acid; [10] hydrogen bond donor count; [11] hydrogen bond acceptor count; [12] topological polar surface area.

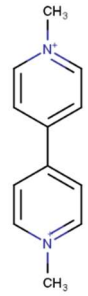
Table C.6: 2D structures of the molecules used in the study.

Name	Chemical structure
Amiodarone	
Carvedilol	
Cyclosporine A	
Doxorubicin	
Pamidronate	
Busulfan	

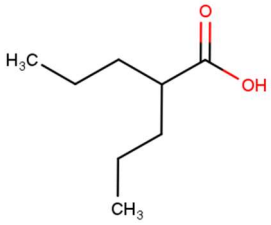
Gentamicin



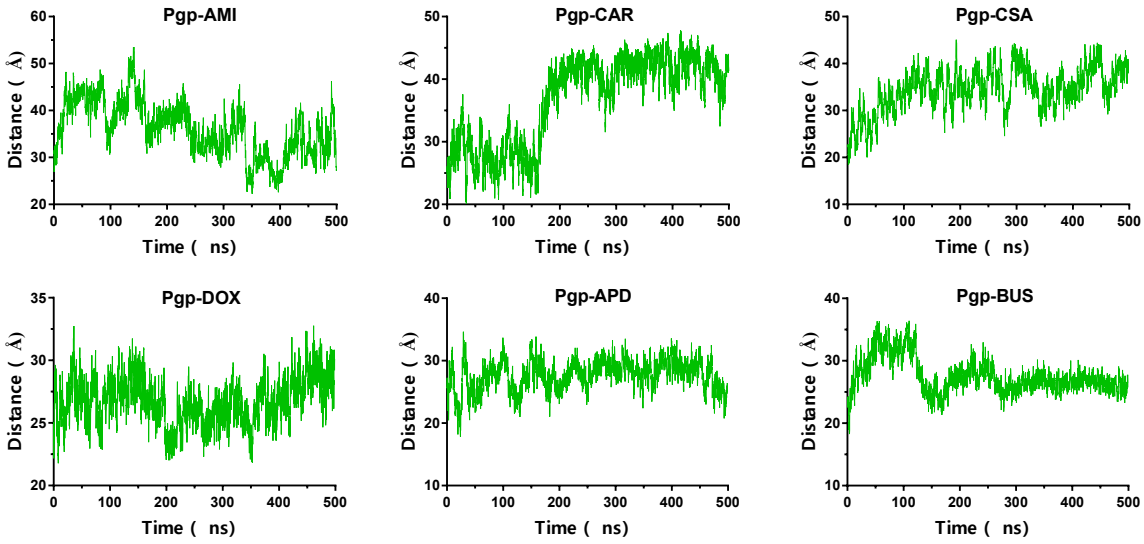
Paraquat



Valproic acid



C.3 NBDs Distance



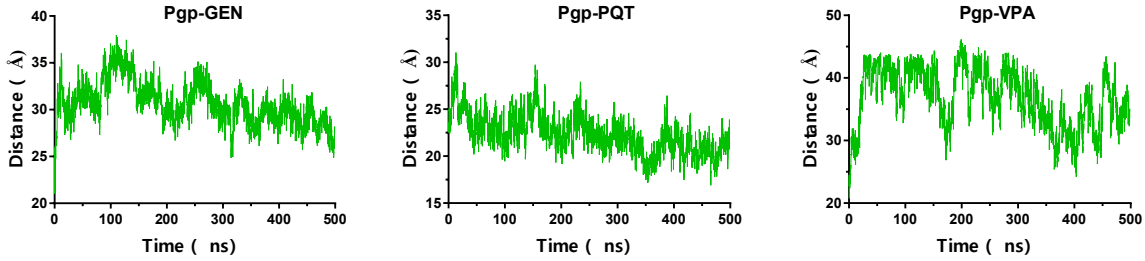


Figure C.3: NBDs distance during the 500 ns production run. The separation was measured by the distance between the N atom in the Lys residue of the Walker A motif in NBD1 and the Cα of the Ser residue in the signature motif of NBD2.

C.4 Principal Component Analysis

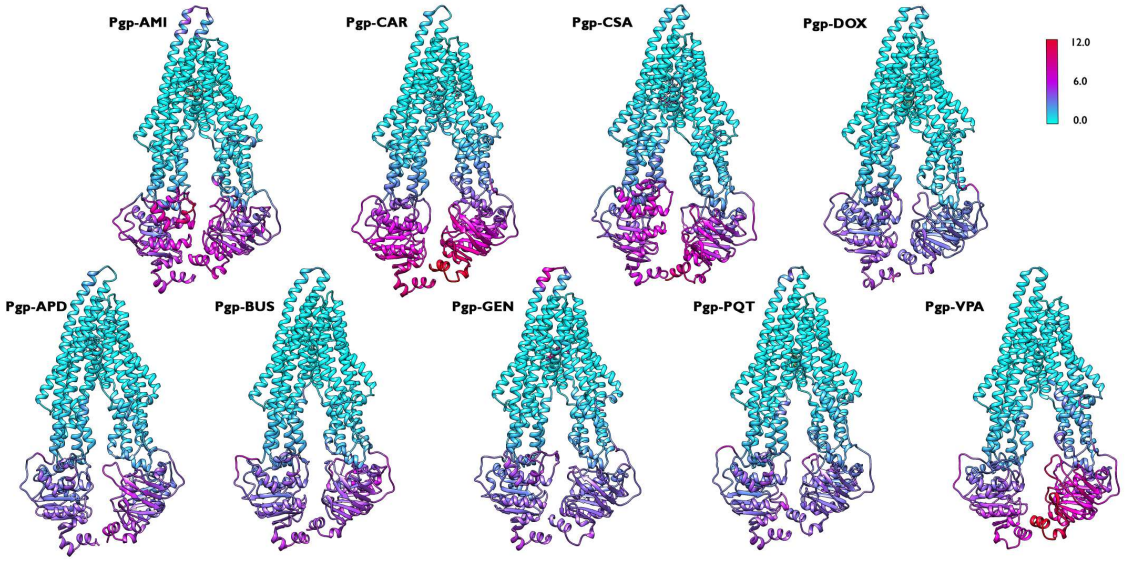


Figure C.4: Cα-RMSF coloured representation for the P-gp–ligand systems along the principal components explaining at least 85% of the flexibility of the systems calculated from the 500 ns production run. The flexibility scale goes from cyan (lower values) to red (higher values).

C.5 Solvent Accessible Surface Area (SASA)

Table C.7: Per residue SASA variations.

Residue number	Average per residue SASA[a]								
	AMI[1]	CAR[2]	CSA[3]	DOX[4]	APD[5]	BUS[6]	GEN[7]	PQT[8]	VPA[9]
33	76.40	76.35	74.60	76.21	78.35	78.14	78.34	77.31	78.38
69	1.32	1.97	1.03	2.12	7.01	6.96	3.26	7.81	5.41
313	0.45	0.49	0.57	0.49	1.60	0.81	0.68	0.91	0.99
331	0.79	0.83	0.78	0.48	1.01	1.51	1.33	4.50	2.03
340	9.37	9.73	0.87	10.06	23.13	22.55	16.07	22.38	23.00
343	19.51	8.13	0.74	9.49	20.70	28.57	22.32	20.01	29.48
367	99.76	105.85	107.04	103.98	110.96	108.63	107.40	114.38	110.99
372	144.74	94.67	146.75	97.70	149.29	165.13	174.33	157.79	164.71
548	57.70	86.00	69.94	82.09	90.66	94.25	91.88	96.92	86.63
727	32.49	35.86	35.74	35.90	36.25	37.53	36.90	36.54	37.48
764	86.19	84.57	85.95	85.44	88.84	91.01	96.07	88.62	89.49
767	65.03	66.33	61.45	64.98	67.18	68.46	70.77	67.62	68.19
953	1.75	6.35	3.97	5.07	8.53	11.03	7.74	9.20	16.38
986	3.88	7.91	5.87	10.83	15.99	29.12	17.45	16.51	27.78

Residue number	Average per residue SASA[a]	
	Min(NA[10]) - Max(A[11])	% Decrease relative to Max(NA)
33	0.91	1.19
69	1.14	53.70
313	0.11	18.37
331	0.18	22.24
340	6.01	59.68
343	0.50	2.55
367	0.36	0.34
372	2.54	1.73
548	0.62	0.73
727	0.35	0.96
764	2.43	2.82
767	0.84	1.27
953	1.40	21.99
986	5.16	47.66

[1] P-gp–AMI; [2] P-gp–CAR; [3] P-gp–CSA; [4] P-gp–DOX; [5] P-gp–APD; [6] P-gp–BUS; [7] P-gp–GEN; [8] P-gp–PQT; [9] P-gp–VPA; [10] Non-active-bound system; [11] Active-bound system; [a] Solvent accessible surface area.

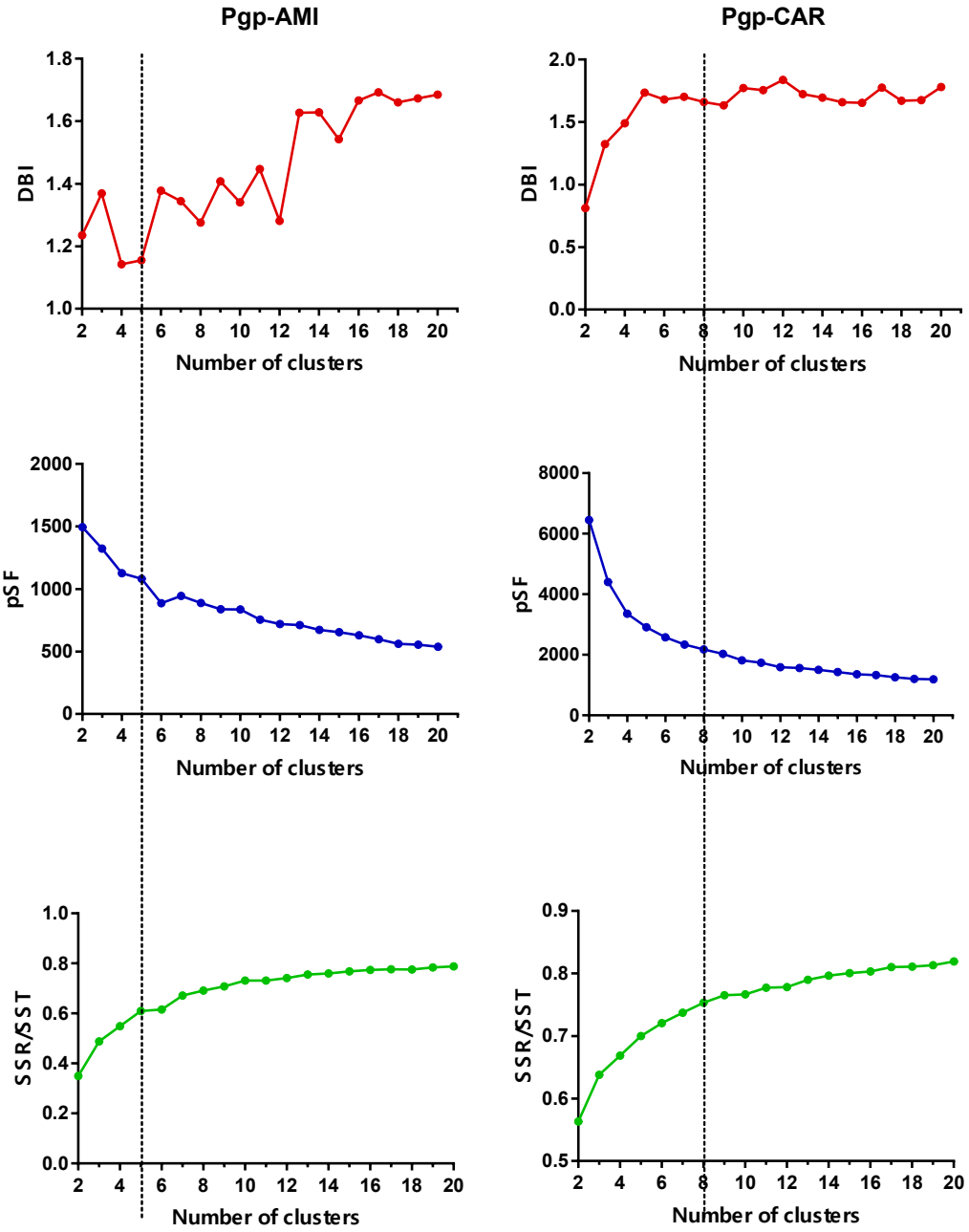
Table C.8: Per residue SASA variations.

