DIPLOMARBEIT / DIPLOMA THESIS

Titel der Diplomarbeit / Title of the Diploma Thesis

„Development of *in silico* models for profiling the selectivity of tariquidar analogs as ligands of p-glycoprotein and breast cancer resistance protein"

verfasst von / submitted by

Priska Litasha Hiswara

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Magistra der Pharmazie (Mag.pharm.)

Wien, 2016 / Vienna, 2016

Acknowledgments

Foremost, I would like to express the deepest appreciation to Univ. Prof. Dr. Gerhard Ecker for giving me the opportunity to join his outstanding research group in the pharmacoinformatic area, and for his guidance, patience, and providing me with an excellent atmosphere for conducting my studies.

My sincere thanks also goes to Floriane Montanari for helping me from the very beginning of the study of my thesis. Thank you for sharing your expertise with me, and for your help, encouragement, understanding and support both in the lab and also outside the lab.

I would also like to thank the rest of the Pharmacoinformatics group "Pharminfomaniacs" for their encouragement, new ideas, and of course unforgettable time together. Thank you, Daria, Doris, Daniela, Lars, Amir, Sankalp, Eva, Melanie, Chonticha, Andreas, Eleni, Michael, Barbara, Taymara, Dennis, Bernhard, Lydia, Marian, Anika, Kathi, Dimitris and all others I forgot to mention!

Special thanks goes to Lydia, Ozotu, Anabela, Jelena and Elisa. Also many thanks to my colleagues and supervisors at the IAEA, especially Laurent H. and Jeannie J. Thank you everyone for your support and understanding. My deepest thanks goes to my *supersayang*. Thank you for your support and sticking by my side through thick and thin, Sebi!

Last but not least, I would like to thank my family: my parents, my older brother, Gunar, and his wife, Nadhila, for supporting me spiritually throughout writing this thesis and my life in general.

Terima kasih!

Table of Contents

List of Figures

List of Tables

For my beloved parents, Eri Hiswara and Suzie Darmawati.

1. Introduction

1.1 General Information – Human ATP-binding Cassette (ABC) Transporters

The ATP-binding cassette (ABC) transporter superfamily of genes is one of the most important membrane transport protein groups, whose main role is to transport diverse substances, including nutritional elements, inorganic ions, metabolites, etc. across membranes. As the name suggests, ABC transporters bind and hydrolyse ATP at the nucleotide-binding domains (NBDs) and utilize energy from it to translocate a large number of molecules across extra- and intracellular membranes. To date, there are 49 known human ABC transporters, which are categorized into seven subfamilies, identified as ABCA through ABCG by the Human Genome Organization.

Each of the NBDs consists of the so-called Walker A and B motifs and is very characteristic for all ABC proteins. Additionally, a C motif, also called the "signature" exists and is enclosed by the Walker A and B regions (Figure 1). What distinguishes the ABC subfamily from other subfamilies is the amino acid sequences found in the NBDs. While it has been recognized that the NBDs are highly conserved among ABC transporters [1], transmembrane domains (TMDs), in contrary, vary between classes and only possess a similar structure within a transporter class.
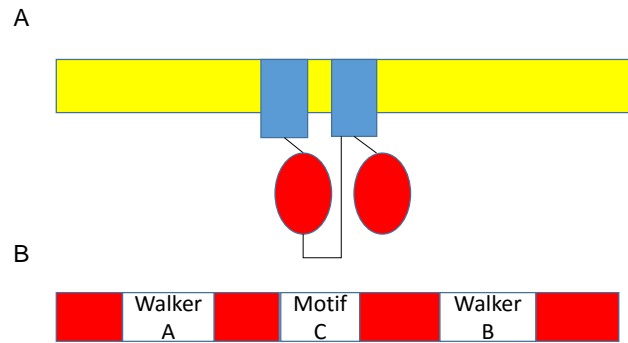
Figure 1: Illustration of typical ABC proteins. A) ABC protein within a lipid bilayer (yellow), the TMDs are depicted in blue, and the NBD in red. B) Regions within an NBD, with Walker A and B and motif C in between.

One unit of ABC transporter contains four functional domains, two transmembrane domains (TMD1 and TMD2) and two nucleotide-binding domains (NBD1 and NBD2) (see Figure 2). However, in some of the subfamilies there is only one TMD and one NBD in one unit. It is therefore known as "half-transporter" and requires homodimerization, as in the case of ABCG2 protein, or heterodimerization to function completely. The number of TM helices also differs between classes, typically containing 6 TM helices [1] [2].
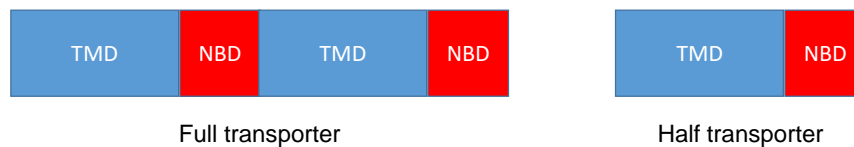
Figure 2: General Architecture of half- and full transporters

A common domain arrangement of a full transporter is composed of NH_2-TMD-NBD-TMD-NBD-COOH. However, other motifs such as NBD-TMD-NBD-TMD, TMD-NBD, and NBD-TMD also exist [2].

The fact that ABC transporters translocate various compounds across membranes has been widely known for several decades. However, the detailed

efflux mechanism is still not fully understood. Several models have been proposed on how ABC transporters extrude the compounds out of the cell. One of the early models proposed was the *flippase* model, in which the translocation of the substrates occurred from the inner leaflet and followed by a flip to the outer membrane leaflet of the cell membrane [3]. The next proposed model was hydrophobic *vacuum cleaner* [4]. This suggests that drugs enter into the membrane, in which Pgp may interact with its substrates and efflux them into the external medium [5] . Another possible theory considers this protein as a unidirectional flux [6], in which substrates are directly extruded out of the cell. This is the so-called *membranar pore* model [7] . The three proposed models are illustrated in Figure 3.
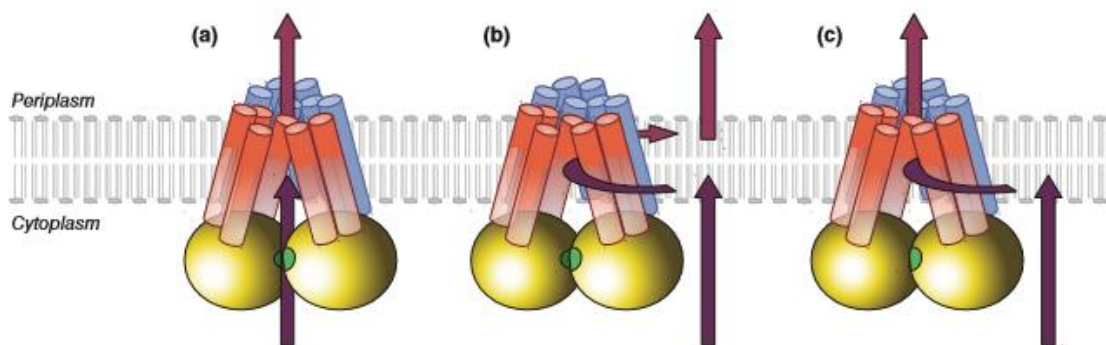
Figure 3: Proposed different models for ABC transporters efflux: (a) membranar pore; (b) flippase; (c) vacuum cleaner. Taken from [7].

The structure of bacterial multidrug transporters has been important historically. The first one published was the prototypical ABC architecture of S. aureus ADP-bound Sav1866 exporter [7]. Later on, a corrected MsbA structure from *E. coli* was available. In 2009, the first murine crystal structure of Pgp, which allows more insights on the MDR transporter, was published. Shortly after this publication of the first mouse Pgp structures, a *Caenorhabditis elegans* crystallographic Pgp structure became available, which demonstrates the different orientation of some of the TM helices. All mentioned crystallographic structures are depicted in the

Figure 4. Recently a refined structure of mouse Pgp was published in the PDB (PDB ID: 4M1M).
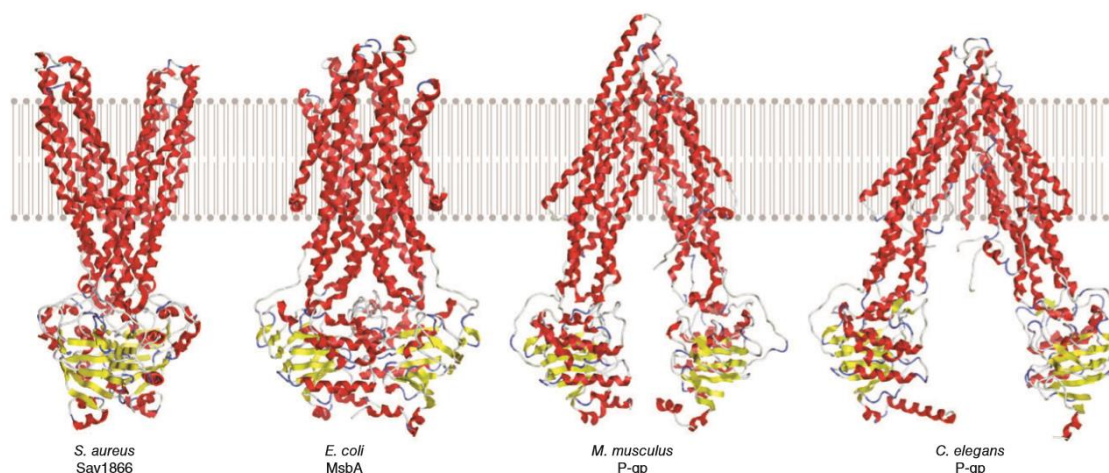
Figure 4: Different crystallographic structures of ABC transporters. Taken from [7].

Some ABC proteins can export many structurally diverse compounds, including cytotoxic molecules from cells. Such transporters are referred to as multidrug transporters and play a very important role in cancer therapy. According to various clinical data, the multi-drug resistance phenotype of tumors is associated with the overexpression of these multidrug transporters, termed MDR Proteins [8]. MDR tumors are resistant to a wide range of structurally unrelated xenobiotics, including anti-cancer drugs, such as vinca alkaloids (vinblastine, vincristine), anthracyclines (doxorubicin, daunorubicin), and taxanes [5].

1.2 ABC Multidrug Resistance Proteins - The Players

The multiple drug resistance phenotype in tumor cells is mostly linked with an overexpression of certain ABC transmembrane proteins: ABCB1 (MDR1/P-glycoprotein), ABCC1 (MRP1), and ABCG2 (MXR, BCRP) [1][2][8][9]. Both ABCB1 and ABCG2 are located at the apical compartment, while ABCC1 at the basolateral side of the plasma membrane of polarized cells [10] (Figure 5). In

general, these MDR transporters contribute to actively translocating a variety of chemically diverse amphipathic compounds including bulky lipophilic anionic, cationic, and neutrally charged drugs, amino acids, polysaccharides, steroids, peptides, toxins as well as conjugated organic anions across the membranes using the energy from ATP hydrolysis [5]. The expression of MDR proteins is mainly in organ tissues that are important for absorption (e.g. the small intestine, gut and lung) and distribution (the blood-brain and placental barriers), as well as metabolism and elimination (liver and kidney) of xenobiotics drugs [11][12]. Sufficient knowledge concerning the relevant physicochemical properties of ligands and how these ABC transporter proteins function is required and there is still a wide range of unanswered questions and insufficient information about the transport mechanism.

Gene term	Protein name (abbreviations, synonyms)
ABCB1	P-glycoprotein (Pgp, Multidrug Resistance Protein 1, MDR1)
ABCC1	Multidrug Resistance Associated Protein 1 (MRP1)
ABCG2	Breast cancer resistance protein (BCRP, Mitoxantrone Resistance Protein, MXR)
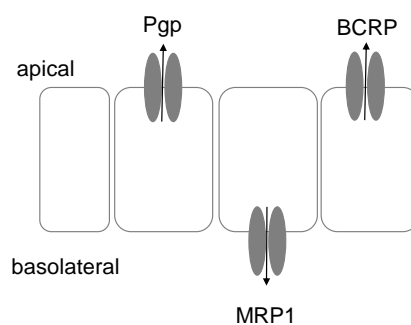
Figure 5: List and localization of the important MDR transporter proteins.

1.2.1 ABCB1 (MDR1/Pgp)

The first ABC transporter that plays a significant role in the multidrug resistance phenomenon is ABC transporter subfamily B member 1 (ABCB1), also known as Pgp (P-glycoprotein) or multidrug resistance protein 1 (MDR1). It is a 170 kDa transmembrane. Pgp was the first to be discovered almost 40 years ago by scientists in Canada [13]. See Table 1 for various Pgp substrates and inhibitors that have been discovered.

Pgp seems to form a functional heterodimer and consists of 2 transmembrane domains (TMDs), with 12 transmembrane helices (TMHs), and two nucleotide-binding sites (NBDs) within one polypeptide chain, as depicted in Figure 6. It is located at the apical (luminal) side of many cells, including the epithelial cells of the intestine, the biliary canalicular membrane of hepatocytes and the proximal tubules of the kidney [8]. A high expression of Pgp has been found in many types of cancer, including renal and colon carcinomas, melanomas, and central nervous system tumors [5].
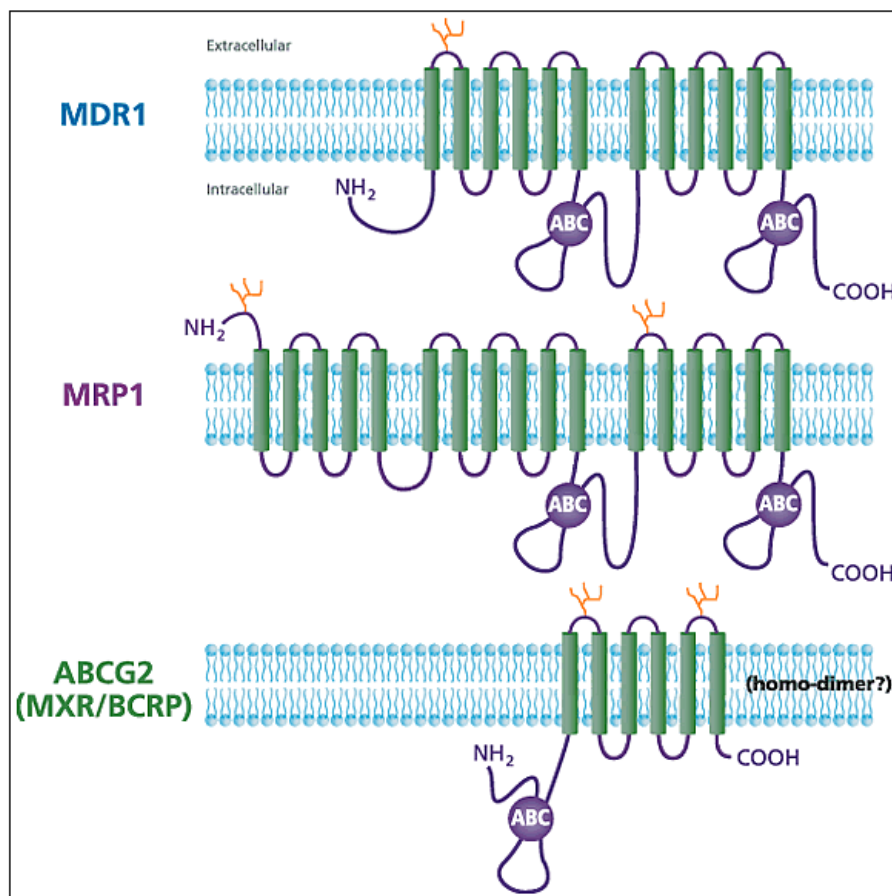
Figure 6: Membrane topology structures of cancer MDR-related ABC transporters.
Taken from [8]

1.2.2 ABCG2 (MXR/ABCP1/BCRP)

ATP-binding cassette subfamily G member 2 (ABCG2), also known as Breast cancer resistance protein (BCRP), has also been identified as playing a major role in multidrug resistance in cancer therapy. It is an approximately 75 kDa plasma membrane protein and it was named based on the fact that it was first isolated from the breast cancer cell lines MCF7 [14]. Additionally, ABCG2 is highly expressed in the placenta, hence the name ABC-Placenta (ABCP1) [15]. Moreover, ABCG2 is also known as mitoxantrone resistance protein (MXR) based on various reports, in which numerous mitoxantrone-resistant cell lines were found, but no indication that ABCB1 and ABCC1 are overexpressed in these cells [16]. Apart from the placenta, this half-size transporter is highly expressed in the gut and biliary tract, in the kidney, liver, small intestine, blood-brain barrier (BBB) and stem cells [12][17] (Table 1).

To date, only a few potent and specific BCRP inhibitors have been developed. In 1998, fumitremorgin C (FTC), a mycotoxin isolated from fumigates, was found to be extremely effective in reversing multidrug resistance in human colon carcinoma cells transfected with the breast cancer resistance protein [18]. Ko143, an analog of FTC, was later discovered [19]. It has been reported to be rather less toxic, and enhance the bioavailability of orally administered anticancer drugs, e.g. topotecan [20]. Other BCRP inhibitors include gefitinib (Iressa, ZD1839), imatinib (both are tyrosine kinase inhibitors) [21], and reserpine [22] (see Table 1).

Unlike ABCB1 and ABCC1, the human ABCG subfamily in general and ABCG2 in particular is a half-transporter, which means that this transporter only consists of one transmembrane domain and one ATP-binding domain. It has a different arrangement pattern compared to the majority of the ABC family. The NBD, in which ATP binding and hydrolysis takes place, locates at the N terminus and the TMD at the C terminus.

Previous studies demonstrated that this transporter forms a homodimer to function at the plasma membranes [5][12]. However, recent studies suggested that this multidrug transport protein likely exists as functional homotetramer [23], or even possibly as homododecamer [24].

ABC transporter	Substrates	Inhibitors	Tissue distribution
Pgp	adamycin [25]	verapamil [26]	kidney
	daunorubicin [25]	cyclosporine A [26]	liver
	epirubicin [25]	PSC 8333 [26]	intestine
	paclitaxel [26]	tariquidar [27] [28]	adrenal gland
	docetaxel [26]	zosuquidar [29][30]	blood-brain barrier
	vincristine [26]	laniquidar [31]	
	vinblastine [26]	ONT-093 [32]	
BCRP	natural substrates	FTC [18]	placenta cells
	xenobiotics	Ko143 [19]	hepatocyte
	chemotherapeutics	elacridar [33]	intestine
		tariquidar [27] [28]	gut and biliary tract
		reserpine [22]	kidney
		imatinib [21]	blood-brain barrier
		gefitinib [26]	stem cells

Table 1: Prominent substrates, inhibitors and tissue distribution of Pgp and BCRP.

The third MDR-related protein is called multidrug-associated protein 1 (MRP1) and belongs to ABC-transporter subfamily C (gene ABCC1) [34]. However, this protein was not analyzed in this work.

1.3 Reversal of MDR-related ABC Transporters

The discussed proteins are likely to cause unsuccessful cancer therapy, because they are able to efflux many cytotoxic drugs commonly used in chemotherapy. The inhibition of these proteins may thus be a promising approach to substantially increase the efficacy of the treatment for cancer [35]. Some strategies to supress

the Pgp function include transporter inhibition by co-administering chemotherapy drugs and modulators, development of new anti-cancer drugs that interact differently with Pgp and are not effluxed by it, down-regulation of the transporter expression, as well as blockage of its disease-related up-regulation [5].

From all options that can be taken under consideration, the application of co-administered Pgp modulator with anti-cancer drugs, e.g. paclitaxel is very likely to be the most favoured and achievable approach to increase bioavailability of these drugs [5]. Therefore, many clinical trials have been conducted to investigate the potential of the co-administration of modulators for cancer therapy. Until now, Pgp inhibitors can be categorized into three generations:

The first generation of Pgp inhibitors included compounds like verapamil, a calcium channel blocker, [36] and cyclosporin A, an immunosuppressant [37]. Various studies have revealed that co-administration of verapamil or cyclosporine A with antineoplastic agents results in an increased bioavailability of the drugs [38][39]. However, high plasma levels were required to reverse MDR in cancer cells, thereby enhancing the likeliness of side effects.

Due to the limited clinical success rate of the first generation of MDR modulators, the second generation was developed. In 1991, the discovery of PSC8333, a non-immunosuppressive cyclosporine D analog, also known as valspodar was introduced by Boesch et al. [40]. The results demonstrated that it was 10-fold more potent than cylosporin A (Sandimmune), suggesting to be an advantageous agent in inhibiting Pgp more effectively. Unfortunately, this agent also inhibited CYP3A4, an important enzyme involved in drug metabolism. This led to an increased levels of antineoplastic agents in various crucial organs for metabolism and excretion, including liver and kidney [41].

These findings point out the need of designing compounds of third generation modulating agents that are more potent, but less toxic for MDR reversal.

In 1993, an acridonecarboxamide derivative GF120918, also known as elacridar, was developed [33]. Furthermore, a series of anthranilamide derivatives, was

developed by Xenova Group, Ltd [27][28]. One of the prominent members of these compounds is XR9576, also called tariquidar. It has been demonstrated that a complete reversal of several human and murine cell lines resistance was achieved by this anthranilamide-based compound [42]. It also has high affinity to bind within the binding pocket in Pgp and inhibits its function to a high degree. Moreover, it has been suggested that it binds to Pgp in a different binding site than other Pgp substrates [35].

However, different from the case of the previous generation of Pgp inhibitors, the enzyme CYP3A4 was not inhibited by third-generation modulating agents. In other words, despite high plasma concentrations of the third generation modulating agents, the plasma levels of antineoplastic agents, e.g. paclitaxel, still remain at therapeutically relevant concentrations [35].

Other third-generation inhibitors [43] that are highly potent and specific for Pgp inhibition include LY335979 (zosuquidar) [29][30], R101933 (laniquidar) [31], and *ONT-093* [32] (see Table 1)*.*

However, to date, there is still no selective modulator against MDR proteins on the market as all third generation modulators failed in phase III clinical trials [41][44].

1.4 Tariquidar and Analogs

As outlined above, tariquidar seems to be promising in inhibiting Pgp without causing any significant increase of plasma levels of anticancer drugs. Hence, after the discovery of tariquidar, a large number of experiments on structural modification of tariquidar in query of a more potent inhibitor of Pgp have been extensively carried out. It has been demonstrated in the previous studies that the two methoxy groups in positions 6 and 7 of the tetrahydroisoquinolinylamide substructure play an important role in inhibiting Pgp [45] and a tertiary amine

group is also important for many Pgp substrates and modulators [46]. The structure of tariquidar is depicted in Figure 7.
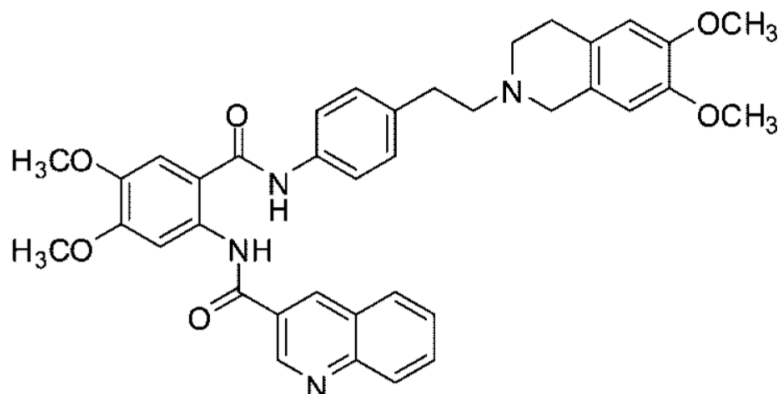
Figure 7: Structure of Tariquidar (XR 9576, N-[2-[[4-[2-(6,7-dimethoxy-3,4-dihydro-1H-isoquinolin-2yl)ethyl]phenyl]carbamoyl]-4,5-dimethoxyphenyl]quinoline-3-carboxamide).

The MDR modulators tariquidar and elacridar are amongst the most potent inhibitors of ABC transporters. They are known to inhibit both Pgp and BCRP with a preference for Pgp in case of tariquidar [47]. However, it has been demonstrated that the introduction of the quinolone-3-carboxamide in the *para* substituent of the benzamide ring instead of in the original meta position of the tariquidar structure and the substitution of 3,4-dimethoxy moiety with a 4-methoxycarbonylbenzoyl moiety resulted in the higher affinity of the molecule towards BCRP [48]. This modification is shown in Figure 8. Since these findings, numerous experiments and measurements have been carried out for tariquidar-like modulators in inhibiting the function of apically localized MDR protein transporters [20][47][49][50][51].
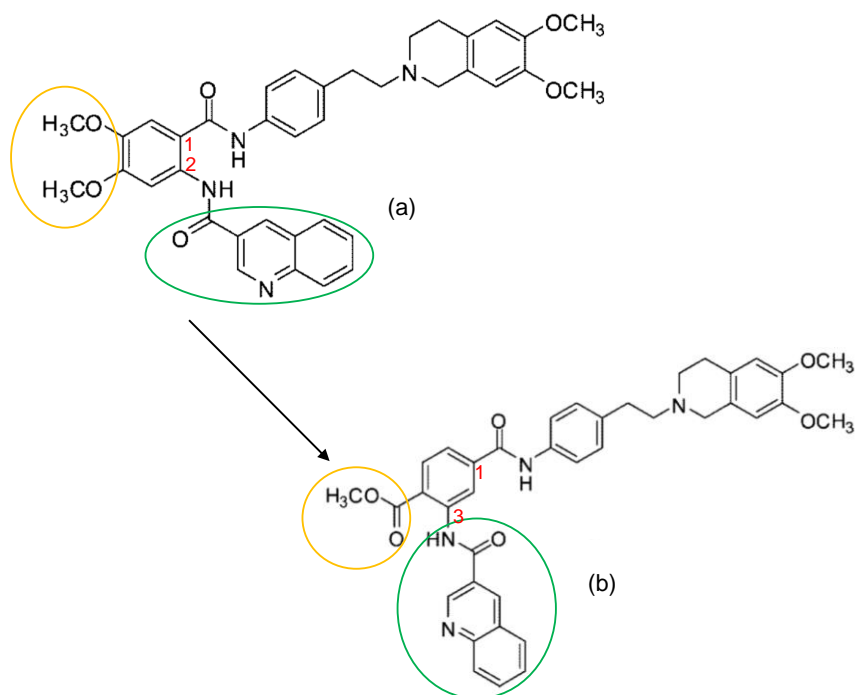
Figure 8: Structure of tariquidar (a) and a more potential modulator of tariquidar analog (b). Minimal structural modification is carried out, including the position of the quinolone-3-carboxamide (circled in blue) in benzamide ring and 3,4-dimethoxy moiety is substituted with a 4-methoxycarbonylbenzoyl moiety (circled in orange). Taken from [48].

2. Aim of the Study

As mentioned in the introduction section, both Pgp and BCRP are located in the apical (luminal) side e.g. of the epithelial cells of the intestine. However, several experiments revealed that a small modification of tariquidar structure influenced the preference of the molecule towards BCRP. Thus, the objective of this work is to analyze and find the structure activity relationship of tariquidar analogs and elucidate the molecular basis of differences encountered in the derivative's affinities towards Pgp and BCRP by using both ligand- and structure-based *in silico* models

This work focused on finding the key factors of this findings using *in silico* methods. There are two types of approaches that can be applied to computationally predict molecular interactions between proteins and ligands: ligand-based and structure-based approaches.

Briefly, predictive models constructed in the ligand-based method do not require the information of the target protein, as they are only based on the structure of ligands in question and the corresponding biological activity measured. In contrast, in the structure-based approach, such as molecular docking, the computational structure prediction of the complexity of interactions between protein and ligands under study requires a high resolution structure of the protein and the ligand.

3. Ligand-based Approach

3.1 Introduction

To date, there is still little structural information about ABC transporters, because they are transmembrane proteins, and therefore difficult to crystallize. For that reason, ligand-based approaches have been carried out extensively. In the ligand-based method, the information of the target (protein) is not required. Therefore, it is the recommended method when the structural information of the protein is not well known.

One of the methods commonly used is Quantitative Structure Activity Relationship (QSAR) analysis [52]. In general, QSAR is applied to form a quantitative connection between the chemical structure and the (biological) activity of a molecule [53]. The approaches include both 2D- and 3D QSAR studies. QSAR studies can be applied both for prediction using linear regression (e.g. partial least square (PLS) or partial component regression (PCR)) and nonlinear methods of classification (k-nearest neighbor, decision tree, random forest, etc).

Its aim is to predict the activity from the structure in a quantitative way without performing *in vitro* or *in vivo* approaches, but instead *in silico* methods by using machine learning methods. In order to perform QSAR studies, the physicochemical properties of structures and their activity measured in an assay are required to make reliable, interpretable models. There are many aspects that can affect the models. One of them is selection of suitable descriptors, which represent the chemical structure. They encode chemical properties of a molecule as numerical descriptions.

This is important to make the models have a high predictive reliability for new sets of compounds, and to prevent over-fitting of the original data. In this regard, one

can check on correlation matrix of the selected descriptors and exclude intercorrelated descriptors.

As mentioned above, QSAR models can be constructed for prediction using linear regression, e.g. PLS, of calculated values of descriptors for a training set. In QSAR regression models, the independent variables (X) are the physicochemical properties of the ligands and the dependent variables (Y) are the biological activity.

The aim of the ligand-based approach was to make selectivity profiling of Pgp and BCRP modulators with QSAR methods. The resulting information about why certain tariquidar analogs have a higher affinity for Pgp or BCRP is necessary to decide which compounds are suitable for further studies and applications.

3.2 Methods

The classic way of finding structure-activity relationships of compounds is by applying QSAR. Before starting with QSAR studies, a set of compounds was first selected. Subsequently, one of the many computational methods that are available for the ligand-based approach was then selected. The method applied in this work was 2D-QSAR, mainly linear regression, using the molecular descriptors available in MOE and also applying respective descriptors in WEKA. Furthermore, as already mentioned above, fingerprint descriptors calculated with PaDEL-Descriptors and jCMapper were also used to make models in MOE and WEKA.

For data preprocessing, the fragmentation approach was also applied (explained later in section 3.2.1.2), where the structure is divided into multiple fragments that are analyzed separately and added back together.

The workflow of the QSAR model development is shown in Figure 9.

15

```
                    ┌─────────────────────┐
                    │  Data preparation   │
                    └─────────────────────┘
                              │         • Compile the ABCB1 and ABCG2 datasets
                              │         • Conversion of the activity value
                              │         • Fragmentation approach
                              ▼
        ┌───────────────────────────────────────┐
        │   Select molecular descriptors and     │
        │            fingerprints                 │
        └───────────────────────────────────────┘
                              │         • Choose 2D descriptors available in MOE
                              ▼
        ┌───────────────────────────────────────┐
        │       Select statistical method        │
        └───────────────────────────────────────┘
                              │         • PLS
                              ▼
        ┌───────────────────────────────────────┐
        │      Validation obtaines of the        │
        │               model                     │
        └───────────────────────────────────────┘
                                        • Internal validation by leave-one-out cross-
                                          validation
```
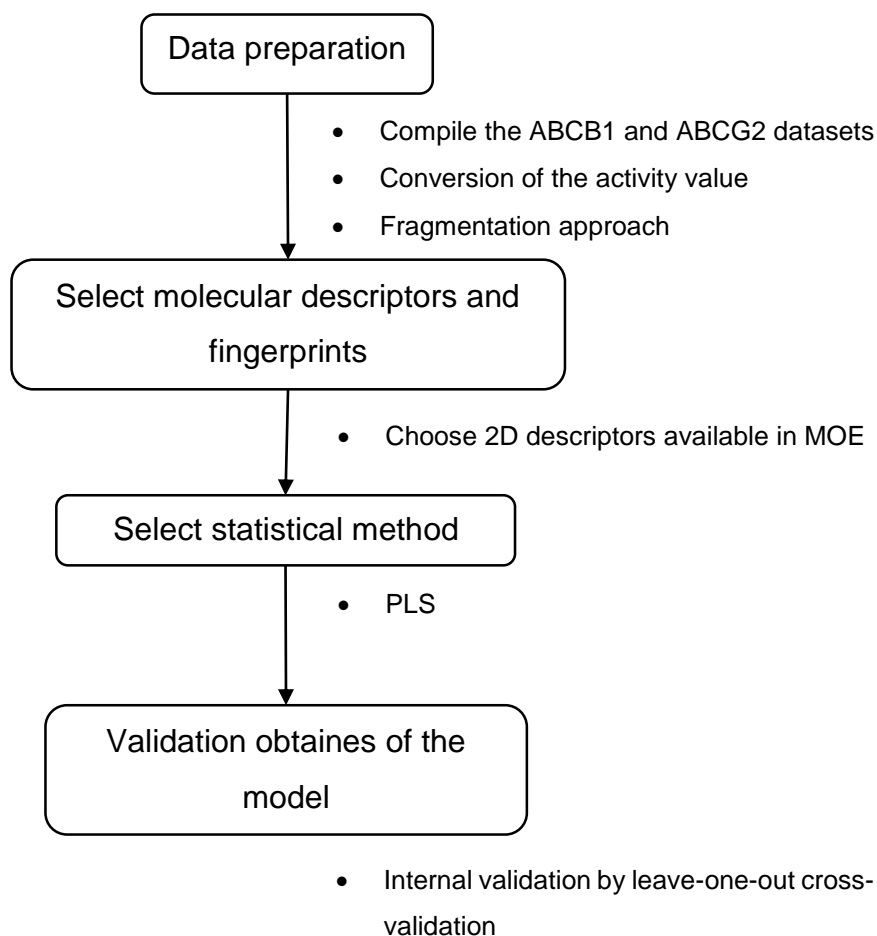
Figure 9: Overview of the workflow of the QSAR model development.

3.2.1 Data Preparation

3.2.1.1 Fragmentation Approach

This approach emphasised on the importance of fragmenting the ligands to gain more insight about the structure activity relationship. This is based on the observation that each of the substituent could be separately analysed for its contribution for the final model. By applying fragmentation, important local molecular features that influence the final model can be analysed. The formula and illustration of this approach are shown in Figure 10.

16

$$pIC50 = a_0 + \sum a_{core} F_{core} + \sum a_{S1} F_{S1} + \sum a_{S2} F_{S2} + \sum a_{S3} F_{S3}$$
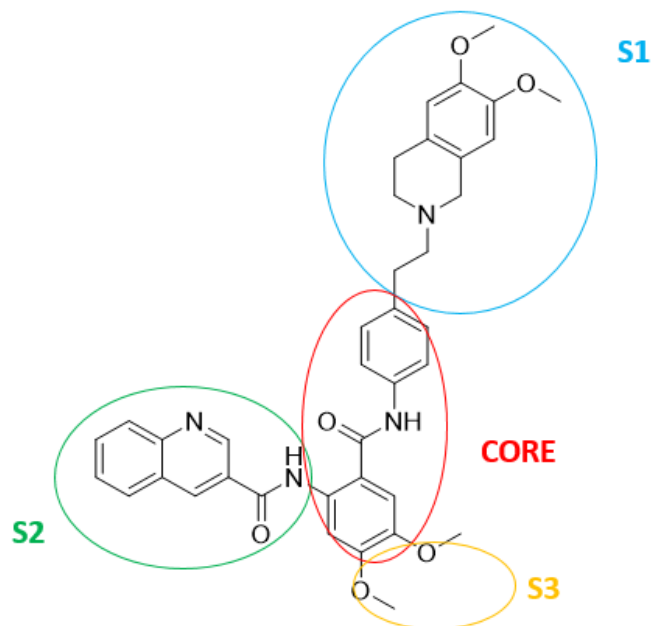
Figure 10: The formula and illustration of the fragmentation approach. Tariquidar is fragmented into 4 parts: red encircled is the core. Blue represents S1, green S2 and orange S3.

Before the approach was performed, the 81 compounds in the ABCG2 database were filtered. Some compounds that have missing fragments were removed from the dataset, because the result could be biased by the incomplete structures. Moreover, the compounds with long substituent(s) were also excluded. At the end, there were 55 compounds in total that have been used for this approach.

As depicted in Figure 10, each of the molecules was divided into 4 fragments: core, S1, S2, and S3. In the end, there are in total of 55 compounds, containing 11 different cores. Regarding the side chains, there are 13, 10, and 11 different fragments for S1, S2 and S3, respectively.

Then, two attempts using two different sets of MOE descriptors were performed. The applied descriptors are listed in Table 2.

Attempt	MOE descriptors
1st	a_acc, a_don, b_rotN, logP(o/w), TPSA
2nd	apol, logP(o/w), mr, TPSA, vdw_vol

Table 2: BCRP models of the fragmentation approach.

The descriptors that were used in the fragmentation approach are listed in Table 3.

Descriptor	Definition
logP(o/w)	Log of the octanol/water partition coefficient [54].
a_acc	The number of hydrogen bond acceptor atoms [54].
a_don	The number of hydrogen bond acceptor atoms [54].
apol	The sum of atomic polarizabilities [54].
b_rotN	The number of rotatable bonds. (A rotatable bond is only taken into account if it is not located in a heterocyclic ring, and has at least two heavy neighbors) [54].
TPSA	Topological polar surface area [55].
mr	Molecular refractivity [54].
vdw_vol	Van der Waals volume [54].