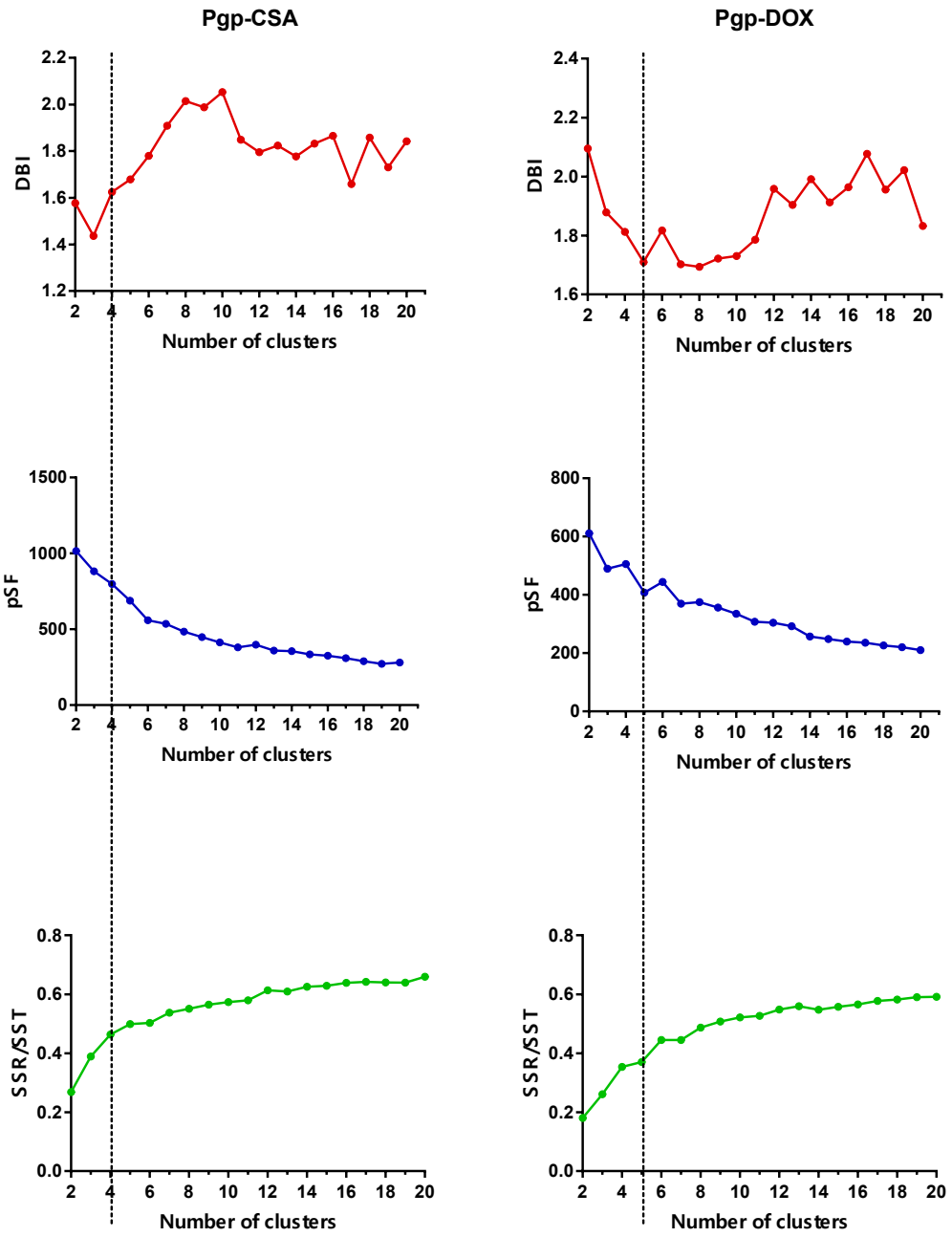
Residue number	Average per residue SASA[a]								
	AMI[1]	CAR[2]	CSA[3]	DOX[4]	APD[5]	BUS[6]	GEN[7]	PQT[8]	VPA[9]
38	47.78	46.18	48.30	47.63	42.49	45.82	45.53	46.00	35.60
184	39.37	49.12	67.59	40.81	36.45	25.02	34.34	30.22	32.35
188	38.64	35.27	39.26	36.05	32.10	29.39	21.38	26.46	31.23
320	8.44	7.69	8.26	8.29	7.51	6.21	5.24	5.32	4.81
391	49.94	39.60	40.22	40.36	38.24	31.79	37.17	36.95	36.40
420	30.49	38.18	32.18	33.64	10.85	5.52	29.27	7.32	28.27
559	62.70	95.91	73.74	71.91	44.18	10.17	5.78	49.84	39.52
580	45.61	42.93	56.46	39.21	23.57	13.20	22.59	18.83	23.87
581	2.58	0.44	1.16	0.40	0.07	0.10	0.20	0.13	0.25
601	4.97	5.84	6.96	6.29	1.69	4.89	2.92	4.58	2.28
1021	118.28	147.02	126.67	138.48	90.40	108.25	109.98	86.89	103.94
1040	0.12	0.09	0.12	0.10	0.01	0.06	0.07	0.02	0.05
1049	102.06	125.42	111.60	117.66	80.14	94.11	87.75	85.74	88.55
1090	124.00	123.14	124.24	124.84	110.54	121.46	122.47	118.19	121.07
1154	2.03	0.86	0.96	0.99	0.41	0.66	0.67	0.83	0.72

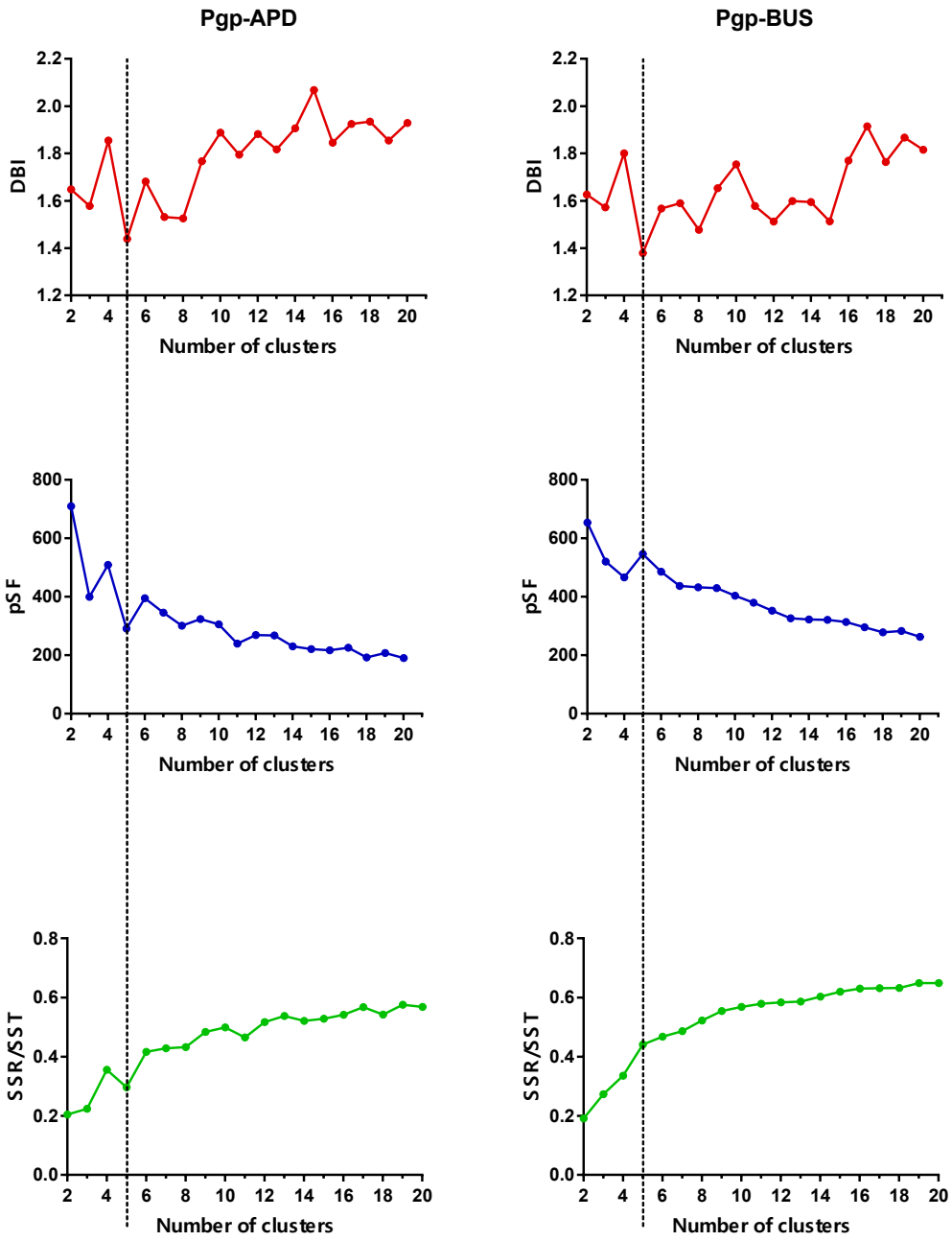
Residue number	Average per residue SASA[a]	
	Min(A[11]) - Max(NA[10])	% Decrease relative to Max(NA)
38	0.18	0.40
184	2.91	7.99
188	3.16	9.85
320	0.18	2.37
391	1.36	3.55
420	1.22	4.17
559	12.86	25.80
580	15.35	64.29
581	0.15	58.04
601	0.08	1.68
1021	8.30	7.55
1040	0.02	28.27
1049	7.95	8.45
1090	0.67	0.55
1154	0.03	3.53