Table 3: Detailed descriptor list used for QSAR studies with fragmentation approach.

In each of the attempts, the descriptors were calculated for each of the fragments. Subsequently, all descriptors of each of the fragments were put back together in one line, so that in the end each compound has 20 descriptors in total (Figure 11).

Figure 11: A partial example of the dataset compilation for the 2nd attempt of building models. Each of the encircled box depicts the 5 calculated descriptors for each of the individual fragment.

After the dataset was compiled completely, the 2D-QSAR models using PLS regression analysis in MOE and WEKA were then constructed.

3.2.2 Tariquidar and Analogs

All pharmacological data for tariquidar analogs were provided by the laboratory of Prof. Dr. Armin Buschauer (University of Regensburg, Germany). The inhibitory activity of some compounds against the Pgp and BCRP transporters was determined in different assays as inhibitory concentration (IC$_{50}$) value (explained below), including a flow cytometric calcein-AM efflux assay using ABCB1-overpressing Kb-V1 cells, Hoechst 33342 microplate assay and flow cytometric mitoxantrone efflux assay using ABCG2-overexpressing MCF-7/Topo cells [50][51][47][20][56][57].

IC$_{50}$ value is the concentration of a substance needed to reach half of the maximal inhibition of a biological or biochemical function and usually measured in nanomolar or micromolar. It is commonly used to evaluate the inhibitory capacity of a substance towards the target, e.g. protein.

Unfortunately, the activity of IC$_{50}$ is not reported for all compounds. Hence, the ligands which didn't show activity of either Pgp or BCRP or of even both were excluded.

19

After compiling the dataset, the IC_{50} value that was available in the initial dataset of all compounds was converted to the pIC_{50} scale, which is the negative log of the IC_{50} value in molar.

$$pIC_{50} = -log_{10}\ (IC_{50})$$

Subsequently, all 38 compounds of the Pgp dataset are sorted by the pIC_{50} value.

At the end, there were 87 tariquidar analogs in total. All of the 87 compounds were then divided into three subsets, as shown in Table 4:

Subset	Contents	Total no. of compounds
1	Compounds with activity measured only for Pgp	6
2	Compounds with activity measured only for BCRP	49
3	Compounds with activity measured for both Pgp and BCRP	32

Table 4: Three subsets for Pgp and BCRP datasets.

In order to build the model for Pgp, the subset 1 and subset 3 were added together because in these two subsets the IC_{50} for Pgp is known (Table 5). The same approach was performed for building the model for BCRP, this time using the subsets 2 and 3.

Dataset	Total no. of compounds	Range of the activity (pIC50)
Pgp	38	3.98 – 6.83
BCRP	81	4.20 – 7.37

Table 5: Overview of the Pgp and BCRP datasets.

3.2.3 Descriptors

In general, descriptors represent the physicochemical properties of a ligand. In this work, both molecular descriptors and molecular fingerprints were used to build QSAR models. As mentioned in the introduction of this chapter, one of the

important steps in constructing a model is the selection of descriptors. Thereby a good, interpretable model can be built, and over-fitting of dataset can therefore be prevented.

3.2.3.1 Molecular Descriptors

The aim of a molecular descriptor is to encode chemical properties of a molecule as numerical descriptions. The descriptors are derived either solely from the chemical structure itself, e.g. sum of the halogen atoms, molecular weight (denoted 1D descriptors), or by taking the configuration of the molecules into consideration, also called 2D descriptors, or furthermore, taking the conformation of the respective molecule into account, as well as the 3D coordinates, the so-called 3D descriptors. These descriptors are used to form a mathematical equation, along with the physio-chemical properties of a compound, e.g. solubility.

3.2.3.2 Molecular Fingerprints

The fingerprints methodology, in which specific substructures present in each molecule will be encoded, was also used in this work. For the fingerprint computation some programs, such as PaDEL-Descriptors [58] and jCompoundMapper (jCMapper) [59] were employed.

In general, molecular fingerprints encode particular molecular structure as a series of binary digits 0 or 1 that symbolize the absence or the presence, respectively, of a particular substructure in the molecule.

There are two types of fingerprints used in this work, Molecular ACCess System (MACCS) fingerprints and extended connectivity fingerprints (ECFP) calculated in PaDEL-Descriptors and jCMapper, respectively.

The MACCS fingerprints consist of a set of 166 predefined structural patterns. The other type of fingerprint used was ECFP fingerprints. Unlike MACCS fingerprints, ECFP fingerprints are not predefined, rather the keys are developed

from the molecule itself [60]. Each specific substructure is first transformed into hash codes using a hashing algorithm, resulting in a large integer value. Subsequently, the hash codes will be folded into smaller number bit strings, according to the number generation of a defined length, normally 1024. Finally, the absence and the presence of a particular substructure is marked as "0" and "1", respectively.The implementation of ECFP in jCMapper is based on the description by Rogers and Hahn [61]. They described that the algorithm starts with the initial atom identifier of the center atom and is expanded into a circular substructure around this particular atom (Figure 12). This process is then repeated according to the defined number of iterations [59].
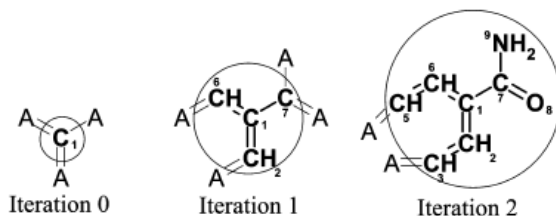
Figure 12: A schematic explanation of ECFP. The atom number 1 is considered as the initial atom identifier. In iteration 0, it only contains of information about atom 1 and its bonds. In iteration 1, the identifier now contains information about atom 1's direct neighbors. After two iterations, the circulated area has grown one atom further. "A" represents any type of atom other than hydrogen. Taken from [61]

3.2.4 Linear Regression

While analyzing the activity, linear regression was applied. This statistical method tries to model the relationship between a set of independent variables X and dependent (response) variable Y in a linear way by constructing a straight line (in case of 2D) that fits the observed data. When there is only one independent variable, it is called simple linear regression. The equation of a linear regression line is:

$$Y = a + bX,$$

whereas X is the independent variable, Y is the dependent variable, a describes the slope of the line and b is the y axis intercept when x = 0. In the case of multiple regression, X is a matrix, and b is a vector.

In order to fit a regression line, there are several methods that can be chosen. The most common method is the least-squares method. It calculates the best-fitting line for the sampled data that minimizes the sum of the squared differences between the estimated values and the actual values, denoted as *sum of squared errors (SSE)*.

3.2.5 Partial Least Square (PLS)

As mentioned in the beginning of this chapter, this work was focused on 2D-QSAR to analyse the relationship between the chemical structure and the biological activity. The statistical method that was applied for the QSAR analysis was PLS regression analysis [62]. This is a linear regression method, which attempts to find a good predictive model that fits the observed data. PLS analysis is normally used when there is a large number of explanatory variables that are often correlated with each other [63]. In this case, models are often over-fitted, i.e. they achieve a good fit but use too many independent variables and are possibly unable to predict new data.

The complete list of the software that was used in this work is shown in Table 6:

Software	Function
jCMapper	for fingerprints calculation
PaDEL-Descriptors	for fingerprints calculation
MOE	for building the models
WEKA	for building the models

Table 6: List of the softwares used for ligand-based studies.

3.2.6 Model Validation

In order to view the fit of the model to the sampled data, one could plot the constructed regression line over the actual observations to evaluate the results.

The result of the fit can also be shown by the root mean square error (RMSE). This indicates how close or far the observed data values are to the predicted values of the model. The coefficient of determination, denoted as r^2, describes how well the regression fits the data. The q^2 value is the cross-validated value of r^2 for model validation. Both r^2 and q^2 values can vary from 0 to 1, with 1 corresponding to an ideal fit. The closer the value to 1, the better the regression line fits the observed data. The q^2 value could be referred to as a quality parameter of a model. A QSAR model that has q^2 value greater than 0.6 could be regarded as satisfactory, whereas a value below 0.6 indicates a low reliability [64]. The formula used for the calculation for the q^2 value is as follows:

$$q^2 = 1 - \frac{PRESS}{TSS} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2},$$

whereas the value of PRESS (Predictive Error Sum of Squares) is calculated by the sum of the squared difference between the actual experimental response y and the response predicted by the regression model.

The TSS (Total Sum of Squares) value can be obtained by the sum of the squared difference of the responses of the experiment and the average experimental responses. Finally, the q^2 value can be determined as one minus the PRESS divided by TSS.

The software MOE and WEKA were used to build the models, using the descriptors implemented in MOE. Finally, the models undergo careful verification procedures in order to validate the quality and prove the reliability and predictive ability of the models on a test set of compounds by applying either internal validation or external validation.

In the internal validation, the original dataset is divided into a training and test set. A model is then built using only the training set and applied to the test set in order to see how well it predicts the data. In contrast, external validation uses all the original data to build a model and the predictability of the resulting model is tested by applying it to a second dataset.

For all approaches in this work, the predictability of the models is calculated by leave-one-out-method (LOO-method). This means that for each experiment N-1 observations are used as the training set and the remaining observation as the test set. This process is then repeated until each of the samples was left out once. This method is very suitable in order to train on as many examples as possible, particularly in sparse datasets.

3.3 Results and Discussion

In the first several attempts, some adequate BCRP models with a high q² value could be built. The equation of the best Pgp and BCRP models is shown in Table 7.

Model	Equation	Total no. of descriptors	Total no. of compounds	r^2	q^2
Pgp	pIC50= 15.65285 **+10.00719 * vsa_hyd** -8.41081 * vdw_area -0.74863 * diameter +1.57997 * bpol +7.19931 * a_IC -9.52772 * a_count	6	38	0.56	0.39
BCRP	pIC50 = 7.86282 +20.60493 * a_nC +13.29378 * a_IC **-21.77544 * a_hyd** +0.53976 * a_don -1.04624 * a_aro -2.12296 * a_acc -20.71580 * bpol -10.44899 * Kier1 +19.09126 * KierA1 +1.26807 * logP(o/w) +0.40926 * PEOE_VSA_NEG -0.60344 * SlogP +4.45880 * TPSA +2.96313 * vsa_acc -2.43893 * vsa_other -5.58947 * vsa_pol	16	81	0.77	0.60

Table 7: Pgp and BCRP models.

2D-QSAR analysis revealed the importance of the hydrophobic surface area and the number of hydrophobic atoms for Pgp and BCRP models, respectively.
This is very interesting, as previous studies already suggested a strong correlation between the compounds' lipophilicity and their inhibitory activity on Pgp [65]. Our results supports this proposition. Since all compounds in the Pgp and BCRP datasets are tariquidar analogs, we would expect similar results in the

BCRP model. However, the number of hydrophobic atoms was demonstrated to have a negative effect in the BCRP model. This indicates that hydrophobic atoms may only contribute to tariquidar-like ligands binding to the protein, but are not the key factor for MDR-related proteins inhibition.

	a_count	a_IC	bpol	diameter	vdw_area	vsa_hyd
a_count	100	99	99	92	99	100
a_IC	99	100	99	91	100	98
bpol	99	99	100	90	99	98
diameter	92	91	90	100	92	92
vdw_area	99	100	99	92	100	99
vsa_hyd	100	98	98	92	99	100

Table 8: Correlation matrix of the Pgp model.

As shown in Table 8, the Pgp model was built with intercorrelated descriptors. This means that the models were constructed with redundant descriptors. As a consequence, although the models may have a q^2 value higher than 0.6, they don't have the ability to accurately predict activities of an external test set.

Conclusively, no good models could be created by using the molecular descriptors implemented in MOE. This may be due to the limited number of compounds in the dataset. Alternatively, one can build QSAR models in WEKA using embedded feature selection.

The next method in 2D-QSAR studies was the fragmentation approach. This approach was focused on elucidating the influence of local physicochemical parameters of each fragment on the biological activity of the molecule. The best developed models only have a cross-validated q^2 value below 0.6

The equations of the estimated linear QSAR models for the two data sets of descriptors are shown in Table 9:

Descriptor	Equation	r^2	q^2
a_acc, a_don, b_rotN, logP(o/w), TPSA	pIC50 = 8.92025 +0.73519 * a_acc core +0.93559 * a_don core -0.87205 * b_rotN core -0.54476 * logP(o/w) core **+0.96634 * a_acc S1** -0.89569 * b_rotN S1 +0.33886 * logP(o/w) S1 -0.18552 * b_rotN S2 +0.49398 * logP(o/w) S2 +0.29119 * TPSA S2 +0.25302 * a_acc S3 -0.65729 * a_don S3 +0.13623 * b_rotN S3	0.72	0.38
apol, logP(o/w), mr, TPSA, vdw_vol	pIC50 = -0.99323 -7.95307 * apol +1.29559 * mr +6.50316 * vdw_vol -3.37690 * apol S1 +2.99255 * logP(o/w) S1 +1.27400 * mr S1 +0.30527 * TPSA S1 S1) +1.24021 * vdw_vol S1 **+11.56918 * apol S2** +3.96067 * logP(o/w) S2 +0.24794 * mr S2 +0.98656 * TPSA S2 -8.39750 * vdw_vol S2 -0.20960 * apol S3 -0.17674 * TPSA S3 +0.34756 * vdw_vol S3	0.49	0.089

Table 9: Results of the BCRP models using fragmentation approach.

We immediately observed that q^2 value is too low in the second model to usefully analyze it further. Hence, we focused our attention on the first model in our analysis. According to its data set, a_acc s1, which represents the number of hydrogen bond acceptor atoms of the first fragment, e.g. tetrahydroisoquinoline, and a_don, which represents the number of hydrogen bond donor atoms core fragment, are the most contributing descriptor for biological activity. The descriptor of apol s2, which represents the sum of atomic polarizabilities of the second fragment, e.g. quinolone-3-carboxamide, is the most important descriptor in the second data set of the BCRP model.

Moreover, analysis of the tendency of each of the fragments showed that hydrophobicity and the number of hydrogen bond donor atoms highly contribute to the ligand-protein interactions (Figure 13).

Interestingly, the result showed that the 3rd fragment, which always contains hydrogen bond acceptor atoms, negatively contributes to the molecule interaction with Pgp.
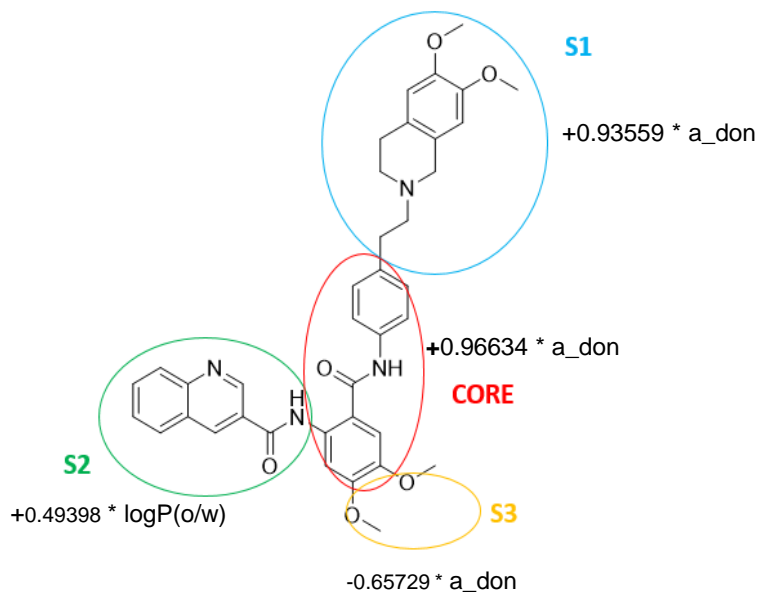
Figure 13: Tendency of each of the fragments.

Finally, QSAR models were constructed in WEKA for both Pgp and BCRP. The methods that were applied were linear regression function, with embedded feature selection. For further analysis, the molecular descriptors and fingerprints used to build the models in WEKA were then applied in MOE. Some satisfactory QSAR models were found. The summary of the results for Pgp and BCRP models is shown in Table 10.

Dataset	Model	WEKA		MOE		Type of descriptor*
		r^2	q^2	r^2	q^2	
Pgp	1st	0.54	0.29	0.54	0.41	MACCS fingerprints
	2nd	0.91	**0.35**	0.91	**0.69**	ECFP fingerprints
BCRP	1st	0.81	-0.01	0.43	0.07	ECFP fingerprints

Table 10: Results of the models built using PLS regression analyses and leave-one-out method for internal validation.

*List of the FP descriptors is shown in Appendix.

At first glance, the 2nd Pgp model's q² value obtained from MOE seems to be reasonable. However, it is important to mention that the methodology used needs to be taken into consideration carefully, since the model was first created using WEKA with all available fingerprint descriptors. Subsequently, the important descriptors that contributed for building the model in WEKA were then selected in MOE instead of selecting all descriptors and building the model from them. As a consequence, a good result of q² value in MOE could be obtained certainly. The problem is that the molecular descriptors used to build the Pgp model in MOE were already selected in the previous attempt in WEKA. This means that the descriptors are already known for their role in making the best model, resulting in a biased, unrealistic q² value.

The approach in selecting descriptors in MOE should also be considered carefully. One should bear in mind that the higher the number of descriptors compared to the number of compounds in the dataset, the more likely it is to select those that could give high q² values. Moreover, as Golbraikh and Tropsha highlighted in their paper, one has to be careful in using only the cross-validated q² value for internal validation of models. They found that there was no correlation between good cross-validated q² values and predictive ability for the test set [66]. Therefore, it is more important to apply external validation, using an external test set to validate the predictive ability of the model, rather than carrying out only internal validation. Since only a sparse dataset was used in this work, external

validation method using an external test set could not be applied here. In cases with sufficient datasets, the dataset can be split into three parts: training set, test set and validation set (also called external test set). While the training set is used to calculate the best-fitting parameters for the sampled data, the test set is used to assess the performance of a fully-trained model. Furthermore, a validation set can be used for external validation in order to see how well the model predicts a new dataset.

As shown in Table 9 and Table 10, all models constructed with different types of descriptors are quite poor or insignificant. This may be due the fact that many of the compounds in the Pgp and BCRP datasets are both Pgp and BCRP inhibitors and are therefore difficult to distinguish.

In future studies, one could also apply selectivity index for this purpose, whereby the yielded models could be interpreted specifically for the determination of Pgp or BCRP preference. Moreover, the number of ligands in the dataset should be sufficient in order to achieve a high degree of correlation of the measured data and the ligands.