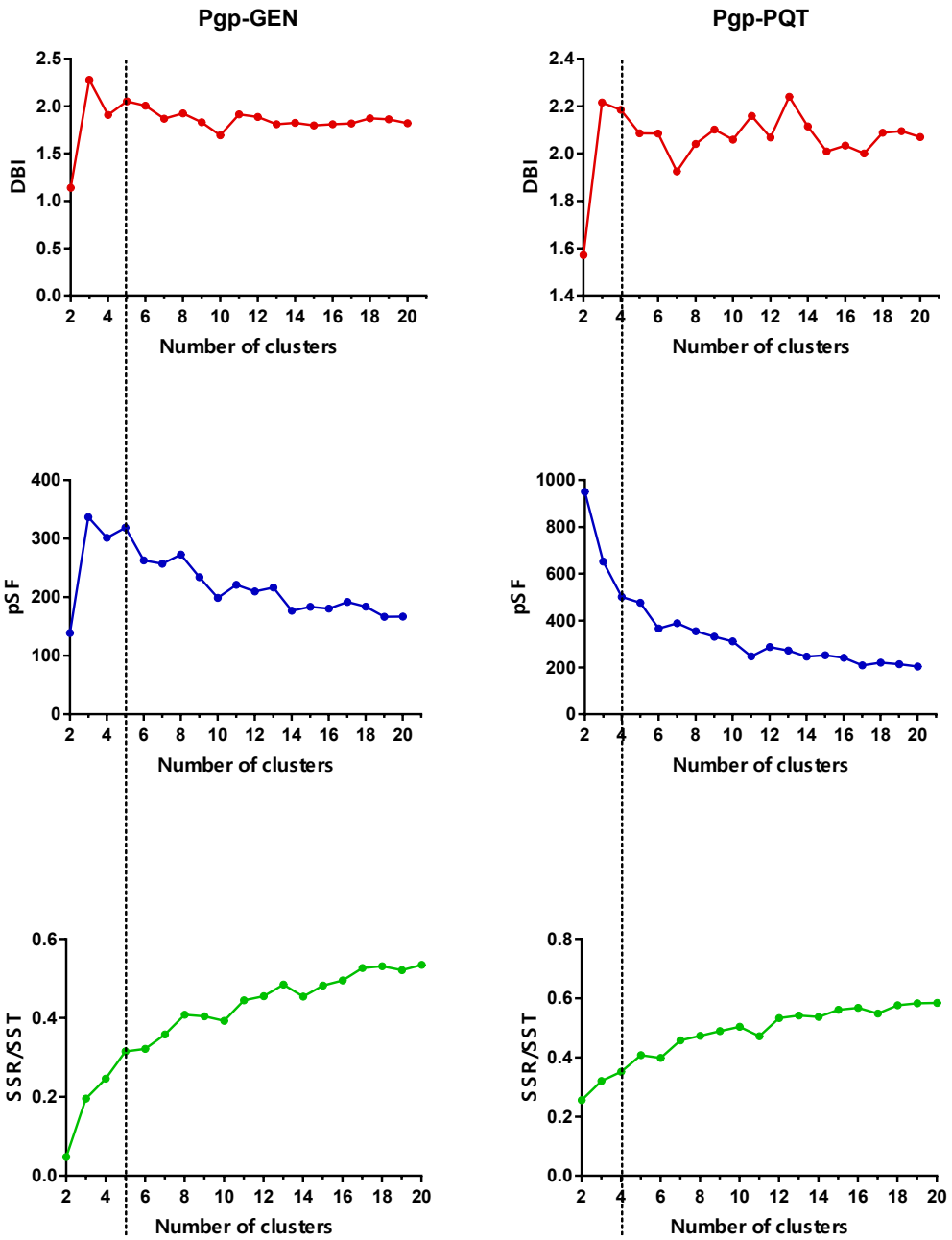
[1] P-gp–AMI; [2] P-gp–CAR; [3] P-gp–CSA; [4] P-gp–DOX; [5] P-gp–APD; [6] P-gp–BUS; [7] P-gp–GEN; [8] P-gp–PQT; [9] P-gp–VPA; [10] Non-active-bound system; [11] Active-bound system; [a] Solvent accessible surface area.

C.6 Metrics of the Clustering Analysis




Pgp-CSA
Pgp-DOX
DBI
pSF
SSR/SST
Number of clusters


Pgp-APD
Pgp-BUS
DBI
pSF
SSR/SST
Number of clusters


Pgp-GEN
Pgp-PQT
DBI
pSF
SSR/SST
Number of clusters

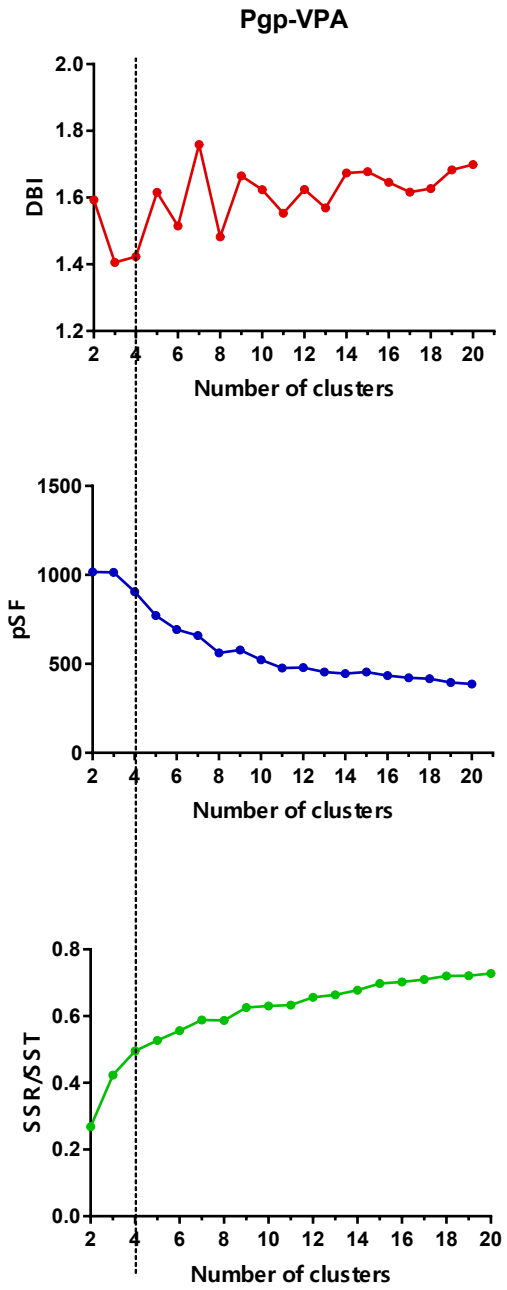


Figure C.5: Metrics used to select of the optimal number of clusters of each ligand–P-gp system. The dashed lines in the graph indicate the selected number of clusters.

C.7 Energy of the System

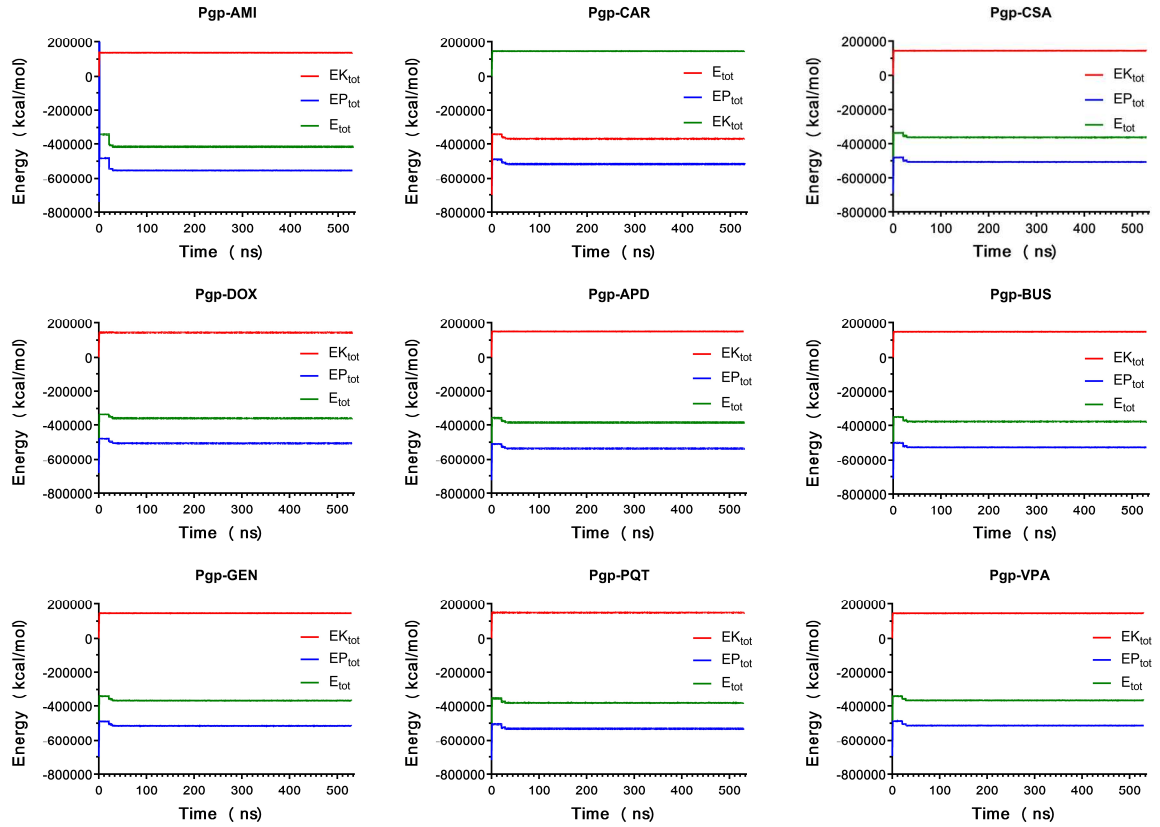


Figure C.6: Energy of the simulated systems during the 500 ns production run. The red line shows the kinetic energy, and the blue line shows the potential energy. The green line shows the total energy.

C.8 MM/PBSA Free Energies of Binding

Table C.9: Free energies of binding and the various MM/PBSA terms. Estimate of the overall binding free energies for the ligand–P-gp complexes studied, using MM/PBSA calculations.

Name	$\mathbf{\Delta G_{Binding}}$ **(kcal/mol)**	$\mathbf{E_{VDW}}$	$\mathbf{E_{elec}}$	$\mathbf{G_{non\text{-}polar}}$	$\mathbf{G_{Disper}}$	$\mathbf{\Delta G_{Gas}}$	$\mathbf{\Delta G_{Solv}}$
CSA[1]	-55.0940	-105.2820	-19.4685	-80.3234	149.9800	-124.7505	69.6565
AMI[2]	-31.2295	-53.4501	-10.3573	-39.3306	71.9084	-63.8074	32.5778
CAR[3]	-23.9604	-41.3819	-10.6671	-32.2420	60.3305	-52.0490	28.0885
DOX[4]	-24.7911	-47.6753	-9.8549	-36.7381	69.4773	-57.5302	32.7392
GEN[5]	-20.0385	-41.0188	-7.5270	-35.1429	63.6502	-48.5458	28.5073
BUS[6]	-12.7677	-25.2568	-3.2940	-18.9787	34.7618	-28.5508	15.7831
APD[7]	-15.4427	-21.1230	-9.6178	-14.8442	30.1422	-30.7407	15.2980
PQT[8]	-18.2873	-20.0311	-13.7060	-15.0945	30.5442	-33.7371	15.4497
VPA[9]	-9.1548	-17.7954	-2.5063	-15.4406	26.5875	-20.3017	11.1469

[1] Cyclosporine A; [2] amiodarone; [3] carvedilol; [4] doxorubicin; [5] gentamicin; [6] busulfan; [7] pamidronate; [8] paraquat; [9] valproic acid.

References

Accelrys, P. E. (2014). BIOVIA Pipeline Pilot (Version 9.2). San Diego, CA, USA: Dassault Systèmes. Retrieved from http://www.3dsbiovia.com/products/collaborative-science/biovia-pipeline-pilot/

Accelrys, P. E. (2017). Discovery Studio Modeling Environment (Version 4.1). San Diego, USA: Dassault Systèmes BIOVIA. Retrieved from https://www.3dsbiovia.com/products/collaborative-science/biovia-discovery-studio/

Akiyama, S.-I., Cornwell, M. M., Kuwano, M., Pastan, I., & Gottesman, M. M. (1988). Most drugs that reverse multidrug resistance also inhibit photoaffinity labeling of P-glycoprotein by a vinblastine analog. *Molecular pharmacology, 33*(2), 144-147.

Alam, A., Kowal, J., Broude, E., Roninson, I., & Locher, K. P. (2019). Structural insight into substrate and inhibitor discrimination by human P-glycoprotein. *Science, 363*(6428), 753-756.

Aller, S. G., Yu, J., Ward, A., Weng, Y., Chittaboina, S., Zhuo, R., . . . Chang, G. (2009). Structure of P-glycoprotein reveals a molecular basis for poly-specific drug binding. *Science, 323*(5922), 1718-1722. doi: 10.1126/science.1168750

Ambudkar, S. V., Kim, I.-W., Xia, D., & Sauna, Z. E. (2006). The A-loop, a novel conserved aromatic acid subdomain upstream of the Walker A motif in ABC transporters, is critical for ATP binding. *FEBS letters, 580*(4), 1049-1055.

Bajusz, D., Rácz, A., & Héberger, K. (2019). Comparison of Data Fusion Methods as Consensus Scores for Ensemble Docking. *Molecules, 24*(15), 2690.

Bakan, A., Meireles, L. M., & Bahar, I. (2011). ProDy: protein dynamics inferred from theory and experiments. *Bioinformatics, 27*(11), 1575-1577.

Ballabio, D., Grisoni, F., & Todeschini, R. (2018). Multivariate comparison of classification performance measures. *Chemometrics and Intelligent Laboratory Systems, 174*, 33-44. doi: 10.1016/j.chemolab.2017.12.004

Beck, W. T., Cirtain, M. C., Glover, C. J., Felsted, R. L., & Safa, A. R. (1988). Effects of indole alkaloids on multidrug resistance and labeling of P-glycoprotein by a photoaffinity analog of vinblastine. *Biochemical and biophysical research communications, 153*(3), 959-966.

Becker, J.-P., Depret, G., Van Bambeke, F., Tulkens, P. M., & Prévost, M. (2009). Molecular models of human P-glycoprotein in two different catalytic states. *BMC structural biology, 9*(1), 3.

Benfenati, E., Manganaro, A., & Gini, G. C. (2013). *VEGA-QSAR: AI Inside a Platform for Predictive Toxicology.* Paper presented at the PAI@ AI* IA.

Benkert, P., Biasini, M., & Schwede, T. (2011). Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics, 27*(3), 343-350. doi: 10.1093/bioinformatics/btq662

Beno, B. R., Yeung, K.-S., Bartberger, M. D., Pennington, L. D., & Meanwell, N. A. (2015). A survey of the role of noncovalent sulfur interactions in drug design. *J. Med. Chem., 58*(11), 4383-4438.

Bikadi, Z., Hazai, I., Malik, D., Jemnitz, K., Veres, Z., Hari, P., . . . Hazai, E. (2011). Predicting P-glycoprotein-mediated drug transport based on support vector machine and three-dimensional crystal structure of P-glycoprotein. *PloS one, 6*(10), e25815.

Borst, P., & Elferink, R. O. (2002). Mammalian ABC transporters in health and disease. *Annual review of biochemistry, 71*(1), 537-592.

Bowie, J. U., Luthy, R., & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science, 253*(5016), 164-170.

Broccatelli, F., Carosati, E., Neri, A., Frosini, M., Goracci, L., Oprea, T. I., & Cruciani, G. (2011). A novel approach for predicting P-glycoprotein (ABCB1) inhibition using molecular interaction fields. *J. Med. Chem., 54*(6), 1740-1751. doi: 10.1021/jm101421d

Bruggemann, E., Germann, U., Gottesman, M. M., & Pastan, I. (1989). Two different regions of P-glycoprotein [corrected] are photoaffinity-labeled by azidopine. *Journal of Biological Chemistry, 264*(26), 15483-15488.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics, 10*, 421. doi: 10.1186/1471-2105-10-421

Cascorbi, I. (2006). Role of pharmacogenetics of ATP-binding cassette transporters in the pharmacokinetics of drugs. *Pharmacology & therapeutics, 112*(2), 457-473.

Chen, J., Lu, G., Lin, J., Davidson, A. L., & Quiocho, F. A. (2003). A tweezers-like motion of the ATP-binding cassette dimer in an ABC transport cycle. *Molecular cell, 12*(3), 651-661.

Chen, L., Li, Y., Yu, H., Zhang, L., & Hou, T. (2012). Computational models for predicting substrates or inhibitors of P-glycoprotein. *Drug discovery today, 17*(7-8), 343-351.

Chen, L., Li, Y., Zhao, Q., Peng, H., & Hou, T. (2011). ADME evaluation in drug discovery. 10. Predictions of P-glycoprotein inhibitors using recursive partitioning and naive Bayesian classification techniques. *Mol. Pharm., 8*(3), 889-900. doi: 10.1021/mp100465q

Cheng, F., Li, W., Zhou, Y., Shen, J., Wu, Z., Liu, G., . . . Tang, Y. (2012). admetSAR: A Comprehensive Source and Free Tool for Assessment of Chemical ADMET Properties. *J. Chem. Inf. Model., 52*(11), 3099-3105. doi: 10.1021/ci300367a

Cianchetta, G., Singleton, R. W., Zhang, M., Wildgoose, M., Giesing, D., Fravolini, A., . . . Vaz, R. J. (2005). A pharmacophore hypothesis for P-glycoprotein substrate recognition using GRIND-based 3D-QSAR. *J. Med. Chem., 48*(8), 2927-2935.

Cirrito, J. R., Deane, R., Fagan, A. M., Spinner, M. L., Parsadanian, M., Finn, M. B., . . . Bales, K. R. (2005). P-glycoprotein deficiency at the blood-brain barrier

increases amyloid-β deposition in an Alzheimer disease mouse model. *J Clin Invest, 115*(11), 3285-3290.

Colabufo, N. A., Berardi, F., Cantore, M., Contino, M., Inglese, C., Niso, M., & Perrone, R. (2010). Perspectives of P-glycoprotein modulating agents in oncology and neurodegenerative diseases: pharmaceutical, biological, and diagnostic potentials. *J. Med. Chem., 53*(5), 1883-1897.

Cole, J. C., Murray, C. W., Nissink, J. W. M., Taylor, R. D., & Taylor, R. (2005). Comparing protein–ligand docking programs is difficult. *Proteins: Structure, Function, and Bioinformatics, 60*(3), 325-332.

Collett, A., Tanianis-Hughes, J., Hallifax, D., & Warhurst, G. (2004). Predicting P-glycoprotein effects on oral absorption: correlation of transport in Caco-2 with drug pharmacokinetics in wild-type and mdr1a (-/-) mice in vivo. *Pharmaceutical research, 21*(5), 819-826.

Colovos, C., & Yeates, T. O. (1993). Verification of protein structures: patterns of nonbonded atomic interactions. *Protein science, 2*(9), 1511-1519.

Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., . . . Kollman, P. A. (1995). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society, 117*(19), 5179-5197.

Cramer, R. D., Patterson, D. E., & Bunce, J. D. (1988). Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *Journal of the American Chemical Society, 110*(18), 5959-5967.

Crowley, E., O'Mara, M. L., Kerr, I. D., & Callaghan, R. (2010). Transmembrane helix 12 plays a pivotal role in coupling energy provision and drug binding in ABCB1. *The FEBS journal, 277*(19), 3974-3985.

Crowley, M. F., Williamson, M. J., & Walker, R. C. (2009). CHAMBER: Comprehensive support for CHARMM force fields within the AMBER software. *International Journal of Quantum Chemistry, 109*(15), 3767-3772.

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

de Lange, E. (2007). Multi drug resistance P glycoprotein and other transporters.

Dearden, J., Al-Noobi, A., Scott, A., & Thomson, S. (2003). QSAR studies on P-glycoprotein-regulated multidrug resistance and on its reversal by phenothiazines. *SAR and QSAR in Environmental Research, 14*(5-6), 447-454.

Demel, M. A., Krämer, O., Ettmayer, P., Haaksma, E. E., & Ecker, G. F. (2009). Predicting ligand interactions with ABC transporters in ADME. *Chem. Biodivers., 6*(11), 1960-1969. doi: 10.1002/cbdv.200900138

Dias, R., de Azevedo, J., & Walter, F. (2008). Molecular docking algorithms. *Current drug targets, 9*(12), 1040-1047.

Dolghih, E., Bryant, C., Renslo, A. R., & Jacobson, M. P. (2011). Predicting binding to p-glycoprotein by flexible receptor docking. *PLoS Comput. Biol., 7*(6), e1002083. doi: 10.1371/journal.pcbi.1002083

Domicevica, L., & Biggin, P. C. (2015). Homology modelling of human P-glycoprotein. *Biochemical Society Transactions, 43*(5), 952-958.

Drgan, V., Zuperl, Š., Vracko, M., Cappelli, C. I., & Novič, M. (2017). CPANNatNIC software for counter-propagation neural network to assist in read-across. *J Cheminform, 9*(1), 30. doi: 10.1186/s13321-017-0218-y

Durrant, J. D., & McCammon, J. A. (2011). Molecular dynamics simulations and drug discovery. *BMC biology, 9*(1), 1-9.

Eckford, P. D., & Sharom, F. J. (2009). ABC efflux pump-based resistance to chemotherapy drugs. *Chemical reviews, 109*(7), 2989-3011.

Ekins, S., Ecker, G., Chiba, P., & Swaan, P. (2007). Future directions for drug transporter modelling. *Xenobiotica, 37*(10-11), 1152-1170.

Ekins, S., Kim, R. B., Leake, B. F., Dantzig, A. H., Schuetz, E. G., Lan, L.-B., . . . Schuetz, J. D. (2002). Three-dimensional quantitative structure-activity relationships of inhibitors of P-glycoprotein. *Molecular pharmacology, 61*(5), 964-973.

Elmeliegy, M., Vourvahis, M., Guo, C., & Wang, D. D. (2020). Effect of P-glycoprotein (P-gp) inducers on exposure of P-gp substrates: Review of clinical drug–drug interaction studies. *Clinical pharmacokinetics, 59*(6), 699-714.

Eyal, S., Lamb, J., Smith-Yockman, M., Yagen, B., Fibach, E., Altschuler, Y., . . . Bialer, M. (2006). The antiepileptic and anticancer agent, valproic acid, induces P-glycoprotein in human tumour cell lines and in rat liver. *British journal of pharmacology, 149*(3), 250-260.

Fardel, O., Lecureur, V., & Guillouzo, A. (1996). The P-glycoprotein multidrug transporter. *General Pharmacology: The Vascular System, 27*(8), 1283-1291.

Favia, A. D. (2011). Theoretical and computational approaches to ligand-based drug discovery. *Frontiers in bioscience (Landmark edition), 16*, 1276-1290.

FDA. (2017). Food and Drug Administration, Center for Drug Evaluation and Research (CDER), U.S. Department of Health and Human Services *Guidance for Industry: Clinical Drug Interaction Studies — Study Design, Data Analysis, and Clinical Implications* (Draft ed.): FDA.

Ferreira, R. J., Ferreira, M.-J. U., & Dos Santos, D. J. (2012). Insights on P-glycoprotein's efflux mechanism obtained by molecular dynamics simulations. *Journal of Chemical Theory and Computation, 8*(6), 1853-1864.