4. Structure-based Approach

4.1 Introduction

Unlike the ligand-based approach, the structural information of the target protein is required in structure-based studies. 3D structural data of the protein can be determined using various methods, including X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy. The problem with ABC transporters in general, and ABCB1 (MDR1/Pgp) and ABCG2 (MXP/BCRP) in particular is that they are embedded in the membrane [67]. Such membrane-bound proteins are difficult to crystallize. In order to circumvent this problem, a so-called 'homology model' from a protein with a high sequence identity to human Pgp is built and used for molecular docking instead of an actual human Pgp. Possible templates of such proteins can be found in Protein Data Bank (PDB), e.g. mouse Pgp. The first X-ray structures of mouse Pgp have been resolved in 2009 (PDB Code: 4G5U, 3.8 Å) [68] and then refined in 2014 (PDB Code: 4M1M, 3.8 Å) [69]. Since then, homology models built using the structure could be applied for docking studies. However, in case of BCRP, there is no sufficient sequence identity (>20%) of the transmembrane (TM) domain with any existing ABC structure available in PDB. For this reason, only homology models for human Pgp are available for *in silico* studies.

The objective of the structure-based approach is to gain more valuable information about the drug binding and transport mechanism of tariquidar analogs in Pgp.

4.2 Methods

4.2.1 Homology Model

In order to circumvent the problem that there is no high-resolution X-ray protein structure available, there are some different techniques that can be applied to predict protein structures (i.e. building a 3D protein model):

1) Comparative modeling,

2) *de novo* method and

3) Protein threading or fold recognition.

Depending on the information available from the known structure database, the appropriate approach is applied. When there are similar or identical sequences of known structures available in the protein data bank (PDB) that the target protein can be compared to, the method applied is comparative building, also known as homology modeling.

In the cases where there is no homologous protein with known structure available, both *de novo* and protein threading method are available.

De novo method, which can only be applied for very small peptide chains, can be further subdivided into two different methods: *ab initio* and knowledge-based, which have a rather high complexity and are computationally very expensive [70].

Protein threading, also known as fold recognition, is suitable in cases where there is only low sequence identity between target structure and potential templates. It tries to compare the common folds of target sequences and that of proteins of known structures through amino acids alignment. This technique is comparatively fast and computationally inexpensive compared to the *de novo* method.

In short, homology modeling tries to predict structures similar to the target protein by comparing the sequence of the known homologous protein structure and target

sequence. Based on the assumption that two proteins with high sequence identity would share similar structures, a homology model can be generated, in which a known protein structure sharing a high sequence identity with the target protein sequence could be identified and used as a template. Ideally, the sequence identity of the 3D structure of the template to the target protein should be at least 50% to build a reliable structural model [71].

As depicted in Figure 14, the homology modeling steps include search for homologous sequences of the target protein, identification of the template structure based on the sequence identity with the query protein, the alignment of the target and template sequences, building the homology model, and finally evaluating the model. If the model is not satisfactory, the steps following the finding of the homologous sequence can be repeated again until an acceptable model is obtained.
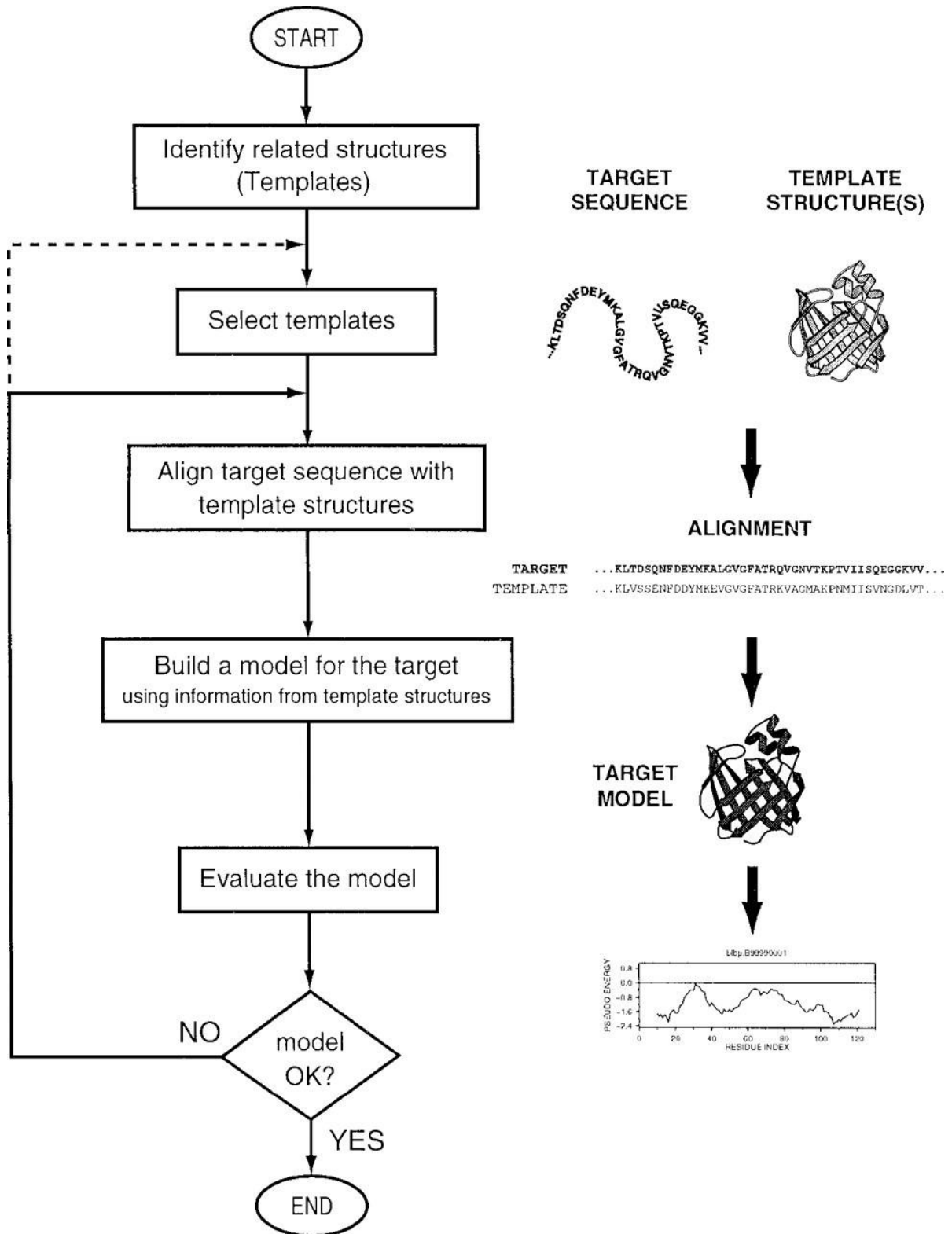
Figure 14 : Homology modeling workflow. Taken from [72].

In March 2009, the X-ray structure of mouse Pgp was available for the first time (PDB Code: 3G5U, resolution 3.8 Å) [68]. This structure is a good template for homology modeling due to its high sequence identity of 87% to human Pgp. Since then, a large number of structure-based studies using this 3D structure have been carried out intensively to gain more insight about the molecular mechanism of Pgp-mediated drug transport [73][37][74].

Since January 2014, a refined structure of mouse Pgp is available (PDB Code: 4M1M, resolution 3.8 Å) [69], in which the inward open conformation of Pgp structure is more closed which reveals a slightly different structure than the previous one, resulting in different potential binding sites.

4.2.2 Model Building

In this work, a sequence of human Pgp was first retrieved from the National Center of Biotechnology Information (NCBI) protein database. The sequence acquired was then used to PSI-BLAST against the PDB Database to find the most suitable structural templates. With 88% sequence identity to human Pgp, 4M1M was used as the template. Sequence alignment of query and template was performed using ClustalX version 2.1. In total, 10 models were generated using MODELLER version 9.13.

After the models were created, each of their Discrete Optimized Protein Energy (DOPE) and GA341 scores were analyzed to select the best model according to their energy levels.

The GA341 scores range from 0.0 to 1.0. The closer the score is to 1, the more likely the model is native-like. Models with the highest GA341 score or with the lowest DOPE score have the most stable minimized energy. However, the DOPE score values are more favorable to identify good models with the highest quality [75]. The result of the assessment scores are shown in Table 11.

No. of model	Molpdf	DOPE score	GA341 score
1	5325.57471	-148136.04688	1.00000
2	5486.13477	-147680.07812	1.00000
3	5364.56543	-148017.43750	1.00000
4	5415.12842	-147936.70312	1.00000
5	5444.74658	-146939.23438	1.00000
6	5419.91162	-148036.26562	1.00000
7	5446.01318	**-148182.35938**	1.00000
8	5859.42627	-147211.87500	1.00000
9	5237.12646	-147866.64062	1.00000
10	5279.91504	-148173.87500	1.00000

Table 11: Summary of successfully generated models.

Since all models were assessed with the highest possible GA341 score of 1, the model no. 7 with the lowest DOPE assessment score was selected for further protein structure analysis.

4.2.3 Model Validation and Optimization

4.2.3.1 Ramachandran Plot

After the best model was selected, the evaluation of the stereochemical quality of the model regarding the outliers were checked with PROCHECK, which is included in the PDBsum analysis [76]. The PROCHECK tool used for the final model structure analysis is in form of a diagram, called Ramachandran Plot, and the related G-factors.

In general, Ramachandran plot statistics are used to represent the final distributions of stereochemical parameters of the final model structure, which are two dihedral angles (φ and ψ) in the backbone of a protein. Additionally, sterically allowed regions for these angles can also be identified [77].

The evaluation of the homology model is done by looking at the percentage distribution of the four important regions: most favoured, additional allowed, generously allowed, and disallowed regions. The higher the percentage distribution in the most favoured area, the better the quality of a 3D structure is. Typically, 90% in most-favoured is desired.

The G-factors show how usual or unusual the stereochemical parameters distribution of the protein main-chain is. While G-factor values below -0.5 indicate unusual properties of the protein structure, values above -0.5 are therefore normally desired.

4.2.3.2 Protein Optimization

After evaluating the model with Ramachandran plot statistics, the structure can be refined in order to ensure chemical correctness and to optimize the protein structures. The Maestro Protein Preparation Wizard can be applied to assign bond orders, add missing hydrogens, and change the protonation state of the residues. The settings for the minimization was left default, using the OPLS 2005 force field with an RMSD cut-off of 0.3 Å.

4.2.4 Docking

The objective of the docking studies is to find out the key molecular factors that determine the activity of inhibitors at Pgp. In order to generate conformations and orientations of a ligand in the binding pocket of a protein, molecular docking could be performed. The aim is to obtain a pose of a ligand, in terms of how a ligand lies within the protein pocket or to 'discover' a new binding hypothesis. Many software packages can be used to perform docking studies, and they differ from each other by their algorithm to create a pose of a ligand, the generation of the grid and the type of scoring function. In the docking studies of this work, 5 compounds were selected. Prior to docking studies, there are some essential

steps to be done, including ligand preparation and grid-generation. Finally, the molecular docking of selected compounds was performed.

4.2.4.1 Ligand Preparation

4.2.4.1.1 Selected Pgp Modulators

In order to investigate the putative binding sites of tariquidar analogs, five compounds were selected based on their pIC_{50} value (see chapter 3.2.2). Compound (Cpd) 5, Cpd 6 and Cpd 94 (tariquidar) which show the highest pIC_{50} value, as well as Cpd 52 and Cpd 53 that have the lowest pIC_{50} value were chosen for further docking studies (Figure 15). In general, all five compounds share a common structure, which is tetrahydroisoquinoline-ethyl-phenylamine as substructure connected to the anthranilamide nucleus. However, different substituents at the central anthranilamide ring can also be observed.
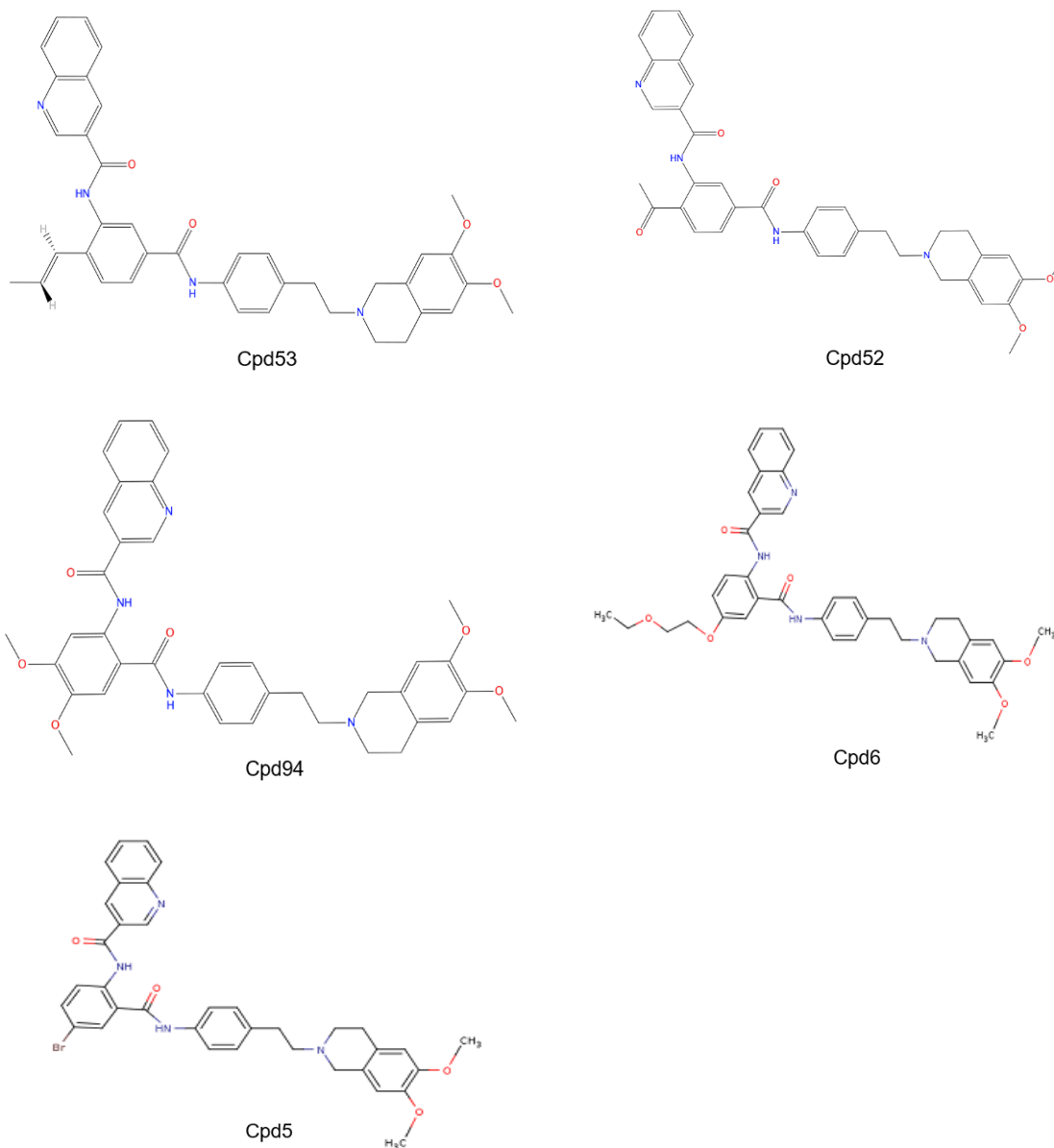
Figure 15: Chemical structure of compounds selected for docking experiments.

The first subset containing three compounds with the highest pIC_{50} value were then regarded as active compounds and the second subset, which includes two compounds with the lowest pIC_{50} value, were regarded as inactive compounds (Table 12).

Compound Name	pIC_{50} value	Regarded as
Cpd 53	3.99	inactive compound
Cpd 52	4.02	inactive compound
Cpd 94 (tariquidar)	6.65	active compound
Cpd 6	6.74	active compound
Cpd 5	6.84	active compound

Table 12: pIC_{50} value of the 5 compounds of Pgp dataset selected for docking studies.

The 3D structure generation of all five compounds was obtained using CORINA with the default settings [78]. Then, database minimization was performed in MOE. The RMS gradient, which indicates the deviation from an optimized structure, was changed to 0.1, and the existing chirality was preserved.

The charge state of the nitrogen atom was checked on the public web resource developed by ChemAxon [79], and it was suggested that the protonated type of the nitrogen in the tetrahydroisoquinoline of each of the tariquidar-like modulators predominates under physiological conditions. Hence, it was regarded as being protonated throughout the docking studies.

The ligands were then prepared with the *LigPrep* protocol in Schrödinger. All of the settings were left to default, including generation of possible ionization states at a target pH of 7.0 ± 2,0, tautomers, and possible stereoisomers of each ligand. The generation of the low energy conformations was set to 1 per ligand.

4.2.4.2 Grid-Generation

In this work, for the determination of the drug-binding region in Pgp Glide from Schrödinger [80] was used. The receptor grid for binding poses for the ligands was defined by coordinates (X=22,4349 , Y=64,5613 and Z=10,7297). All other settings were kept as default. Figure 16 depicts the grid generation for the docking studies.
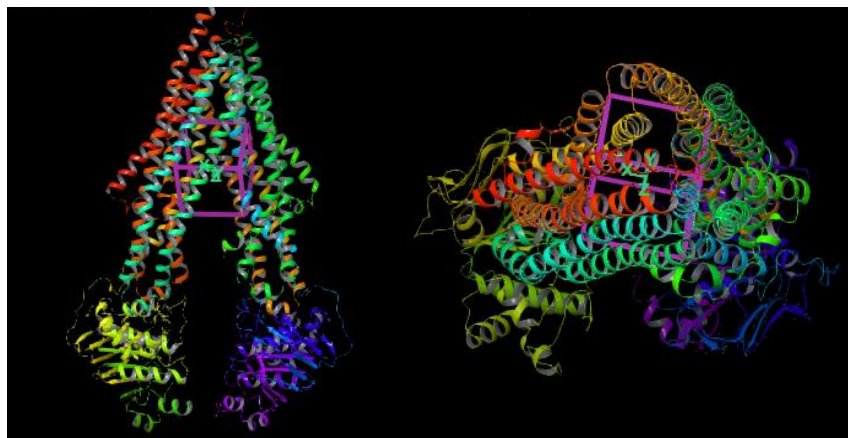
Figure 16: Grid generation for the molecular docking; left side: front view, right side: upper-side view.