Ferreira, R. J., Ferreira, M.-J. U., & dos Santos, D. J. V. A. (2013a). Assessing the Stabilization of P-Glycoprotein's Nucleotide-Binding Domains by the Linker, Using Molecular Dynamics. *Mol Inform, 32*(5-6), 529-540. doi: 10.1002/minf.201200175

Ferreira, R. J., Ferreira, M.-J. U., & dos Santos, D. J. V. A. (2013b). Molecular Docking Characterizes Substrate-Binding Sites and Efflux Modulation Mechanisms within P-Glycoprotein. *J. Chem. Inf. Model., 53*(7), 1747-1760. doi: 10.1021/ci400195v

Fjodorova, N., Novič, M., Roncaglioni, A., & Benfenati, E. (2011). Evaluating the applicability domain in the case of classification predictive models for carcinogenicity based on the counter propagation artificial neural network. *J. Compu.t Aided Mol. Des., 25*(12), 1147-1158. doi: 10.1007/s10822-011-9499-9

Freeman, J. A., & Skapura, D. M. (1991). *Neural Networks: Algorithms, applications, and programming techniques*. Redwood City, CA, USA: Addison Wesley Longman Publishing Co., Inc.

Fromm, M. (2000). P-glycoprotein: a defense mechanism limiting oral bioavailability and CNS accumulation of drugs. *Int. J. Clin. Pharmacol. Ther, 38*(2), 69-74. doi: 10.5414/CPP38069

Gao, J., Murase, O., Schowen, R. L., Aubé, J., & Borchardt, R. T. (2001). A functional assay for quantitation of the apparent affinities of ligands of P-glycoprotein in Caco-2 cells. *Pharmaceutical research, 18*(2), 171-176.

Geick, A., Eichelbaum, M., & Burk, O. (2001). Nuclear receptor response elements mediate induction of intestinal MDR1 by rifampin. *Journal of Biological Chemistry, 276*(18), 14581-14587.

George, A. M., & Jones, P. M. (2012). Perspectives on the structure–function of ABC transporters: the switch and constant contact models. *Progress in biophysics and molecular biology, 109*(3), 95-107.

Giroud, M., Harder, M., Kuhn, B., Haap, W., Trapp, N., Schweizer, W. B., . . . Diederich, F. (2016). Fluorine Scan of Inhibitors of the Cysteine Protease Human Cathepsin L: Dipolar and Quadrupolar Effects in the π-Stacking of Fluorinated Phenyl Rings on Peptide Amide Bonds. *ChemMedChem, 11*(10), 1042-1047.

Giroud, M., Ivkovic, J., Martignoni, M., Fleuti, M., Trapp, N., Haap, W., . . . Schirmeister, T. (2017). Inhibition of the Cysteine Protease Human Cathepsin L by Triazine Nitriles: Amide··· Heteroarene π-Stacking Interactions and Chalcogen Bonding in the S3 Pocket. *ChemMedChem, 12*(3), 257-270.

Goodwin, B., Hodgson, E., & Liddle, C. (1999). The orphan human pregnane X receptor mediates the transcriptional activation of CYP3A4 by rifampicin through a distal enhancer module. *Molecular pharmacology, 56*(6), 1329-1339.

Greiner, B., Eichelbaum, M., Fritz, P., Kreichgauer, H. P., von Richter, O., Zundler, J., & Kroemer, H. K. (1999). The role of intestinal P-glycoprotein in the interaction of digoxin and rifampin. *J Clin Invest, 104*(2), 147-153. doi: 10.1172/JCI6663

Güner, O. F. (2000). *Pharmacophore perception, development, and use in drug design* (Vol. 2): Internat'l University Line.

Hansson, T., Oostenbrink, C., & van Gunsteren, W. (2002). Molecular dynamics simulations. *Current opinion in structural biology, 12*(2), 190-196.

Harder, M., Kuhn, B., & Diederich, F. (2013). Efficient stacking on protein amide fragments. *ChemMedChem, 8*(3), 397-404.

Héberger, K., & Kollár-Hunek, K. (2011). Sum of ranking differences for method discrimination and its validation: comparison of ranks with random numbers. *Journal of Chemometrics, 25*(4), 151-158. doi: 10.1002/cem.1320

Hediger, M. A., Romero, M. F., Peng, J.-B., Rolfs, A., Takanaga, H., & Bruford, E. A. (2004). The ABCs of solute carriers: physiological, pathological and therapeutic implications of human membrane transport proteins. *Pflügers Archiv, 447*(5), 465-468.

Heller, H., Schaefer, M., & Schulten, K. (1993). Molecular dynamics simulation of a bilayer of 200 lipids in the gel and in the liquid crystal phase. *The Journal of Physical Chemistry, 97*(31), 8343-8360.

Higgins, C. F., & Gottesman, M. M. (1992). Is the multidrug transporter a flippase? *Trends in biochemical sciences, 17*(1), 18-21.

Higgins, C. F., & Linton, K. J. (2004). The ATP switch model for ABC transporters. *Nature structural & molecular biology, 11*(10), 918-926.

Higgins, C. F., Rosenberg, M. F., Callaghan, R., & Ford, R. C. (1997). Structure of the multidrug resistance P-glycoprotein to 2.5 nm resolution determined by electron microscopy and image analysis. *Journal of Biological Chemistry, 272*(16), 10685-10694.

Humphrey, W., Dalke, A., & Schulten, K. (1996). VMD: visual molecular dynamics. *Journal of molecular graphics, 14*(1), 33-38.

Ichiro, N., Kimitoshi, K., Junko, K., Shin-Ichi, A., Akira, K., Ken-Ichi, S., . . . Gottesman, M. M. (1989). Analysis of structural features of dihydropyridine analogs needed to reverse multidrug resistance and to inhibit photoaffinity labeling of P-glycoprotein. *Biochemical Pharmacology, 38*(3), 519-527.

Jain, A. N. (2004). Virtual screening in lead discovery and optimization. *Current opinion in drug discovery & development, 7*(4), 396-403.

Jin, M. S., Oldham, M. L., Zhang, Q., & Chen, J. (2012). Crystal structure of the multidrug transporter P-glycoprotein from Caenorhabditis elegans. *nature, 490*(7421), 566-569.

Jo, S., Cheng, X., Islam, S. M., Huang, L., Rui, H., Zhu, A., . . . Vanommeslaeghe, K. (2014). CHARMM-GUI PDB manipulator for advanced modeling and simulations of proteins containing nonstandard residues. *Advances in protein chemistry and structural biology, 96*, 235-265.

Jo, S., Kim, T., Iyer, V. G., & Im, W. (2008). CHARMM-GUI: a web-based graphical user interface for CHARMM. *Journal of computational chemistry, 29*(11), 1859-1865.

Jones, G., Willett, P., & Glen, R. C. (1995). Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *Journal of molecular biology, 245*(1), 43-53.

Jones, G., Willett, P., Glen, R. C., Leach, A. R., & Taylor, R. (1997). Development and validation of a genetic algorithm for flexible docking. *Journal of molecular biology, 267*(3), 727-748.

Jones, P. M., & George, A. M. (1999). Subunit interactions in ABC transporters: towards a functional architecture. *FEMS microbiology letters, 179*(2), 187-202.

Jones, P. M., & George, A. M. (2002). Mechanism of ABC transporters: a molecular dynamics simulation of a well characterized nucleotide-binding subunit. *Proceedings of the National Academy of Sciences, 99*(20), 12639-12644.

Jones, P. M., & George, A. M. (2009). Opening of the ADP-bound active site in the ABC transporter ATPase dimer: Evidence for a constant contact, alternating sites model for the catalytic cycle. *Proteins: Structure, Function, and Bioinformatics, 75*(2), 387-396.

Jones, P. M., & George, A. M. (2012). Role of the D-loops in allosteric control of ATP hydrolysis in an ABC transporter. *The journal of physical chemistry A, 116*(11), 3004-3013.

Jorgensen, W. L. (2004). The many roles of computation in drug discovery. *Science, 303*(5665), 1813-1818.

Jouan, E., Le Vee, M., Mayati, A., Denizot, C., Parmentier, Y., & Fardel, O. (2016). Evaluation of P-Glycoprotein Inhibitory Potential Using a Rhodamine 123 Accumulation Assay. *Pharmaceutics, 8*(2). doi: 10.3390/pharmaceutics8020012

Juliano, R. L., & Ling, V. (1976). A surface glycoprotein modulating drug permeability in Chinese hamster ovary cell mutants. *Biochimica et Biophysica Acta (BBA)-Biomembranes, 455*(1), 152-162.

Karplus, M., & McCammon, J. A. (2002). Molecular dynamics simulations of biomolecules. *Nature structural biology, 9*(9), 646-652.

Kartner, N., Riordan, & Ling, V. (1983). Cell surface P-glycoprotein associated with multidrug resistance in mammalian cell lines. *Science, 221*(4617), 1285-1288. doi: 10.1126/science.6137059

Kazemi, F., Karimi, I., & Yousofvand, N. (2021). Molecular docking study of lignanamides from Cannabis sativa against P-glycoprotein. *In Silico Pharmacology, 9*(1), 1-7.

Kim, I.-W., Peng, X.-H., Sauna, Z. E., FitzGerald, P. C., Xia, D., Müller, M., . . . Ambudkar, S. V. (2006). The conserved tyrosine residues 401 and 1044 in ATP sites of human P-glycoprotein are critical for ATP binding and hydrolysis: evidence for a conserved subdomain, the A-loop in the ATP-binding cassette. *Biochemistry, 45*(24), 7605-7616.

Kim, K. H. (2001). 3D-QSAR analysis of 2, 4, 5-and 2, 3, 4, 5-substituted imidazoles as potent and nontoxic modulators of P-glycoprotein mediated MDR. *Bioorganic & medicinal chemistry, 9*(6), 1517-1523.

Kim, Y., & Chen, J. (2018). Molecular structure of human P-glycoprotein in the ATP-bound, outward-facing conformation. *Science, 359*(6378), 915-919.

Kirchmair, J., Markt, P., Distinto, S., Wolber, G., & Langer, T. (2008). Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection—what can we learn from earlier mistakes? *J. Compu.t Aided Mol. Des., 22*(3-4), 213-228.

Kitchen, D. B., Decornez, H., Furr, J. R., & Bajorath, J. (2004). Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature reviews Drug discovery, 3*(11), 935-949.

Klebe, G. (2000). Recent developments in structure-based drug design. *Journal of molecular medicine, 78*(5), 269-281.

Klepsch, F., Chiba, P., & Ecker, G. F. (2011). Exhaustive sampling of docking poses reveals binding hypotheses for propafenone type inhibitors of P-glycoprotein. *PLoS Comput Biol, 7*(5), e1002036. doi: 10.1371/journal.pcbi.1002036

Klepsch, F., & Ecker, G. F. (2010). Impact of the recent mouse P-glycoprotein structure for structure-based ligand design. *Mol Inform, 29*(4), 276-286.

Kliewer, S. A., Moore, J. T., Wade, L., Staudinger, J. L., Watson, M. A., Jones, S. A., . . . Zetterström, R. H. (1998). An orphan nuclear receptor activated by pregnanes defines a novel steroid signaling pathway. *Cell, 92*(1), 73-82.

Kodan, A., Yamaguchi, T., Nakatsu, T., Sakiyama, K., Hipolito, C. J., Fujioka, A., . . . Hiratake, J. (2014). Structural basis for gating mechanisms of a eukaryotic P-glycoprotein homolog. *Proceedings of the National Academy of Sciences, 111*(11), 4049-4054.

Krivov, G. G., Shapovalov, M. V., & Dunbrack Jr, R. L. (2009). Improved prediction of protein side-chain conformations with SCWRL4. *Proteins: Structure, Function, and Bioinformatics, 77*(4), 778-795.

Lacher, S. E., Gremaud, J. N., Skagen, K., Steed, E., Dalton, R., Sugden, K. D., . . . Woodahl, E. L. (2014). Absence of P-glycoprotein transport in the pharmacokinetics and toxicity of the herbicide paraquat. *Journal of Pharmacology and Experimental Therapeutics, 348*(2), 336-345.

Laskowski, R. A., MacArthur, M. W., Moss, D. S., & Thornton, J. M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of applied crystallography, 26*(2), 283-291.

Leardi, R. (2003). *Nature-inspired methods in chemometrics: genetic algorithms and artificial neural networks*. Amsterdam: Elsevier Science.

Leitner, I., Nemeth, J., Feurstein, T., Abrahim, A., Matzneller, P., Lagler, H., . . . Zeitlinger, M. (2011). The third-generation P-glycoprotein inhibitor tariquidar may overcome bacterial multidrug resistance by increasing intracellular drug concentration. *Journal of Antimicrobial Chemotherapy, 66*(4), 834-839.

Leslie, E. M., Deeley, R. G., & Cole, S. P. (2005). Multidrug resistance proteins: role of P-glycoprotein, MRP1, MRP2, and BCRP (ABCG2) in tissue defense. *Toxicol. Appl. Pharmacol., 204*(3), 216-237. doi: 10.1016/j.taap.2004.10.012

Li, D., Chen, L., Li, Y., Tian, S., Sun, H., & Hou, T. (2014). ADMET evaluation in drug discovery. 13. Development of in silico prediction models for P-glycoprotein substrates. *Mol. Pharm., 11*(3), 716-726. doi: 10.1021/mp400450m

Li, Y., & Zhang, Y. (2009). REMO: A new protocol to refine full atomic protein models from C-alpha traces by optimizing hydrogen-bonding networks. *Proteins: Structure, Function, and Bioinformatics, 76*(3), 665-676.

Lindorff-Larsen, K., Maragakis, P., Piana, S., Eastwood, M. P., Dror, R. O., & Shaw, D. E. (2012). Systematic validation of protein force fields against experimental data. *PloS one, 7*(2), e32131.

Litman, T., Zeuthen, T., Skovsgaard, T., & Stein, W. D. (1997). Competitive, non-competitive and cooperative interactions between substrates of P-glycoprotein as

measured by its ATPase activity. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease, 1361*(2), 169-176.

Lockhart, A. C., Tirona, R. G., & Kim, R. B. (2003). Pharmacogenetics of ATP-binding Cassette Transporters in Cancer and Chemotherapy1. *Mol. Cancer Ther., 2*(7), 685-698.

Loo, T. W., Bartlett, M. C., & Clarke, D. M. (2003a). Simultaneous binding of two different drugs in the binding pocket of the human multidrug resistance P-glycoprotein. *Journal of Biological Chemistry, 278*(41), 39706-39710.

Loo, T. W., Bartlett, M. C., & Clarke, D. M. (2003b). Substrate-induced conformational changes in the transmembrane segments of human P-glycoprotein Direct evidence for the substrate-induced fit mechanism for drug binding. *Journal of Biological Chemistry, 278*(16), 13603-13606.

Loo, T. W., Bartlett, M. C., & Clarke, D. M. (2006). Transmembrane segment 7 of human P-glycoprotein forms part of the drug-binding pocket. *Biochemical Journal, 399*(2), 351-359.

Loo, T. W., & Clarke, D. M. (1997). Identification of residues in the drug-binding site of human P-glycoprotein using a thiol-reactive substrate. *Journal of Biological Chemistry, 272*(51), 31945-31948.

Loo, T. W., & Clarke, D. M. (2001). Defining the drug-binding site in the human multidrug resistance P-glycoprotein using a methanethiosulfonate analog of verapamil, MTS-verapamil. *J Biol Chem, 276*(18), 14972-14979. doi: 10.1074/jbc.M100407200

Loo, T. W., & Clarke, D. M. (2002). Location of the rhodamine-binding site in the human multidrug resistance P-glycoprotein. *J Biol Chem, 277*(46), 44332-44338. doi: 10.1074/jbc.M208433200

Lugo, M. R., & Sharom, F. J. (2005). Interaction of LDS-751 and rhodamine 123 with P-glycoprotein: evidence for simultaneous binding of both drugs. *Biochemistry, 44*(42), 14020-14029.

Lum, B. L., Gosland, M. P., Kaubisch, S., & Sikic, B. I. (1993). Molecular targets in oncology: implications of the multidrug resistance gene. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy, 13*(2), 88-109.

Ma, X. H., Jia, J., Zhu, F., Xue, Y., Li, Z. R., & Chen, Y. Z. (2009). Comparative analysis of machine learning methods in ligand-based virtual screening of large compound libraries. *Combinatorial chemistry & high throughput screening, 12*(4), 344-357.

Martin, C., Berridge, G., Mistry, P., Higgins, C., Charlton, P., & Callaghan, R. (2000). Drug binding sites on P-glycoprotein are altered by ATP binding prior to nucleotide hydrolysis. *Biochemistry, 39*(39), 11901-11906.

Martin, C., Higgins, C. F., & Callaghan, R. (2001). The vinblastine binding site adopts high-and low-affinity conformations during a transport cycle of P-glycoprotein. *Biochemistry, 40*(51), 15733-15742.

Marzolini, C., Paus, E., Buclin, T., & Kim, R. B. (2004). Polymorphisms in human MDR1 (P-glycoprotein): recent advances and clinical relevance. *Clinical Pharmacology & Therapeutics, 75*(1), 13-33.

Mauri, A., Consonni, V., Pavan, M., & Todeschini, R. (2017). Dragon (software for molecular descriptor calculation) (Version 7.0.8): Kode srl. Retrieved from https://chm.kode-solutions.net

Mavromoustakos, T., Durdagi, S., Koukoulitsa, C., Simcic, M., G Papadopoulos, M., Hodoscek, M., & Golic Grdadolnik, S. (2011). Strategies in the rational drug design. *Current medicinal chemistry, 18*(17), 2517-2530.

McCammon, J. A., Gelin, B. R., & Karplus, M. (1977). Dynamics of folded proteins. *nature, 267*(5612), 585-590.

McGibbon, R. T., Beauchamp, K. A., Harrigan, M. P., Klein, C., Swails, J. M., Hernández, C. X., . . . Pande, V. S. (2015). MDTraj: a modern open library for the analysis of molecular dynamics trajectories. *Biophysical journal, 109*(8), 1528-1532.

Miller III, B. R., McGee Jr, T. D., Swails, J. M., Homeyer, N., Gohlke, H., & Roitberg, A. E. (2012). MMPBSA. py: an efficient program for end-state free energy calculations. *Journal of Chemical Theory and Computation, 8*(9), 3314-3321.

Montanari, F., & Ecker, G. F. (2015). Prediction of drug-ABC-transporter interaction--Recent advances and future challenges. *Adv. Drug Deliv. Rev., 86*, 17-26. doi: 10.1016/j.addr.2015.03.001

Mora Lagares, L., Minovski, N., Caballero Alfonso, A. Y., Benfenati, E., Wellens, S., Culot, M., . . . Novič, M. (2020). Homology modeling of the human P-glycoprotein (ABCB1) and insights into ligand binding through molecular docking studies. *International journal of molecular sciences, 21*(11), 4058. doi: 10.3390/ijms21114058

Mora Lagares, L., Minovski, N., Drgan, V., Marjan, T., & Novič, M. (2019). *P-gp transport activity in connection to the efflux of toxicants or drugs.* Paper presented at the Conferentia Chemometrica 2019, Karcag, Hungary.

Mora Lagares, L., Minovski, N., & Novič, M. (2018, 11-15th June 2018). *P-glycoprotein modelling : development of an in silico prediction model for substrates, inhibitors and non-interacting compounds.* Paper presented at the International Conference on QSAR in Environmental and Health Sciences, Bled, Slovenia.

Mora Lagares, L., Minovski, N., & Novič, M. (2019). Multiclass Classifier for P-Glycoprotein Substrates, Inhibitors, and Non-Active Compounds. *Molecules, 24*(10). doi: 10.3390/molecules24102006

Mordalski, S., Witek, J., Smusz, S., Rataj, K., & Bojarski, A. J. (2015). Multiple conformational states in retrospective virtual screening–homology models vs. crystal structures: beta-2 adrenergic receptor case study. *J Cheminform, 7*(1), 13.

Moro, S., Bacilieri, M., & Deflorian, F. (2007). Combining ligand-based and structure-based drug design in the virtual screening arena. *Expert opinion on drug discovery, 2*(1), 37-49.

Morris, G. M., & Lim-Wilby, M. (2008). Molecular docking *Molecular modeling of proteins* (pp. 365-382): Springer.

Morris, J. H., Huang, C. C., Babbitt, P. C., & Ferrin, T. E. (2007). structureViz: linking Cytoscape and UCSF Chimera. *Bioinformatics, 23*(17), 2345-2347.

Motherwell, W. B., Moreno, R. B., Pavlakos, I., Arendorf, J. R., Arif, T., Tizzard, G. J., . . . Aliev, A. E. (2018). Noncovalent interactions of π systems with sulfur: The atomic chameleon of molecular recognition. *Angewandte Chemie, 130*(5), 1207-1212.

Myint, K. Z., & Xie, X.-Q. (2010). Recent advances in fragment-based QSAR and multi-dimensional QSAR methods. *International journal of molecular sciences, 11*(10), 3846-3866.

Nicklisch, S. C., Rees, S. D., McGrath, A. P., Gökirmak, T., Bonito, L. T., Vermeer, L. M., . . . Chang, G. (2016). Global marine pollutants inhibit P-glycoprotein: Environmental levels, inhibitory effects, and cocrystal structure. *Science advances, 2*(4), e1600001.

Nobili, S., Landini, I., Mazzei, T., & Mini, E. (2012). Overcoming tumor multidrug resistance using drugs able to evade P-glycoprotein or to exploit its expression. *Medicinal research reviews, 32*(6), 1220-1262.

O'Mara, M. L., & Tieleman, D. P. (2007). P-glycoprotein models of the apo and ATP-bound states based on homology with Sav1866 and MalK. *FEBS letters, 581*(22), 4217-4222.

O'Mara, M. L., & Mark, A. E. (2012). The effect of environment on the structure of a membrane protein: P-glycoprotein under physiological conditions. *Journal of Chemical Theory and Computation, 8*(10), 3964-3976.

OECD. (2007). OECD Environment Health and Safety Publications, Series on Testing and Assessment No 69. Retrieved 23 May 2019 https://read.oecd-ilibrary.org/environment/guidance-document-on-the-validation-of-quantitative-structure-activity-relationship-q-sar-models_9789264085442-en#page1

Ou-Yang, S.-s., Lu, J.-y., Kong, X.-q., Liang, Z.-j., Luo, C., & Jiang, H. (2012). Computational drug discovery. *Acta Pharmacologica Sinica, 33*(9), 1131-1140.

Pajeva, I. K., & Wiese, M. (2002). Pharmacophore model of drugs involved in P-glycoprotein multidrug resistance: explanation of structural variety (hypothesis). *J. Med. Chem., 45*(26), 5671-5686.

Palmeira, A., Rodrigues, F., Sousa, E., Pinto, M., Vasconcelos, M. H., & Fernandes, M. X. (2011). New uses for old drugs: pharmacophore-based screening for the discovery of P-glycoprotein inhibitors. *Chem. Biol. Drug Des., 78*(1), 57-72. doi: 10.1111/j.1747-0285.2011.01089.x

Palmeira, A., Sousa, E., Vasconcelos, M. H., & Pinto, M. (2012). Three decades of P-gp inhibitors: skimming through several generations and scaffolds. *Current medicinal chemistry, 19*(13), 1946-2025.