Several docking experiments of Pgp inhibitors also used this approximate drug binding region [73][81] These findings are in agreement with some possible binding sites found using the site finder tool in MOE, which are depicted in Figure 17.
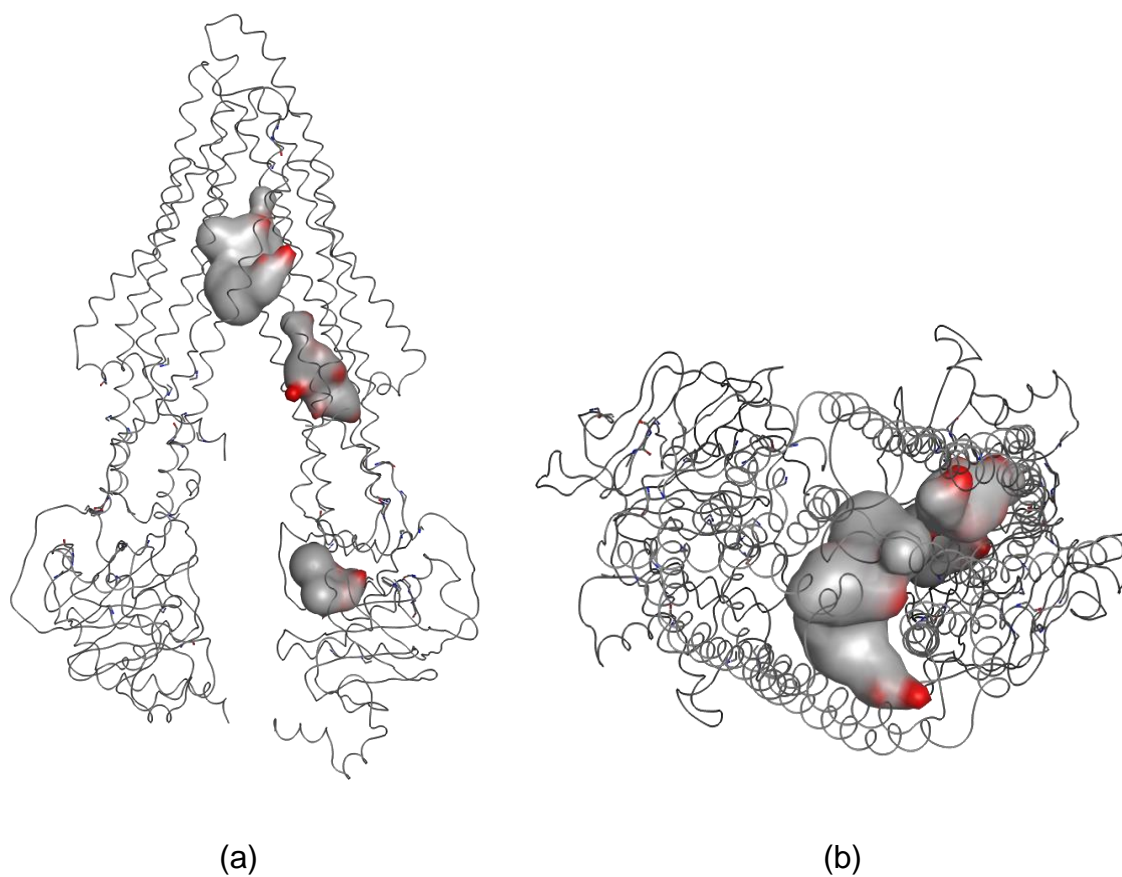
Figure 17: Some possible active sites of P-glycoprotein calculated by site finder tool of MOE (a) front view (b) upper-side view.

Some important residues, obtained from SiteFinder, that are involved in the ligand-protein interactions, can be seen in the Appendix.

4.2.4.3 Docking

Before the docking protocol was carried out, the ASN/GLN/HIS flip state was determined with the web application MolProbity [82]. The docking experiments of five selected compounds were performed with the default settings.

The standard-precision procedure was used because extremely enriched precision was not required for this work. The other type of precision is Glide extra-precision (XP), which allows the top-scoring 10%-30% of the generated ligand poses to be examined more thoroughly with an advanced scoring [83].

For the output quantity and file type, the settings were specified as follows: 100 poses per ligands to be written out at most, and the number of poses per ligand to include was set to 100.

4.2.5 Common Scaffold Clustering

In order to identify groups of objects in a dataset so that members of the group are more similar to each other in specified ways, one can apply clustering analysis [84]. Unique individuals are divided into subgroups, also called clusters, according to specified variables. There are several ways to form clusters. One of them is the hierarchical clustering method, which can be further divided in agglomerative hierarchical clustering and divisive hierarchical clustering. Agglomerative hierarchical clustering is also called a bottom-up approach, because each of the objects in a dataset starts as its own cluster. Two clusters with the shortest distance to each other (again, according to specified variables) are merged into a single new cluster containing both previous clusters. Depending on the characteristics of these clusters at successive steps, the process is repeated until all objects are in one cluster. At the end, a dendogram with a defined cut-off point can be produced. The cut-off point can be defined by desired height, which stands for the distance metrics, e.g. RMSD value between the clusters.

In contrast to the bottom-up approach, the divisive hierarchical clustering begins with a single cluster containing all data entries, and subsequently splits it into smaller clusters. This step can be repeated until each data entry is a cluster on its own. Hence, this method is called a top-down approach. Figure 18 illustrates both agglomerative and divisive clustering.
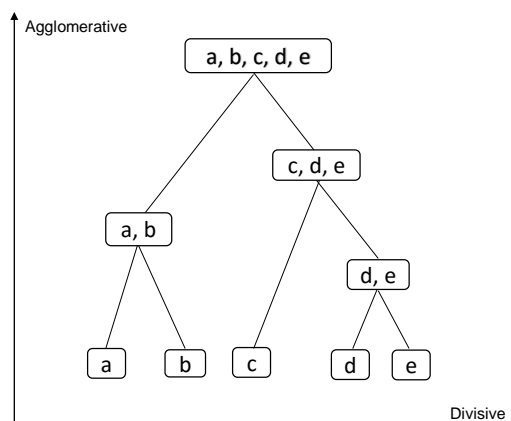

Figure 18: An illustration of hierarchical agglomerative and divisive clustering.

In this work, the clustering of docking results was performed by applying the agglomerative hierarchical clustering and using common scaffold clustering (CSC) [73]. The aim of this method is to group a large number of orientations of ligands with various substituents but with a common core structure (scaffold). This means that compounds with the same scaffold are expected to have a common binding mode. With this approach, highly active compounds and lower active compounds can be distinguished. The concept of common binding mode is based on the idea that active compounds demonstrate a certain binding pose, which inactive compounds don't share.

First, the common scaffold structure for all obtained poses was constructed using MOE (see Figure 19). Then, an in-house SVL script was applied to extract the common scaffold of each of the docking poses. Subsequently, the distance matrix for all generated data was calculated using another SVL script in which the heavy atoms RMSD of the common core are calculated. Finally, an in-house R script

was used to perform the hierarchical clustering. The docking poses are clustered based on the RMSD matrix of the heavy atoms of the common scaffold as defined by the user and is measured in Å.

In this work, the docking poses were clustered at different intracluster distances, ranging from 1 to 7Å. An important assumption in using common scaffold clustering is that all ligands should have the same binding mode [73].
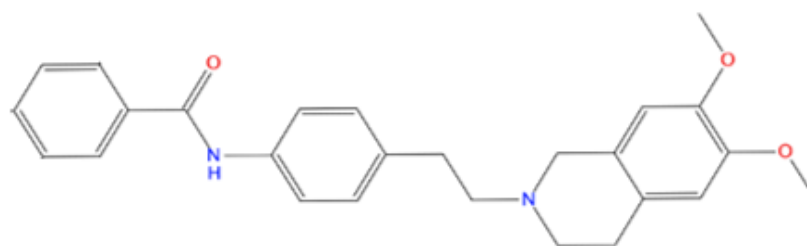

Figure 19: The structure in common with all compounds in the dataset.

Based on the assumption that a common binding pose exists for these five compounds, only those clusters will be considered in further analysis, in which all five compounds were present. Various orientations of the substructures of the compounds that were not part of the common scaffold were then analysed to gain information regarding the molecular differences and the respective measured activity.

4.2.6 Scoring and Rescoring

The assessment process for the generated clusters of the docking poses was done by a combination of the scoring functions in Schrödinger and MOE as well as published experimental mutagenesis data [68] for prioritizing the ligand-protein binding hypotheses [85].

4.2.6.1 Scoring Functions

In order to evaluate the generated docking pose of a ligand in the binding pocket, scoring functions have to be applied, which rank obtained ligand poses. There are three different basic types of scoring functions: force-field based, empirical and knowledge-based scoring functions.

Force-field based scoring functions evaluate the affinity between two binding partners by calculating the energy of intermolecular van der Waals and electrostatic interactions between atoms of the protein and the ligand with a particular distance cut-off to reduce the computational costs [86]. The energy parameters of this type of functions are derived both from the data of complexes and *ab initio* quantum mechanical calculations [87]. Several force-field based scoring functions (although there is sometimes a mild transition to empirical scoring functions) include London dG, GBVI/WSA dG and Affinity dG, which are used among others for the docking studies in this work.

Empirical scoring functions estimate the free energy of binding for a protein–ligand complex. Various energy terms such as van der Waal interaction, electrostatics, number of hydrogen bonds, hydrophobicity, etc. are derived by fitting experimental data of a protein-ligand complex to known binding affinities [88][89]. Due to simple energy terms, this type of scoring functions is computationally faster than force-field based scoring functions. An example of an empirical scoring function is Glide SP/XP [90], which is applied in this work.

The last class of the scoring functions is knowledge-based, or also known as statistical-potential based scoring functions. Energy potentials are derived from large crystal-structure data and the frequency of occurrences of the complexes is observed to calculate the statistical potential for the evaluation of ligand-protein binding [87]. Furthermore, a consensus scoring could also be applied, which combines several functions in order to overcome any potential limitations of each of the above discussed scoring functions [91].

In this work, consensus scoring was applied. For this, both Glide and MOE scoring functions were used.

For all scoring functions, more favourable poses are indicated with lower score. An overview of the applied scoring functions is shown in Table 13.

Name of the scoring function	Software	Type of scoring function
GlideScore	Schrödinger	Empirical scoring function. It approximates the ligand binding free energy
London dG	MOE	Force-field based scoring function. It attempts to estimate the binding free energy from a given pose
GBVI/WSA dG	MOE	Force-field based. It estimates the free energy of binding of the ligand from a given pose
ASE	MOE	Shape-based method [92]. It uses a simple Gaussian overlap scoring function
Affinity dG	MOE	Force-field based scoring function. It attempts to estimate the binding free energy from a given pose
Alpha HB	MOE	Shape-based method [92]. It combines geometric fit plus a hydrogen bond term [93]

Table 13 : Overview of applied scoring functions.

4.2.7 Protein-Ligand Interaction Fingerprint (PLIF)

One of the methods that can be applied to analyse the important residues for the interactions between the ligand and protein is the Protein-Ligand Interaction Fingerprint (PLIF). The individual interactions, such as hydrogen bonds both in the sidechain and backbone, ionic interactions and surface contacts, can be displayed and observed visually, and transformed into a binary fingerprint scheme. This is organised in bits, in which the present interactions will be marked as designated letter and the absence of the corresponding interaction will be annotated as "-".Table 14 shows the list of the ligand-protein interactions by PLIF.

This analysis was done after the common scaffold clustering was performed.

Annotated letter	Type of interaction
D	Sidechain hydrogen bond donor
A	Sidechain hydrogen bond acceptor
d	Backbone hydrogen bond donor
a	Backbone hydrogen bond acceptor
O	Solvent hydrogen bond
I	Ionic interaction
c	Surface contact

Table 14: List of the ligand-protein interactions in PLIF.

4.3 Results and Discussion

The inward-facing human Pgp homology model built based on the new X-ray structure of Pgp was validated with Ramachandran Plot Statistics. Subsequently, a docking protocol was carried out on this model using five selected compounds that include active and inactive compounds. Then, all obtained ligand poses were clustered using common scaffold clustering (CSC) in order to group similar compounds into clusters. The clusters were analysed until the maximum intracluster distance of 7 Å. 33 of the 1643 clusters were investigated further. Items in the clusters were ranked with both Schrödinger and MOE and only the six clusters that got the same internal ranking were kept. Finally, the analysis of the studied-ligands' poses was carried out in order to generate more insights about potential binding areas for these compounds.

4.3.1 Homology Model - Model Validation

The model with the lowest DOPE score was further evaluated with the Ramachandran Plot Statistics. 93.8 % of the residues are to be found in most

favoured regions, 4.2 % in additional allowed regions, 1.3 % in generously allowed regions and only 0.7 % in disallowed regions (see Figure 20).
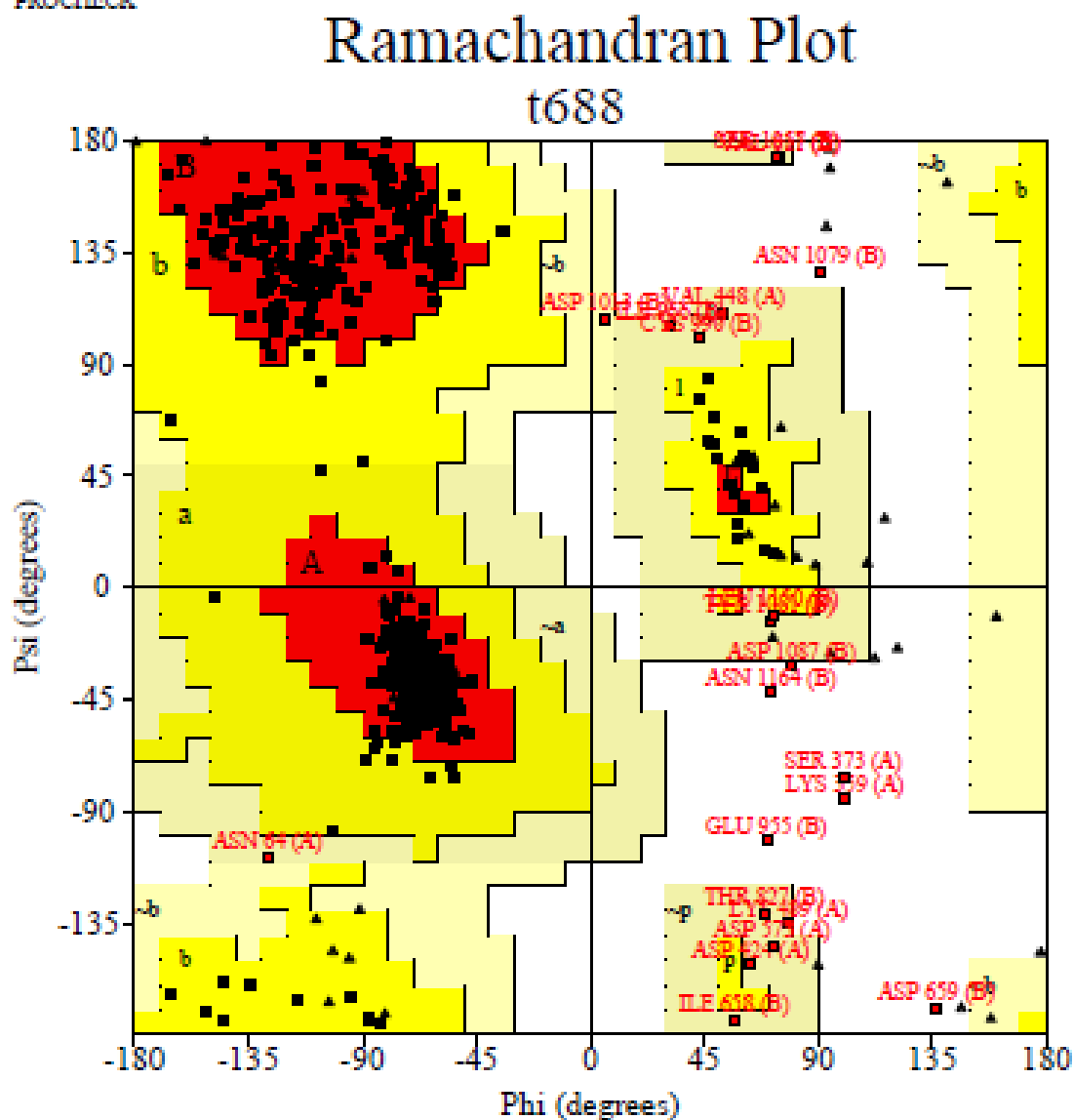
Figure 20 : Ramachandran plot of the selected homology model.

The selected homology model was then loaded into the Maestro Protein Preparation Wizard. After the restrained minimization was completed, the result of the final homology model was then again analysed with PROCHECK.

In the Ramachandran Plot for the minimized protein 92.8 % of the residues are to be found in most favoured regions, 5.4 % in additional allowed regions, 1.3 % in generously allowed regions and only 0.6 % in disallowed regions (see Figure 21).
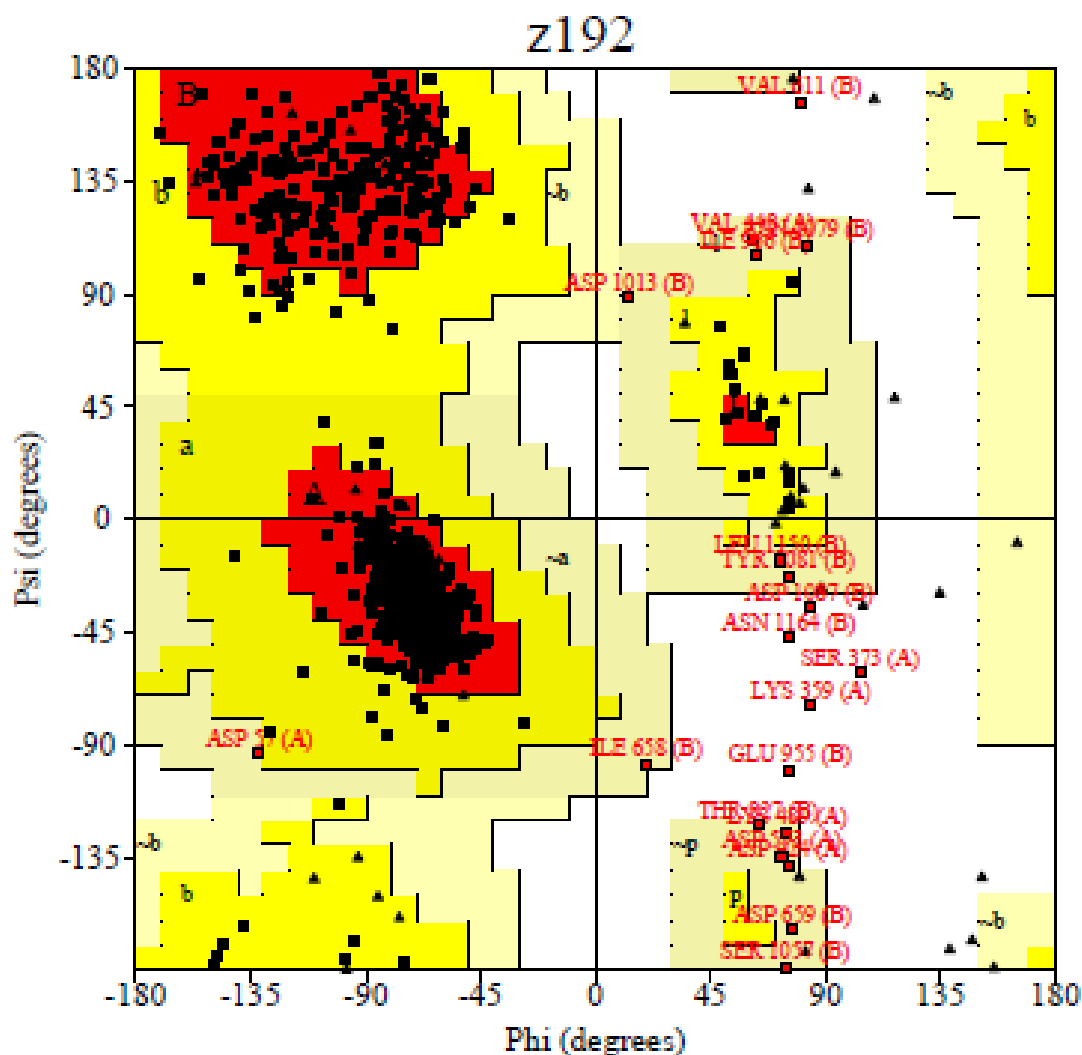
Figure 21 : Ramachandran plot of the final homology model after the restrained minimization.