Perkins, R., Fang, H., Tong, W., & Welsh, W. J. (2003). Quantitative structure-activity relationship methods: Perspectives on drug discovery and toxicology. *Environmental Toxicology and Chemistry: An International Journal, 22*(8), 1666-1679.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). UCSF Chimera—a visualization system for exploratory research and analysis. *Journal of computational chemistry, 25*(13), 1605-1612.

Pleban, K., Kopp, S., Csaszar, E., Peer, M., Hrebicek, T., Rizzi, A., . . . Chiba, P. (2005). P-glycoprotein substrate binding domains are located at the transmembrane domain/transmembrane domain interfaces: a combined photoaffinity labeling-protein homology modeling approach. *Molecular pharmacology, 67*(2), 365-374.

Polli, J. W., Wring, S. A., Humphreys, J. E., Huang, L., Morgan, J. B., Webster, L. O., & Serabjit-Singh, C. S. (2001). Rational use of in vitro P-glycoprotein assays in drug discovery. *Journal of Pharmacology and Experimental Therapeutics, 299*(2), 620-628.

Pontius, J., Richelle, J., & Wodak, S. J. (1996). Deviations from standard atomic volumes as a quality measure for protein crystal structures. *Journal of molecular biology, 264*(1), 121-136.

Prathipati, P., Dixit, A., & Saxena, A. K. (2007). Computer-aided drug design: integration of structure-based and ligand-based approaches in drug design. *Current Computer-Aided Drug Design, 3*(2), 133-148.

Qu, Q., Russell, P. L., & Sharom, F. J. (2003). Stoichiometry and affinity of nucleotide binding to P-glycoprotein during the catalytic cycle. *Biochemistry, 42*(4), 1170-1177.

Ramu, A., & Ramu, N. (1992). Reversal of multidrug resistance by phenothiazines and structurally related compounds. *Cancer Chemotherapy and Pharmacology, 30*(3), 165-173. doi: 10.1007/bf00686306

Ravna, A. W., Sylte, I., & Sager, G. (2007). Molecular model of the outward facing state of the human P-glycoprotein (ABCB1), and comparison to a model of the human MRP5 (ABCC5). *Theoretical Biology and Medical Modelling, 4*(1), 33.

Riordan, J. R., Deuchars, K., Kartner, N., Alon, N., Trent, J., & Ling, V. (1985). Amplification of P-glycoprotein genes in multidrug-resistant mammalian cell lines. *nature, 316*(6031), 817-819.

Roe, D. R., & Cheatham III, T. E. (2013). PTRAJ and CPPTRAJ: software for processing and analysis of molecular dynamics trajectory data. *Journal of Chemical Theory and Computation, 9*(7), 3084-3095.

Romsicki, Y., & Sharom, F. J. (1998). The ATPase and ATP-binding functions of P-glycoprotein: Modulation by interaction with defined phospholipids. *European Journal of Biochemistry, 256*(1), 170-178.

Rosenberg, M. F., Velarde, G., Ford, R. C., Martin, C., Berridge, G., Kerr, I. D., . . . Linton, K. J. (2001). Repacking of the transmembrane domains of P-glycoprotein during the transport ATPase cycle. *The EMBO Journal, 20*(20), 5615-5625.

Roy, A., Kucukural, A., & Zhang, Y. (2010). I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc, 5*(4), 725-738. doi: 10.1038/nprot.2010.5

Saeki, T., Ueda, K., Tanigawara, Y., Hori, R., & Komano, T. (1993). Human P-glycoprotein transports cyclosporin A and FK506. *Journal of Biological Chemistry, 268*(9), 6077-6080.

Safa, A. R. (2004). Identification and characterization of the binding sites of P-glycoprotein for multidrug resistance-related drugs and modulators. *Current Medicinal Chemistry-Anti-Cancer Agents, 4*(1), 1-17.

Šali, A., & Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *Journal of molecular biology, 234*(3), 779-815.

Salonen, L. M., Bucher, C., Banner, D. W., Haap, W., Mary, J. L., Benz, J., . . . Diederich, F. (2009). Cation–π interactions at the active site of factor Xa: dramatic enhancement upon stepwise N-alkylation of ammonium ions. *Angewandte Chemie International Edition, 48*(4), 811-814.

Sauna, Z. E., & Ambudkar, S. V. (2007). About a switch: how P-glycoprotein (ABCB1) harnesses the energy of ATP binding and hydrolysis to do mechanical work. *Mol. Cancer Ther., 6*(1), 13-23. doi: 10.1158/1535-7163.MCT-06-0155

Scapin, G. (2006). Structural biology and drug discovery. *Current pharmaceutical design, 12*(17), 2087-2097.

Schmid, D., Ecker, G., Kopp, S., Hitzler, M., & Chiba, P. (1999). Structure–activity relationship studies of propafenone analogs based on P-glycoprotein ATPase activity measurements. *Biochemical Pharmacology, 58*(9), 1447-1456.

Schwede, T., Kopp, J., Guex, N., & Peitsch, M. C. (2003). SWISS-MODEL: an automated protein homology-modeling server. *Nucleic acids research, 31*(13), 3381-3385.

Seelig, A. (1998). A general pattern for substrate recognition by P-glycoprotein. *European Journal of Biochemistry, 251*(1-2), 252-261.

Seelig, A., & Landwojtowicz, E. (2000). Structure–activity relationship of P-glycoprotein substrates and modifiers. *European journal of pharmaceutical sciences, 12*(1), 31-40.

Seelig, A., Landwojtowicz, E., Fischer, H., & Li Blatter, X. (2004). Towards P-glycoprotein structure–activity relationships. In H. van de Waterbeemd, H. Lennernäs & P. Artursson (Eds.), *Drug Bioavailability: Estimation of solubility, Permeability, Absorption and Bioavailability* (pp. 461–492). Weinheim, Germany: Wiley-VCH Verlag, GmbH & Co. KGaA.

Sekhar, A., Vallurupalli, P., & Kay, L. E. (2013). Defining a length scale for millisecond-timescale protein conformational exchange. *Proceedings of the National Academy of Sciences, 110*(28), 11391-11396.

Senior, A. E., Al-Shawi, M. K., & Urbatsch, I. L. (1995). The catalytic cycle of P-glycoprotein. *FEBS letters, 377*(3), 285-289.

Sevin, E., Dehouck, L., Versele, R., Culot, M., & Gosselet, F. (2019). A Miniaturized Pump Out Method for Characterizing Molecule Interaction with ABC Transporters. *International journal of molecular sciences, 20*(22), 5529.

Shao, J., Tanner, S. W., Thompson, N., & Cheatham, T. E. (2007). Clustering molecular dynamics trajectories: 1. Characterizing the performance of different clustering algorithms. *Journal of Chemical Theory and Computation, 3*(6), 2312-2334.

Sharom, F. J. (2008). ABC multidrug transporters- structure, function and role in chemoresistance. *Pharmacogenomics, 9*(1), 105-127. doi: 10.2217/14622416.9.1.105

Sharom, F. J. (2014). Complex interplay between the P-glycoprotein multidrug efflux pump and the membrane: its role in modulating protein function. *Frontiers in oncology, 4*, 41.

Shen, M. Y., & Sali, A. (2006). Statistical potential for assessment and prediction of protein structures. *Protein Sci, 15*(11), 2507-2524. doi: 10.1110/ps.062416606

Sirisha, K., Shekhar, M. C., Umasankar, K., Mahendar, P., Sadanandam, A., Achaiah, G., & Reddy, V. M. (2011). Molecular docking studies and in vitro screening of new dihydropyridine derivatives as human MRP1 inhibitors. *Bioorganic & medicinal chemistry, 19*(10), 3249-3254.

Śledź, P., & Caflisch, A. (2018). Protein structure-based drug design: from docking to molecular dynamics. *Current opinion in structural biology, 48*, 93-102.

Smith, P. C., Karpowich, N., Millen, L., Moody, J. E., Rosen, J., Thomas, P. J., & Hunt, J. F. (2002). ATP binding to the motor domain from an ABC transporter drives formation of a nucleotide sandwich dimer. *Molecular cell, 10*(1), 139-149.

Sparreboom, A., Danesi, R., Ando, Y., Chan, J., & Figg, W. D. (2003). Pharmacogenomics of ABC transporters and its role in cancer chemotherapy. *Drug Resistance Updates, 6*(2), 71-84.

Srinivasan, J., Cheatham, T. E., Cieplak, P., Kollman, P. A., & Case, D. A. (1998). Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate− DNA helices. *Journal of the American Chemical Society, 120*(37), 9401-9409.

Stahura, F. L., & Bajorath, J. (2005). New methodologies for ligand-based virtual screening. *Current pharmaceutical design, 11*(9), 1189-1202.

Storm, J., O'Mara, M. L., Crowley, E. H., Peall, J., Tieleman, D. P., Kerr, I. D., & Callaghan, R. (2007). Residue G346 in transmembrane segment six is involved in inter-domain communication in P-glycoprotein. *Biochemistry, 46*(35), 9899-9910.

Taipalensuu, J., Tavelin, S., Lazorova, L., Svensson, A.-C., & Artursson, P. (2004). Exploring the quantitative relationship between the level of MDR1 transcript, protein and function using digoxin as a marker of MDR1-dependent drug efflux activity. *European journal of pharmaceutical sciences, 21*(1), 69-75.

Takara, K., Tanigawara, Y., Komada, F., Nishiguchi, K., Sakaeda, T., & Okumura, K. (1999). Cellular pharmacokinetic aspects of reversal effect of itraconazole on P-glycoprotein-mediated resistance of anticancer drugs. *Biological and Pharmaceutical Bulletin, 22*(12), 1355-1359.

Tandon, V. R., Kapoor, B., Bano, G., Gupta, S., Gillani, Z., & Kour, D. (2006). P-glycoprotein: Pharmacological relevance. *Indian journal of pharmacology, 38*(1), 13.

Terwogt, J. M. M., Malingré, M. M., Beijnen, J. H., Wim, W., Rosing, H., Koopman, F. J., . . . Schellens, J. H. (1999). Coadministration of oral cyclosporin A enables oral therapy with paclitaxel. *Clinical Cancer Research, 5*(11), 3379-3384.

Thiebaut, F., Tsuruo, T., Hamada, H., Gottesman, M. M., Pastan, I., & Willingham, M. C. (1987). Cellular localization of the multidrug-resistance gene product P-glycoprotein in normal human tissues. *Proceedings of the National Academy of Sciences, 84*(21), 7735-7738.

Tsuruo, T., Iida, H., Tsukagoshi, S., & Sakurai, Y. (1981). Overcoming of vincristine resistance in P388 leukemia in vivo and in vitro through enhanced cytotoxicity of vincristine and vinblastine by verapamil. *Cancer research, 41*(5), 1967-1972.

UCLA-DOE, I. Structural Analysis and Verification Server SAVEs v5.0. Retrieved June 2019, 2019, from https://servicesn.mbi.ucla.edu/SAVES/

Vacirca, D., Barbati, C., Scazzocchio, B., Masella, R., Rosano, G., Malorni, W., & Ortona, E. (2011). Anti-ATP synthase autoantibodies from patients with Alzheimer's disease reduce extracellular HDL level. *Journal of Alzheimer's Disease, 26*(3), 441-445.

Van De Waterbeemd, H., & Gifford, E. (2003). ADMET in silico modelling: towards prediction paradise? *Nature reviews Drug discovery, 2*(3), 192-204.