The residues that are in the disallowed regions include Lys 359, Ser 373, Asp 659, Glu 955, Asn 1079, Asp 1087 and Asn 1164. After the restrained minimization was performed, some of the residues in the disallowed regions in the previous model were shifted to other regions. In the refined homology model,

5 residues from the previous model, including Lys 359, Ser 373, Glu 955, Asp 1087 and Asn 1164, are still in the disallowed regions and additionally, Val 611 is now also in the disallowed region. Further investigation reveals that all residues in the disallowed regions are mainly located in the nucleotide-binding domains (NBDs), which are not the binding regions of Pgp modulators. All previously mentioned residues are depicted in Figure 22.
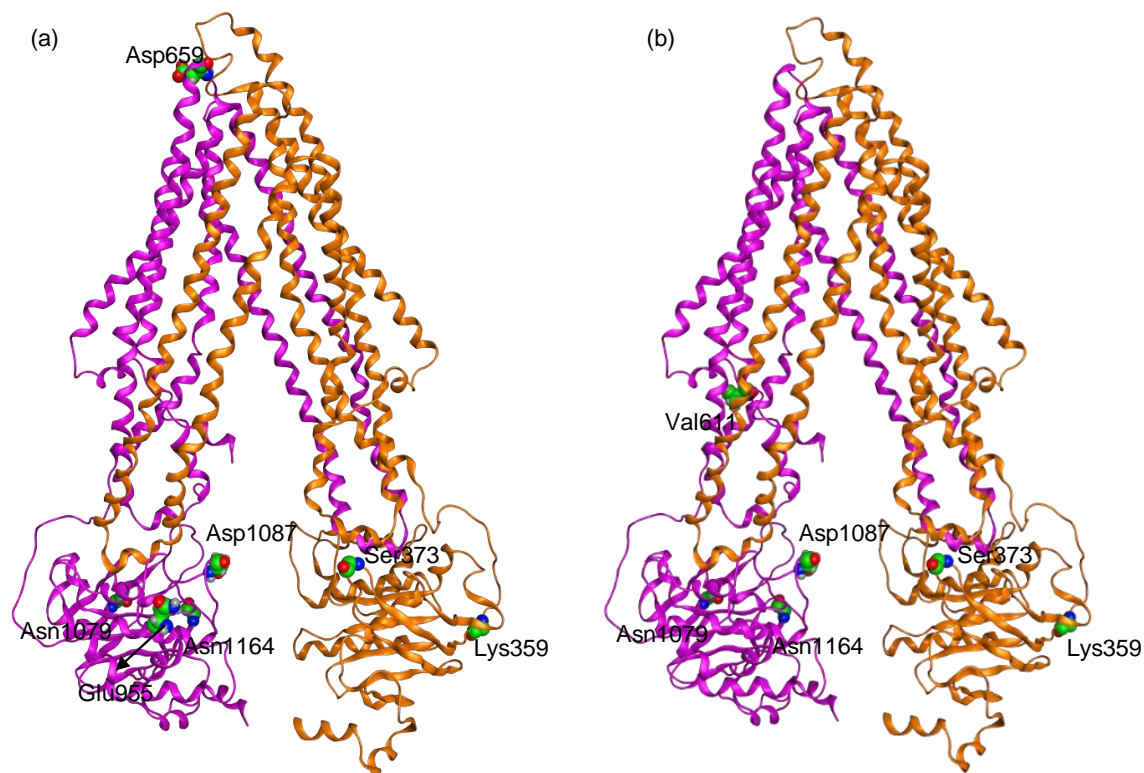
Figure 22: Homology Model before the minimization; (b) Homology model after the minimization.

4.3.2 Docking

After ligand preparation, 2 isomers for each of the active compounds (Cpd 5, 6 and Cpd 94 (tariquidar)) and inactive compounds (Cpd 52 and 53) were generated, resulting in 10 ligands in total. After the docking process, in total 981 poses were obtained.

In terms of the chemical structure of the 5 ligands under study, it is important to note that the shift of the quinolone-3-carbonylamino moiety from position 2 (ortho) to position 3 (meta) results in a less inhibitory activity of Cpd 52 and Cpd 53 towards Pgp (see Figure 15).

All compounds under investigation bind in the common drug-binding pocket. This region is comparable with the "lower" binding site of QZ59-SSS in the x-ray structure [69] and with the possible mechanism of drug transport of Pgp. Moreover, numerous results from cross-linking and photoaffinity labeling, combined with site-directed mutagenesis studies are also in agreement with this location.

In order to comprehend the transport mechanism of Pgp, it is important to identify the binding site of the protein. Knowledge from previous studies regarding the important residues that may be crucial for protein-ligand interactions could be a helpful guidance for docking experiments. It has been suggested that the probable binding site(s) is (are) centralized in the cavity within the plasma membrane [94], and it has been also demonstrated in various experimental studies using different methods [95][73]. Moreover, site-directed mutagenesis experiments and chemical cross-linking studies have identified some residues that are potentially important for substrate transport. Cross-linking experiments with the thiol-reactive substrate tris-(2-maleimidoethyl)amine (TMEA) in SDS-polyacrylamide gel electrophoresis on the human Pgp demonstrated that L339 (TM6) and V982(TM12) are important to bind the substrate [96]. Moreover, ATPase activity of the respective mutants was inhibited by cross-linking with TMEA, and some Pgp substrates (Cyclosporin A, Vinblastine, Colchicine, and Verapamil) inhibited cross-linking by TMEA.

Site-directed mutagenesis of 6 residues with cysteine, including Tyr307 (TM5), Phe343 (TM6), Gln725 (TM7), Phe728 (TM7), Phe978 (TM12) and Val982 (TM12) showed that some Pgp substrates and modulators (QZ59S-SSS, cyclosporine A, tariquidar, valinomycin and 5'-[p-(fluorosulfonyl)benzoyl]adenosine (FSBA) failed to inhibit the labelling of Pgp with

the substrate [(125)I]-Iodoarylazidoprazosin, when Tyr307, Gln725 and Val982 are mutated. However, ATPase activity of the mutant Pgps was not decreased, showing that there is more than one active binding site for each substrate available for the transport [97]. Furthermore, it has also been shown by the same research group a couple of years later that at least two different substrates could bind in the drug-binding pocket of the protein at the same time [98]. Another site-directed labelling approach with a thiol-reactive substrate also demonstrated that Lys339 is important for the inhibition of the transport function of Pgp. [99][100] Other studies using MTS-Verapamil (a thiol-reactive methanathiosulfonate-verapamil) also suggested that the ATPase activity of MTS-verapamil labelled mutant Ile306Cys significantly increased (8-fold higher than untreated controls) [101]. Similar results could also be observed in the cysteine-scanning mutagenesis studies with the thiol-reactive substrate dibromobimane (dBBn) on human Pgp. Some mutants (L339C, A342C, L975C, V982C, and A985C) were protected from inhibition by dBBn with administration of Verapamil, Vinblastine and Colchicine [102]. Findings from other studies [103] showed that the ATPase activity in the MTS-rhodamine labelled mutant Phe343Cys was enhanced 5.8-fold. Moreover, the ATP hydrolysis also increased by adding Verapamil to the MTS-rhodamine-treated mutant. Other studies with MTS-Verapamil also emphasized the significance of residues Phe728, Ile306, Phe343 for contributing to drug binding. The apparent Verapamil affinity decreased when Phe728 is mutated to Cys. However, the ATPase activity of the MTS-verapamil-labelled Phe728Cys mutant was further enhanced by 11.5-fold than those of untreated controls [104].

Taken together, these results suggest that the binding pocket for drug-protein interactions is formed in the interface between TM5, TM6, TM7 and TM12. All above mentioned residues are listed in Table 15 and depicted in Figure 24.

Suggested important residues	Residues in homology model
Tyr307	Tyr277
Phe343	Phe313
Gln725	Gln641
Phe978	Phe894
Val982	Val898
Leu339	Leu309
Ile306	Ile276
Arg342	Arg312
Leu975	Leu891
Arg985	Arg901

Table 15: Suggested critical residues from mutagenesis experiments for Pgp inhibition.

These various experiments with verapamil and analogs are of importance, as verapamil shares a common structure with tariquidar (Figure 23). This observation was also pointed out by a previous studies with tariquidar-like Pgp modulators [105].

R = quinoline-3-carbonylamino
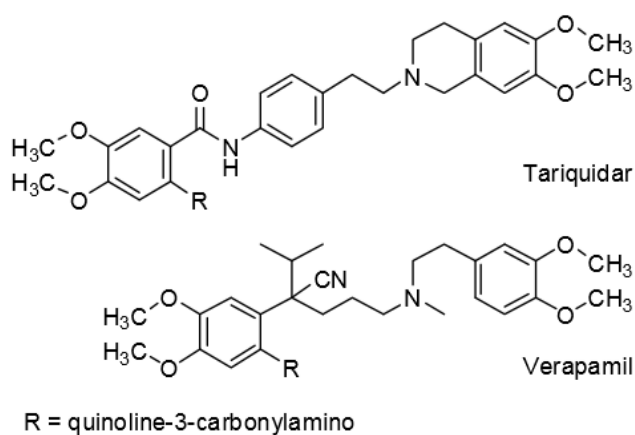
Figure 23 : The structure of tariquidar and verapamil shows common substituent of dimethoxyphenyl moieties and tertiary amine. Taken from[105].
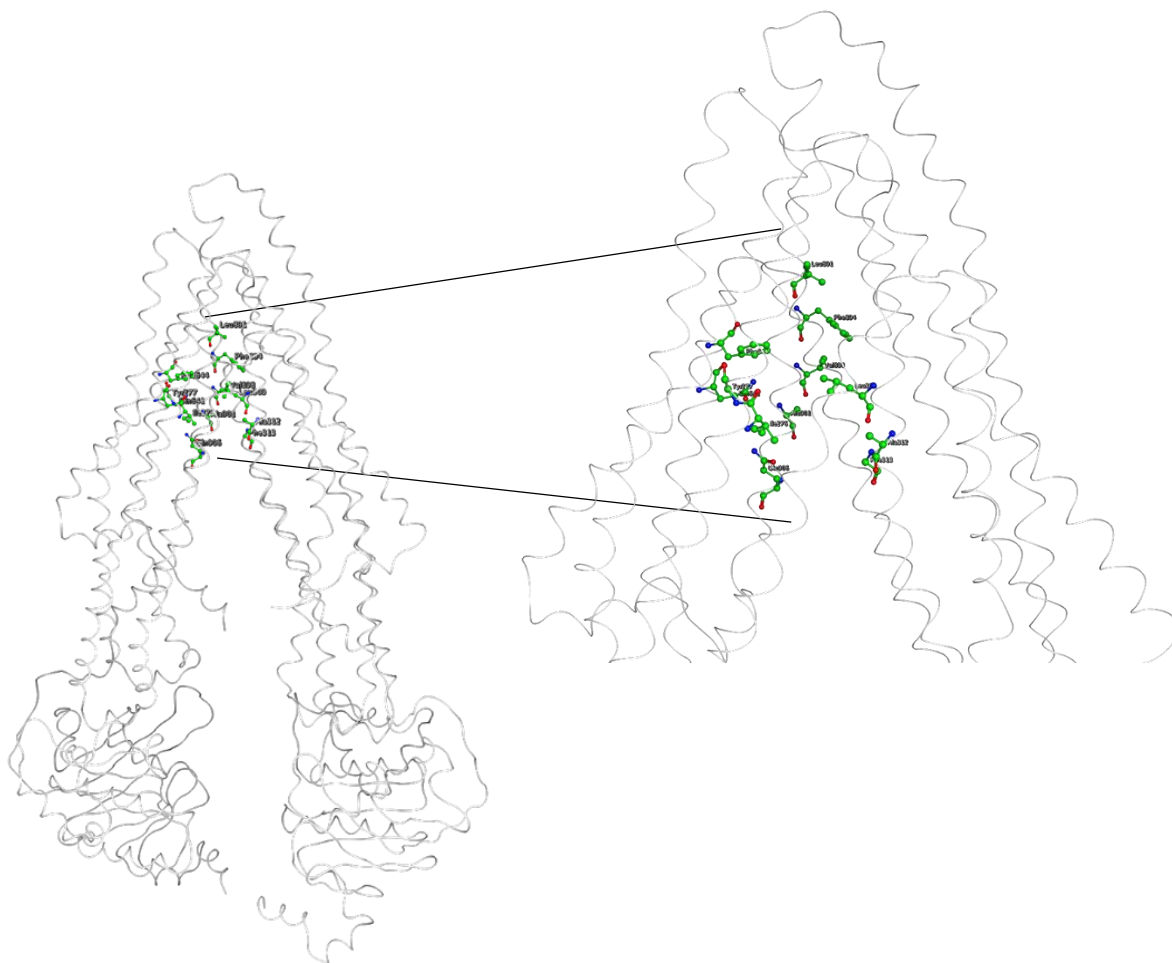
Figure 24 : Critical residues based on various mutagenesis experiments.

In previous studies it has been assumed that tariquidar has the same binding mode as the Pgp substrate Hoechst 33342 [50][35]. Moreover, this finding is in accordance with other studies that revealed some putative drug binding sites of Pgp for flavonoids [67] and propafenone derivatives [73] which share some common structure with tariquidar (tetrahydroisoquinoline substructure). The placement of the grid in this step was therefore based on those studies, which suggest that the binding sites are near to the central cavity of Pgp (see Figure 16). The other two possible binding sites calculated by the *Sitefinder* tool in MOE were disregarded because no evidence was found in the literature that they could be part of putative binding pockets (Figure 17).

4.3.3 Common Scaffold Clustering

As described in the Methods section, a common scaffold clustering (CSC) was performed, and the results of the clustering were analysed up to an intracluster distance of 7 Å. For the analysis, only clusters containing poses of all five ligands were kept. As there are no cluster in the first group (i.e. maximum intracluster distance of 1 Å) which fulfilled this requirement, the analysis started from the distance of 2 Å (see Table 16).

Maximum intracluster distance (Å)	Total clusters	Remaining clusters after elimination
1	657	0
2	388	1
3	236	2
4	149	7
5	98	7
6	66	7
7	49	9
Total	1643	33

Table 16 : Summary of the clusters of obtained docking poses.

4.3.4 Scoring and Rescoring

Prior to analysis, scoring of the poses in all the remaining clusters was done in Schrödinger, using Glidescore, and the rescoring in MOE, using London dG, GBV/WSA dG, ASE, Affinity dG and Alpha HB scoring functions. According to the results of the rescoring, the rank of each of the ligands within a cluster is not always consistent with the measured activity. With the assumption that active compounds are supposed to be better scored than inactive compounds, only those clusters were further analysed, in which the ligands are correctly scored by the six scoring functions. An example of the rescoring can be seen in Table 17:

Maximum intracluster distance (Å)	Cluster	No. of poses	Glide SP	London dG	GBVI/WSA dG	ASE	Affinity	Alpha HB
5	1	46	-	-	-	-	-	-
	7	43	✓	✓	✓	✓	✓	✓
	8	32	-	-	-	-	-	~
	19	21	✓	✓	✓	✓	✓	✓
	28	32	-	✓	~	~	~	~
	49	31	-	✓	~	-	~	-
	55	39	-	~	~	-	-	~

Table 17: Rescoring of the maximum intracluster distance of 5Å.

✓   = ranking preserved

-   = incorrectly ranked

~   = ranking moderately preserved

All the clusters in the maximum intracluster distance (MID) of 2-5 Å were rescored with the above mentioned scoring functions. This resulted in six remaining clusters that were used for a more detailed investigation regarding the ligand poses, which are shown in Table 18:

MID (Å)	Cluster	No. of poses	Cluster ID
2	54	8	I
3	39	11	II
4	7	25	III
	33	16	IV
5	**7**	43	V
	19	21	VI

Table 18: Summary of the 6 selected clusters to be further analysed.

The predicted important residues for the protein-ligand interactions were calculated by PLIF. The binding site consists of predominantly amino acids in the TM α-helical regions 3, 5, 6, 11 and 12 (see Figure 25).

Figure 26 depicts the corresponding residues for the active and inactive compounds in the protein.
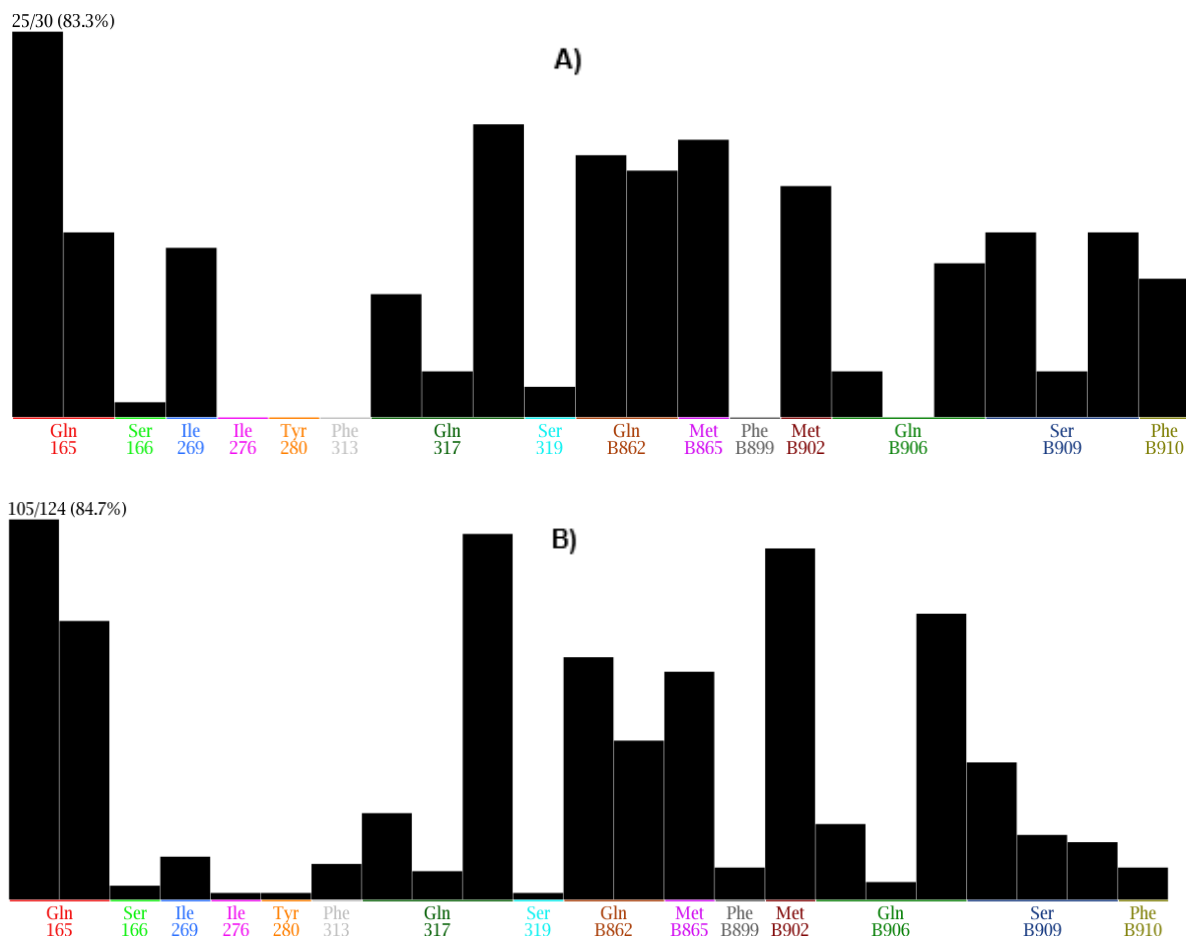
Figure 25: PLIF for all poses of the 6 remaining clusters. Important residues for protein interaction with A) inactive compounds and Panel B) active compounds.

According to the PLIF results, Gln165, Gln862 and Met865 were found to be important residues of Pgp involved in tariquidar binding. These 3 amino acid residues are located near the common scaffold. Thus, interactions with these residues were observed. However, interactions with Ile276, Tyr280, Phe313 and Gln906 only exist with active compounds. It can be observed that there is only a

small number of poses showing interactions with these residues. Interestingly, those poses have a higher docking score compared to other active compounds that don't show interaction with these residues, e.g. a ligand pose that shows interaction with Tyr280 has a higher score than the ligand pose that doesn't show interaction with this residue. Some important residues obtained from the docking experiments, including Ile276, Phe313 and Gln906 (Ile306, Phe343 and Gln986 in the murine Pgp structure, respectively), are in agreement with mutagenesis data. Tyr280 is located near the molecules, indicating that a hydrogen bond between Tyr280 and the carbonyl group of the ligands could be formed, as demonstrated in previous studies [73].
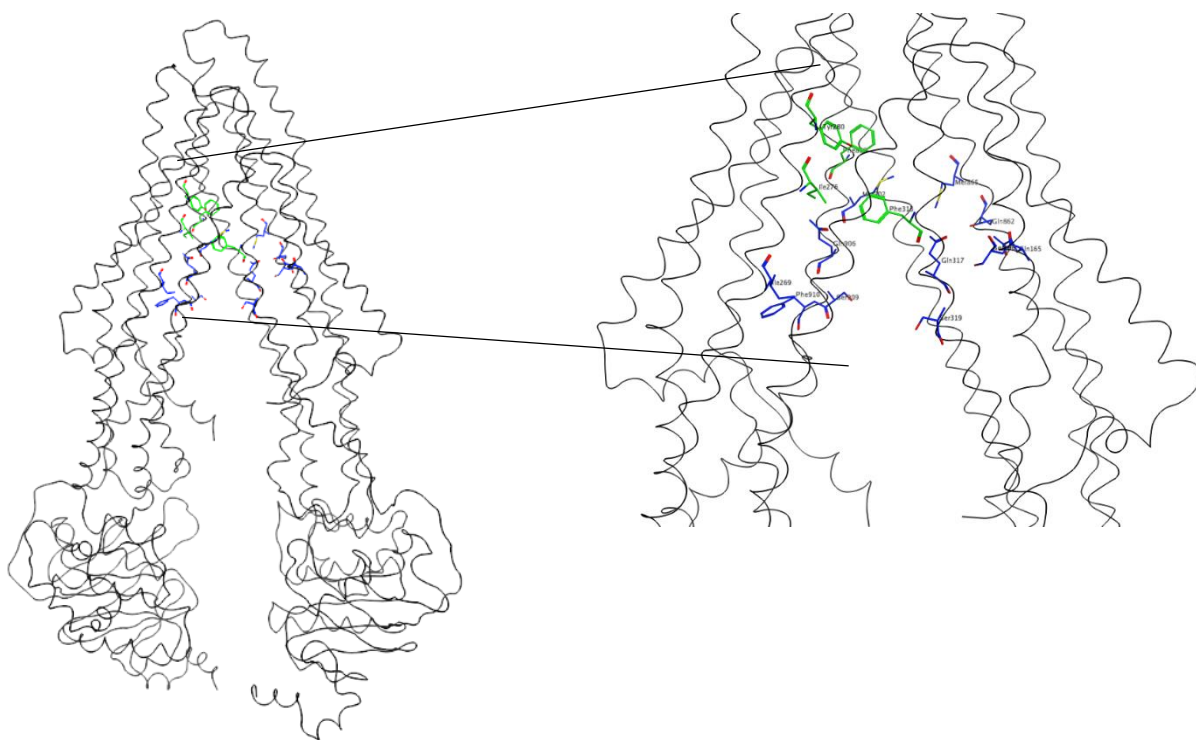
Figure 26: All important residues calculated by PLIF; green: residues for active compounds; blue: important residues for inactive compounds.

The 124 poses of the 6 clusters under study include 30 poses of inactive compounds and 94 poses of active compounds.

Further analysis was done for cluster V (Table 18), which is cluster 7 of maximum intracluster distance of 5 Å. In the hierarchical agglomerative clustering, the closer the cut-off point to the top, the more data points are merged into one cluster (Figure 18). Thus, cluster V, which is in the highest level of the tree, includes also some poses from the lower maximum intramolecular distance. Moreover, cluster V also has the highest population. Cluster V contains 11, 10 and 15 poses for Cpd 5, 6 and 94, respectively. Additionally, 5 and 2 poses for Cpd 52 and 53, respectively, were included in this cluster. A common ligand binding orientation exists for all tariquidar-like compounds, both inactive compounds (e.g. Cpd 52 and 53) as well as active compounds (Cpd 5 and 6). Interestingly, two different ligand binding orientations were observed for Cpd 94 (tariquidar). Finally, the ligand poses with the highest score for each of the compounds in the cluster V were taken as the final binding mode for further analysis.

4.3.5 Analysis of the Poses

In general, the presence of two dimethoxy groups, which are hydrophobic substituents, in positions 6 and 7 of the tetrahydroisoquinolinylamide substructure is important for Pgp inhibition, as it has been shown in previous studies [45]. Moreover, a tertiary amine group also plays a role for many Pgp substrates and modulators [46].

Tariquidar has been extensively analysed since the discovery that the shift of the quinolone-3-carbonylamino substituent of the benzamide ring from para to meta changed the affinity of tariquidar, rendering it more active at BCRP than at Pgp. The three suggested binding sites for tariquidar include the QZ59 binding sites, Hoechst 33342 (Hoechst 33342, H-site) and rhodamine-123 (R123, R-site) [7]. Several exhaustive experiments on tariquidar putative binding modes in an inward-facing Pgp homology model revealed that the location of tariquidar is near

to the lower QZ59 ligand, but also covers the upper site partially and also the H-site [106].

4.3.5.1 Binding Modes of Cpd 5, Cpd 6 and Cpd 94 (tariquidar)

The putative binding modes of Cpd 5, Cpd 6 and Cpd 94 (tariquidar) are shown in Figure 28. In general, the ligand orientation of tariquidar is comparable to the other two active compounds, Cpd 5 and 6. Some polar amino acids were found within 4.5 Å from the molecule. They are predominantly located in TM helix 3 (Q165 and S166), TM helix 5 (Y280), TM helix 6 (Q317 and S319), TM helix 11 (Q862) and TM helix 12 (Q906 and S909). Additionally, hydrophobic interactions are found in TM helix 3 (I269), TM helix 5 (I276), TM helix 6 (F313), TM helix 11 (M865) and TM helix 12 (F899, M902 and F910).

Interestingly, a second binding mode also was observed for tariquidar, suggesting that it also binds in the lower position (See Panel B in Figure 27). While the first binding mode includes 5 poses, the second binding mode includes 22 poses. According to the docking score, the difference between the two binding modes is not significant. This indicates that there is a possibility that tariquidar simultaneously binds to different sites. This finding is in agreement with other studies, which showed that there are more than one multiple transport-active binding sites for Pgp substrates and inhibitors [97]. This observation has also been suggested by other *in silico* studies [7][107]. Additionally, it has also been proposed that Pgp has both primary and secondary sites for each of the R site and H site [108].
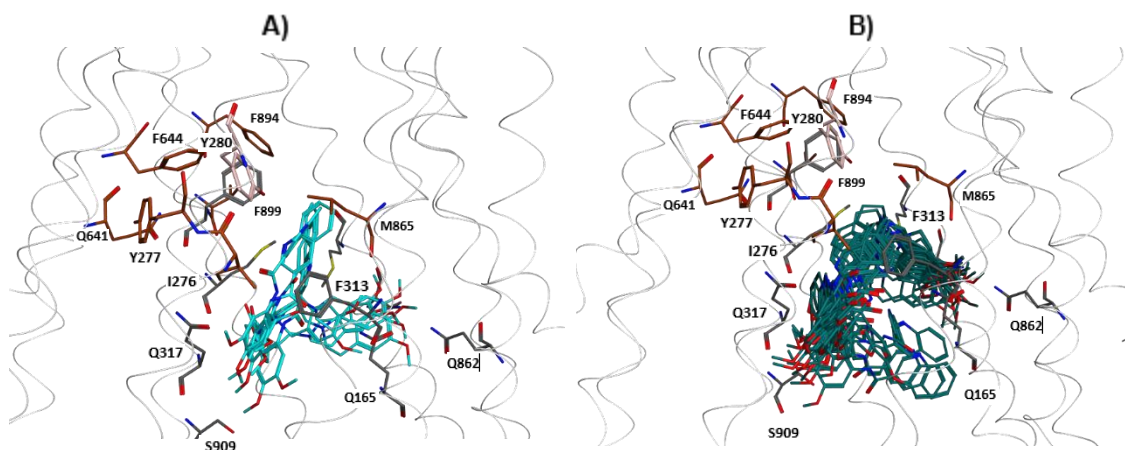
Figure 27: Two different tariquidar poses. Panel A: tariquidar pose that is comparable with Cpd 5 and Cpd 6. Panel B: another tariquidar pose which lies more to the lower position.

PLIF calculation of the two binding modes of tariquidar demonstrated that Gly905 and Ser909 are important residues for ligand-protein interaction only in the second binding mode (Table 19). Some site-directed mutagenesis studies of these residues have been published [109]. However, no studies specifically used tariquidar for the experiments. Thus, further experiments regarding this are needed to confirm this hypothesis. The putative binding mode of all active compounds, including the second tariquidar binding mode, is depicted in Figure 28.

Binding mode 1	Binding mode 2
Phe313	Phe313
Gln317	Gln317
Met865	Met865
Met902	Met902
Gln906	Gln906
	Gly905
	Ser909

Table 19: Calculated PLIF for the two binding modes of tariquidar.
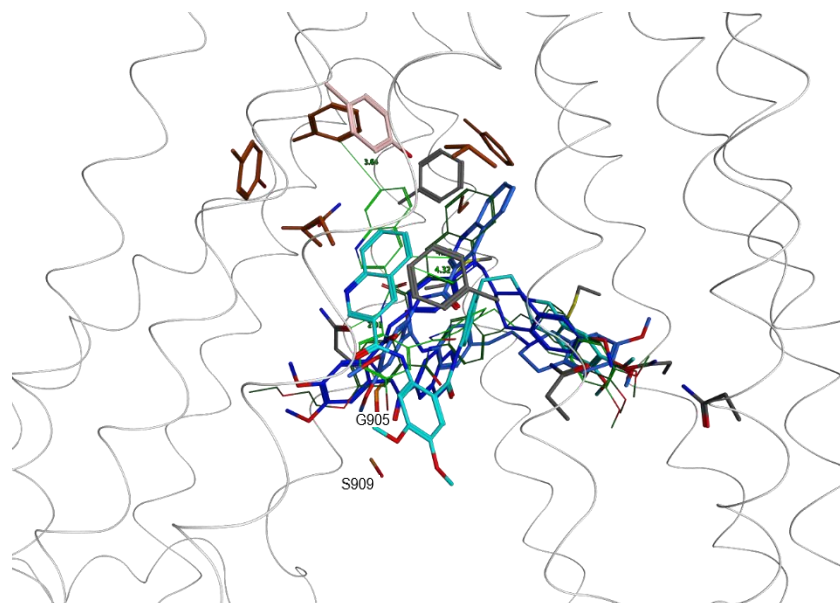
Figure 28: All active compounds, including the second tariquidar binding mode (blue). Cpd 5 (light green), Cpd 6 (dark green), first binding mode of Cpd 94 (cyan). Residues in grey are calculated by PLIF and residues in brown are important residues from mutagenesis data which lie within 4.5 Å around the ligand.

Two of the important residues demonstrated by various mutagenesis experiments were found in all binding poses. They are F343 (F313 in chain A of the model) and I306 (I276 in chain A of the model), which directly participated in interactions with the poses. Furthermore, six other residues are located within a range of 4.5 Å from the compounds, including Y307 (Y277 in chain A of the model), Q725 (Q641 in chain B of the model), F728 (F644 in chain B of the model), F978 (F894 in chain B of the model), V982 (V898 in chain B of the model), and L339 (L309 in chain B of the model).

Moreover, H-bond interactions between Y280 and Y277 with the molecule were not found. However, both residues are located within a distance of 4.5 Å from the molecule. The loss of the interactions could be explained by the fact that some residues of the refined mouse Pgp structure (PDB ID: 4M1M) are significantly shifted from the previous one (PDB ID: 3G5U), including Y277 and V898. In the case of Q641, the side-chain is flipped from the previous structure (see Figure 29).
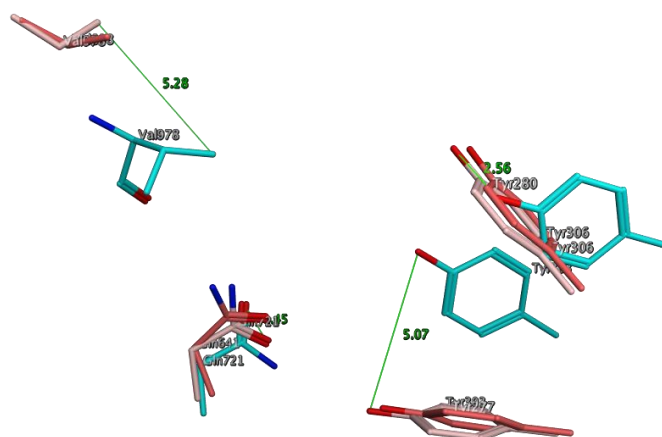
Figure 29: Comparison of some superimposed important residues for Pgp inhibition demonstrated by mutagenesis experiments. Cyan: 3G5U, light pink: 4M1M, rose: Pgp homology model. Tyr303 (Tyr277 in chain A of the model), Tyr306 (Tyr280 in chain A of the model), Gln721 (Gln641 in chain B of the model) and Val978 (Val898 in chain B of the model) for the transport of Pgp.

Despite the above mentioned fact, π–π stacking interactions between Cpd 5 with F644, Cpd 6 with F313 and Cpd 94 with F313 could still exist with a distance of 3.64Å, 3.32Å and 4.28Å, respectively (Figure 29). As shown in Figure 30, the location of the Cpd 5, 6 and 94 (tariquidar) are in the hydrophobic binding pocket.
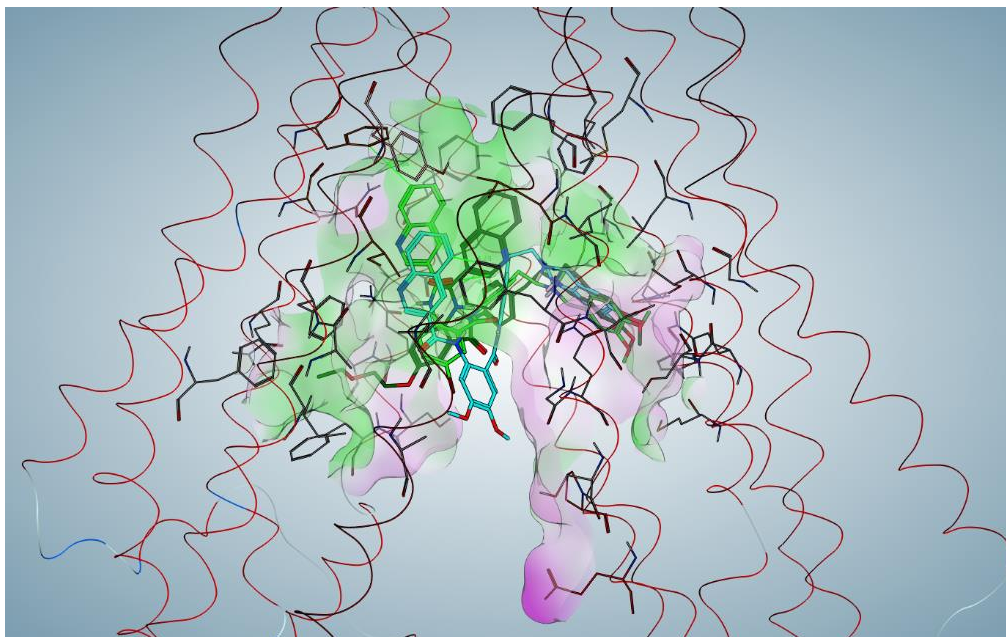
Figure 30: Surface maps of Cpd 5, 6 and 94. The hydrophilicity is depicted in pink and lipophilicity in green.