Vasiliou, V., Vasiliou, K., & Nebert, D. W. (2009). Human ATP-binding cassette (ABC) transporter family. *Human genomics, 3*(3), 281-290. doi: 10.1186/1479-7364-3-3-281

Vedani, A., Briem, H., Dobler, M., Dollinger, H., & McMasters, D. R. (2000). Multiple-conformation and protonation-state representation in 4D-QSAR: the neurokinin-1 receptor system. *J. Med. Chem., 43*(23), 4416-4427.

Vedani, A., & Dobler, M. (2002). 5D-QSAR: the key for simulating induced fit? *J. Med. Chem., 45*(11), 2139-2149.

Verdonk, M. L., Cole, J. C., Hartshorn, M. J., Murray, C. W., & Taylor, R. D. (2003). Improved protein–ligand docking using GOLD. *Proteins: Structure, Function, and Bioinformatics, 52*(4), 609-623.

Verlinde, C. L., & Hol, W. G. (1994). Structure-based drug design: progress, results and challenges. *Structure, 2*(7), 577-587.

Vrbanac, J., & Slauter, R. (2017). ADME in drug discovery *A Comprehensive Guide to Toxicology in Nonclinical Drug Development* (pp. 39-67): Elsevier.

Wagner, J. R., Sørensen, J., Hensley, N., Wong, C., Zhu, C., Perison, T., & Amaro, R. E. (2017). POVME 3.0: software for mapping binding pocket flexibility. *Journal of Chemical Theory and Computation, 13*(9), 4584-4592.

Wandel, C., Kim, R., Kajiji, S., Guengerich, F. P., Wilkinson, G. R., & Wood, A. (1999). P-glycoprotein and cytochrome P-450 3A inhibition: dissociation of inhibitory potencies. *Cancer research, 59*(16), 3944-3948.

Wandel, C., Kim, R., Wood, M., & Wood, A. (2002). Interaction of Morphine, Fentanyl, Sufentanil, Alfentanil, and Loperamide with the Efflux Drug Transporter P-glycoprotein. *Anesthesiology: The Journal of the American Society of Anesthesiologists, 96*(4), 913-920.

Wang, Y.-M., Ong, S. S., Chai, S. C., & Chen, T. (2012). Role of CAR and PXR in xenobiotic sensing and metabolism. *Expert opinion on drug metabolism & toxicology, 8*(7), 803-817.

Wang, Z., Chen, Y., Liang, H., Bender, A., Glen, R. C., & Yan, A. (2011). P-glycoprotein substrate models using support vector machines based on a comprehensive data set. *J. Chem. Inf. Model., 51*(6), 1447-1456. doi: 10.1021/ci2001583

Weiser, J., Shenkin, P. S., & Still, W. C. (1999). Approximate atomic surfaces from linear combinations of pairwise overlaps (LCPO). *Journal of computational chemistry, 20*(2), 217-230.

Weiss, J., Kerpen, C. J., Lindenmaier, H., Dormann, S.-M. G., & Haefeli, W. E. (2003). Interaction of antiepileptic drugs with human P-glycoprotein in vitro. *Journal of Pharmacology and Experimental Therapeutics, 307*(1), 262-267.

Wessler, J. D., Grip, L. T., Mendell, J., & Giugliano, R. P. (2013). The P-glycoprotein transport system and cardiovascular drugs. *Journal of the American College of Cardiology, 61*(25), 2495-2502.

Wiese, M., & Pajeva, I. (1997). Molecular modeling study of the multidrug resistance modifiers cis-and trans-flupentixol. *Die Pharmazie, 52*(9), 679-685.

Wiese, M., & Pajeva, I. K. (2001). Structure-activity relationships of multidrug resistance reversers. *Current medicinal chemistry, 8*(6), 685-713.

Wigler, P. W. (1999). PSC833, cyclosporinA, and dexniguldipine effects on cellular calcein retention and inhibition of the multidrug resistance pump in human leukemic lymphoblasts. *Biochemical and biophysical research communications, 257*(2), 410-413.

Winter, S. S., Lovato, D. M., Khawaja, H. M., Edwards, B. S., Steele, I. D., Young, S. M., . . . Larson, R. S. (2008). High-throughput screening for daunorubicin-mediated drug resistance identifies mometasone furoate as a novel ABCB1-reversal agent. *Journal of biomolecular screening, 13*(3), 185-193.

Wongrattanakamon, P., Lee, V. S., Nimmanpipug, P., Sirithunyalug, B., Chansakaow, S., & Jiranusornkul, S. (2017). Insight into the molecular mechanism of P-glycoprotein mediated drug toxicity induced by bioflavonoids: an integrated computational approach. *Toxicol. Mech. Methods, 27*(4), 253-271. doi: 10.1080/15376516.2016.1273428

Wu, B., Li, H. X., Lian, J., Guo, Y. J., Tang, Y. H., Chang, Z. J., . . . Lu, Z. Q. (2019). Nrf2 overexpression protects against paraquat-induced A549 cell injury primarily by upregulating P-glycoprotein and reducing intracellular paraquat accumulation. *Experimental and therapeutic medicine, 17*(2), 1240-1247.

Wu, G., Robertson, D. H., Brooks III, C. L., & Vieth, M. (2003). Detailed analysis of grid-based molecular docking: A case study of CDOCKER—A CHARMm-based MD docking algorithm. *Journal of computational chemistry, 24*(13), 1549-1562.

Wu, S., & Zhang, Y. (2007). LOMETS: a local meta-threading-server for protein structure prediction. *Nucleic Acids Res, 35*(10), 3375-3382. doi: 10.1093/nar/gkm251

Xiang, M., Cao, Y., Fan, W., Chen, L., & Mo, Y. (2012). Computer-aided drug design: lead discovery and optimization. *Combinatorial chemistry & high throughput screening, 15*(4), 328-337.

Xu, J., Jiao, F., & Berger, B. (2005). *A tree-decomposition approach to protein structure prediction.* Paper presented at the 2005 IEEE Computational Systems Bioinformatics Conference (CSB'05).

Yang, J.-M., Chen, Y.-F., Shen, T.-W., Kristal, B. S., & Hsu, D. F. (2005). Consensus scoring criteria for improving enrichment in virtual screening. *J. Chem. Inf. Model., 45*(4), 1134-1146.

Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., & Zhang, Y. (2015). The I-TASSER Suite: protein structure and function prediction. *Nat Methods, 12*(1), 7-8. doi: 10.1038/nmeth.3213

Ye, J., Osborne, A. R., Groll, M., & Rapoport, T. A. (2004). RecA-like motor ATPases—lessons from structures. *Biochimica et Biophysica Acta (BBA)-Bioenergetics, 1659*(1), 1-18.

Zaitseva, J., Jenewein, S., Wiedenmann, A., Benabdelhak, H., Holland, I. B., & Schmitt, L. (2005). Functional characterization and ATP-induced dimerization of the isolated ABC-domain of the haemolysin B transporter. *Biochemistry, 44*(28), 9680-9690.

Zhang, B., Kang, Z., Zhang, J., Kang, Y., Liang, L., Liu, Y., & Wang, Q. (2021). Simultaneous binding mechanism of multiple substrates for multidrug resistance transporter P-glycoprotein. *Physical Chemistry Chemical Physics, 23*(8), 4530-4543.

Zhang, L., & Hermans, J. (1996). Hydrophilicity of cavities in proteins. *Proteins: Structure, Function, and Bioinformatics, 24*(4), 433-438.

Zhang, Y. (2008). I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics, 9*(1). doi: 10.1186/1471-2105-9-40

Zhang, Y., & Skolnick, J. (2004a). Scoring function for automated assessment of protein structure template quality. *Proteins, 57*(4), 702-710. doi: 10.1002/prot.20264

Zhang, Y., & Skolnick, J. (2004b). SPICKER: a clustering approach to identify near-native protein folds. *Journal of computational chemistry, 25*(6), 865-871.

Zhao, H., & Caflisch, A. (2015). Molecular dynamics in drug design. *European journal of medicinal chemistry, 91*, 4-14.

Zheng, W., Zhang, C., Wuyun, Q., Pearce, R., Li, Y., & Zhang, Y. (2019). LOMETS2: improved meta-threading server for fold-recognition and structure-based function annotation for distant-homology proteins. *Nucleic Acids Res, 47*(W1), W429-W436. doi: 10.1093/nar/gkz384

Zhou, S.-F. (2008). Structure, function and regulation of P-glycoprotein and its clinical relevance in drug disposition. *Xenobiotica, 38*(7-8), 802-832.

Zupan, J., & Gasteiger, J. (1993). *Neural networks for chemists: an introduction*. New York: John Wiley & Sons, Inc.

Bibliography

Publications Related to the Thesis

Journal Articles

Mora Lagares, L., Minovski, N., Caballero Alfonso, A. Y., Benfenati, E., Wellens, S., Culot, M., . . . Novič, M. (2020). Homology modeling of the human P-glycoprotein (ABCB1) and insights into ligand binding through molecular docking studies. *International journal of molecular sciences, 21*(11), 4058. doi: 10.3390/ijms21114058

Mora Lagares, L., Minovski, N., & Novic, M. (2019). Multiclass Classifier for P-Glycoprotein Substrates, Inhibitors, and Non-Active Compounds. *Molecules, 24*(10). doi: 10.3390/molecules24102006

Conference Paper

Mora Lagares, L. M., Nikola; Drgan, Viktor; Tušar, Marjan; Novič, Marjana. (2019). *P-gp transport activity in connection to the efflux of toxicants or drugs.* Paper presented at the Conferentia Chemometrica 2019, Karcag, Hungary.

Mora Lagares, L. M., Nikola; Novič, Marjana. (2018, 11-15th June 2018). *P-glycoprotein modelling: development of an in silico prediction model for substrates, inhibitors and non-interacting compounds.* Paper presented at the International Conference on QSAR in Environmental and Health Sciences, Bled, Slovenia.

Biography

Liadys Mora Lagares is a Computational Chemist currently working in the Theory Department, Laboratory of Cheminformatics at the National Institute of Chemistry in Ljubljana, Slovenia. Her research interests are in the areas of molecular modelling, computer-aided drug design, and computational toxicology. In 2009, she obtained a bachelor's degree in Pharmaceutical Chemistry from the University of Atlántico (Colombia). Later, in 2015, she completed a European Master double degree, the Erasmus Mundus Joint Master's degree (EMJMD) in Theoretical Chemistry and Computational Modelling (TCCM) at the University of Perugia (Italy) and Autonomous University of Madrid (Spain).

In 2017, she joined the laboratory of Prof. Marjana Novič at the National Institute of Chemistry to start her PhD studies within the Marie Skłodowska-Curie Action - Innovative Training Network (MSCA-ITN) project «Integrated *in vitro* and *in silico* tools» (in3). Her PhD work deals with molecular modelling, focusing on the P-glycoprotein, a membrane protein important for the evaluation of pharmacokinetics and toxicity of new drugs. She combined molecular modelling and *in silico* QSAR models to support the integrated interdisciplinary approach to non-animal chemical safety assessment.

During her PhD, she published her work in several international journals, with 2 publications as a first author. She also presented her research in several international conferences, e.g., 20th International Congress on In Vitro Toxicology ESTIV 2018, 21st European Congress on Alternatives to Animal Testing EUSAAT 2018, and EUROTOX Congress 2019. She was awarded a 2-year Erasmus Mundus Scholarship and 3-year Marie Curie ITN fellowship. She is currently a member of the European Society of Toxicology In Vitro (ESTIV).