4.3.5.2 Binding Modes of Cpd 52 and Cpd 53

The shift of the quinolone-3-carbonylamino substituent of the benzamide moiety to the meta-position reduces the affinity for Pgp compared to tariquidar. Additionally, this modification changes the specificity of both Cpd 52 and 53 from Pgp to BCRP. This manner of modification at the substitution position revealed an important factor for determining the selectivity of tariquidar analogs towards Pgp and BCRP.

Moreover, it can be observed that the orientation of the quinolone-3-carbonylamino substituent of Cpd 52 and Cpd 53 is pointing towards the internal cavity of the plasma membrane, due to the meta-position of the benzamide moiety substituent. This caused the substituent to be highly exposed to water and may contribute to a lower inhibitory activity of these Pgp transporters (see Figure 31 and Figure 32). In comparison to the active compounds of 5, 6 and 94 (Figure 33 and Figure 34), the substituent of the inactive compounds is clearly exposed to water.

66
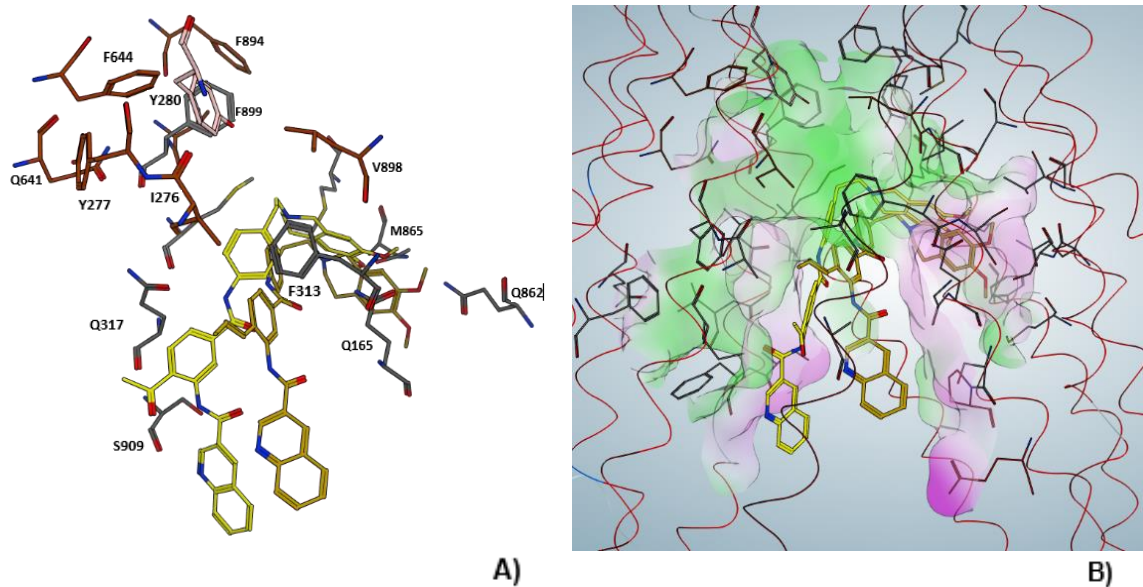
Figure 31: A) Putative binding site of Cpd 52 and 53. Residues in grey are calculated by PLIF and residues in brown are important residues from mutagenesis data which lie within 4.5Å around the ligand. B) Surface maps of Cpd 52 and 53. The hydrophilicity is depicted in pink and lipophilicity in green.
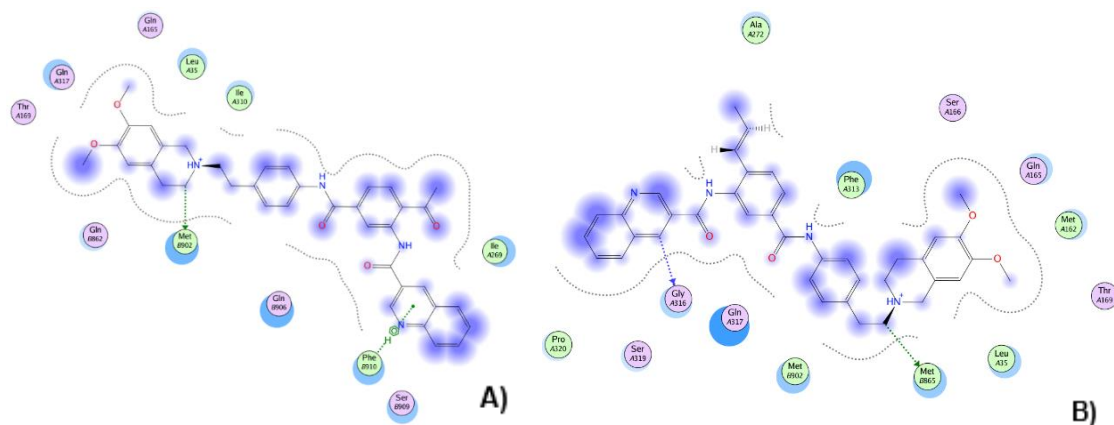
Figure 32: 2D visualization of the inactive compounds in the binding region. High exposure of the quinolone-3-carbonylamino substituent of the benzamide moiety to water. A) Cpd 52. B) Cpd 53.
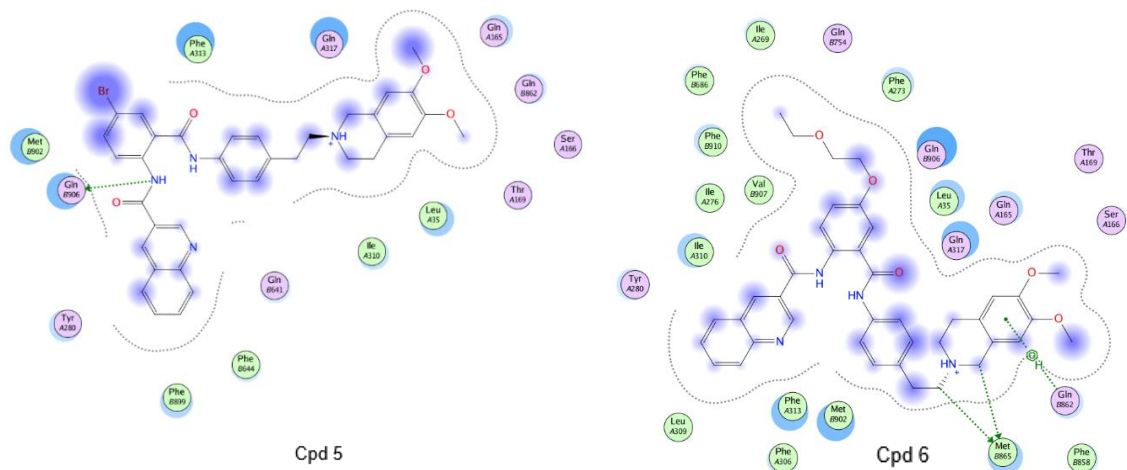
Figure 33: 2D visualization of the active compounds in the binding region. Small exposure of the quinolone-3-carbonylamino substituent of the benzamide moiety to water.
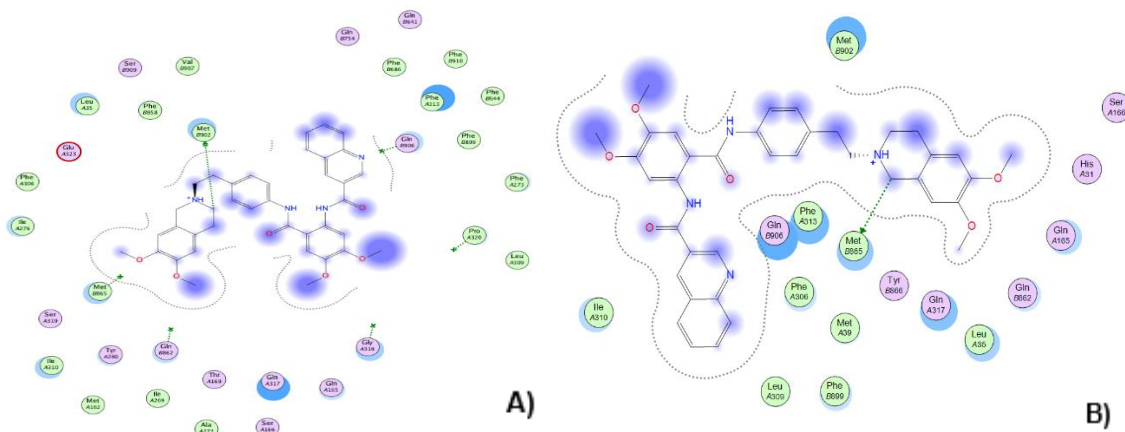
Figure 34: 2D visualization of Cpd 94 (tariquidar) in the binding region. Small exposure of the quinolone-3-carbonylamino substituent of the benzamide moiety to water. A) Binding mode 1. B) Binding mode 2.

4.4 Summary and Conclusion

In summary, the docking experiments in this study focused on identification of the molecular basis of Pgp inhibition by tariquidar-like modulators. The docking studies were carried out with the inward-facing structural model of human Pgp based on the refined crystal structure of mouse Pgp. First, five ligands that represent the most active and least active compounds in the data set were selected for the study. Both binding sites defined by the *Sitefinder* tool in MOE as well as putative binding sites from various studies were taken into consideration. After the docking protocol, subsequent analysis of the obtained orientation of the ligands included hierarchical agglomerative clustering, protein ligand interaction, and consensus scoring to rank the ligand poses. The method used for clustering was Common Scaffold Clustering (CSC). This implies that compounds are supposed to have a common binding mode when a common scaffold exists. Finally, further analysis was based on the "protein ligand interaction fingerprint" (PLIF). The residues that are described to play an important role in protein-ligand interactions were investigated. The resulting 124 poses were visually analysed to distinguish the different orientations between active and inactive compounds.

Results of the structure-based approach propose that the shift of the quinolone-3-carbonylamino from para to meta position of the benzamide ring leads to limitation of the sterical freedom of the benzamide moiety. This may result in a lower inhibitory activity of the compounds towards Pgp. Another reason for the low activity may be the high exposure of the previously mentioned substituent to water.

However, due to this modification, the specificity of both meta-substituted quinolone-3-carbonylamino analogs changed from Pgp to BCRP. This revealed a key factor for profiling the structural determinants for the selectivity of Pgp and BCRP of tariquidar-like modulators. This is in agreement with the observed

selectivity of meta-substituted tariquidar derivatives for ABCG2 transporters (see Appendix: Figure 35).

Other docking experiments with tariquidar analogous revealed some putative binding modes with several critical amino acid residues [105]. These results include Tyr307, Gln725, and Val982, which are in agreement with mutagenesis experiments (see section 4.3.2). Additionally, other *in silico* experiments with propafenone derivatives also demonstrated a H-bond interaction of residue Tyr310 of the protein with the molecule [73]. These docking studies were carried out with a human Pgp homology model based on the old X-ray structures of murine Pgp (PDB ID: 3G5U). The structures show a large internal cavity open to the cytoplasmic surface and the membrane inner leaflet, with a high degree of separation between the two NBDs [68].

It is highly interesting that different results were obtained in this study. Apart from different methods that were applied, it is important to note that the refined structures of murine Pgp (PDB ID: 4M1M) was chosen to build the homology model. Both might explain the different results compared with the previous *in silico* studies.

With regard to the methods, it may be that the applied methods were not appropriate for analysing the interaction, as was previously thought. The selection of methods was based on previous successful studies [73]. Method selection has to be considered in detail for further studies, especially those that use the new X-ray structures.

In future studies, the number of conformations per compound should be increased to cover all binding pose possibilities and further analysis of important residues for the transport should consider various rotamers of charged or polar amino acid residues to reveal more possible interactions of protein residues with the molecule. Furthermore, the unprocessed complexes could be energetically minimized e.g. using LigX.

Another interesting experiment would be to alternatively use GlideScore XP for docking score function. As it is a more precise and strict method, it is computationally more expensive. Therefore, one should apply the SP method first, and the subsequent docking of the top-scoring ligands can be done using XP mode. This will provide higher improvements to the scoring and detection of various interactions, including hydrogen bonds, π–cation and π–π stacking. Subsequently, molecular dynamic (MD) studies for this particular binding mode can be done to reveal the relative binding energies. Moreover, the molecular interactions and molecular features of active and inactive compounds can also be understood in a better way.

Another suggestion would be MD studies without ligands to see how good the homology model is.

Alternatively, one could apply the induced fit docking protocol of the Schrödinger Suite [110] to increase the protein flexibility and thus to obtain more binding pose possibilities. Thereby, the specificity of Pgp and the distinction between active and inactive compounds may be understood to a greater degree by the molecular features. Moreover, if a structure of BCRP would be available, docking of the same compounds in it would allow gaining more clues regarding the specificity of the compounds.

In conclusion, docking studies provide some hypotheses to better understand the structure and function of P-glycoprotein. However, the approximate proposed position of ligand binding only partially corresponded with the suggested mechanism of Pgp inhibition, as well as with reports from cross-linking and site-directed mutagenesis experiments. More exhaustive docking studies and mutagenesis experiments are necessary to confirm the putative binding modes found in this study. The result of this study at least strengthened the importance of residues Gly905 and Ser909 for tariquidar binding to Pgp.

5. Conclusions and Outlook

Despite the fact that a large number of exhaustive and thorough experiments have been carried out since the crystallographic murine Pgp structure was solved in 2009, the big question regarding the polyspecificity of Pgp remains elusive. One still cannot explain the molecular basis of the ligand-protein interactions. Once the key factors regarding mechanism of Pgp drug efflux function are revealed, the efficacy of many cancer therapies would be dramatically increased. This is because of the vital role Pgp plays in cancer phenotypes. Various ligand-based and structure-based studies have been carried out in order to have a better comprehension of which molecular properties a ligand should have to be able to inhibit the function of Pgp. Some highly potent third generation Pgp or BCRP modulators have been discovered, holding promise to scientists and physicians that the long awaited fundamental questions (i.e. how Pgp attracts its substrates to bind or which structural features are responsible for substrate and inhibitor affinity) might finally be answered.

This work focused on finding the selectivity profiling of Pgp and BCRP modulators analogs to tariquidar. It comprises both ligand and structure-based approaches to gain a deeper insight about the molecular basis of ligand-protein interaction.

The 2D-QSAR analysis demonstrated the importance of the hydrophobic surface area and the number of hydrophobic atoms for Pgp and BCRP inhibitory activity, respectively. However, the number of hydrophobic atoms was demonstrated to have a negative effect on BCRP inhibition. This indicates that hydrophobic atoms may contribute binding to the protein, but are not the key factor for transporter inhibition.
Several 2D-QSAR models were also constructed in WEKA for both Pgp and BCRP datasets. The methods that were applied were linear regression function with embedded feature selection. By using various methods to build the models, the result don't clearly indicate the importance of hydrophobicity.

The next approach in the ligand-based studies was the fragmentation approach, whose main objective is to exploit local information which may not be captured by using the whole structure. In other words, this approach uses local features instead of global ones to identify important molecular features responsible for biological activity. The results showed that the number of hydrogen bond acceptor atoms of the first fragment, e.g. tetrahydroisoquinoline, and the number of hydrogen bond donor atoms of the core fragment, are the most contributing descriptors for biological activity at BCRP.

In regard to the dataset studied in this work, one should note that between the two groups (i.e. active and inactive compounds) selected for docking experiments, only approximately two to three orders of magnitude difference in activity exists. Hence, it cannot be assumed that the important descriptors found can be specifically taken as key factors to distinguish active from inactive compounds.

The docking experiments revealed a very important factor for determining the selectivity of tariquidar analogous towards Pgp and BCRP. This is in agreement with the data of the research group that demonstrated that the shift of the quinolone-3-carbonylamino substituent of the benzamide ring from para to meta position changed the affinity of tariquidar from Pgp to BCRP [48].

Moreover, two tariquidar poses suggested that tariquidar may change its conformation while entering the protein without losing its affinity, and that multiple transport-active binding sites might indeed exist, as suggested in previous studies [97]. Two residues Gly905 and Ser909 were found to be essential for tariquidar-Pgp interaction. Both structure-based and ligand-based studies in this work suggested that the hydrophobicity of ligands does not play an important role in Pgp inhibition, but is only important for binding to Pgp.

The results of this work serve as a contribution to reveal the predominant factor in determining Pgp/BCRP-ligand interactions for a set of tariquidar analogs. This has implications for future studies on targeted cancer therapies. Once the missing

piece of the puzzle is solved, it may provide a therapeutic benefit for cancer patients.

6. Bibliography

[1]   S. Wilkens, "Structure and mechanism of ABC transporters," *F1000Prime Rep.*, vol. 7, Feb. 2015.

[2]   M. Dean and M. Dean, *The Human ATP-Binding Cassette (ABC) Transporter Superfamily*. National Center for Biotechnology Information (US), 2002.

[3]   C. F. Higgins and M. M. Gottesman, "Is the multidrug transporter a flippase?," *Trends Biochem. Sci.*, vol. 17, no. 1, pp. 18–21, Jan. 1992.

[4]   Y. Raviv, H. B. Pollard, E. P. Bruggemann, I. Pastan, and M. M. Gottesman, "Photosensitized labeling of a functional multidrug transporter in living drug-resistant tumor cells," *J. Biol. Chem.*, vol. 265, no. 7, pp. 3975–3980, Mar. 1990.

[5]   F. J. Sharom, "Complex interplay between the P-glycoprotein multidrug efflux pump and the membrane: its role in modulating protein function," *Pharmacol. Anti-Cancer Drugs*, vol. 4, p. 41, 2014.

[6]   G. A. Altenberg, C. G. Vanoye, J. K. Horton, and L. Reuss, "Unidirectional fluxes of rhodamine 123 in multidrug-resistant cells: evidence against direct drug extrusion from the plasma membrane.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 91, no. 11, pp. 4654–4657, May 1994.

[7]   "Reversing cancer multidrug resistance: insights into the efflux by ABC transports from in silico studies - Ferreira - 2014 - Wiley Interdisciplinary Reviews: Computational Molecular Science - Wiley Online Library." [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/wcms.1196/abstract. [Accessed: 10-Jan-2016].

[8]   "Enzyme Explorer Drug Metabolism Tech Review," *Sigma-Aldrich*. [Online]. Available: http://www.sigmaaldrich.com/life-science/metabolomics/enzyme-explorer/cell-signaling-enzymes/xenobiotics/drug-metabolism-tech-review.html. [Accessed: 10-Nov-2015].

[9]   E. M. Leslie, R. G. Deeley, and S. P. C. Cole, "Multidrug resistance proteins: role of P-glycoprotein, MRP1, MRP2, and BCRP (ABCG2) in tissue defense," *Toxicol. Appl. Pharmacol.*, vol. 204, no. 3, pp. 216–237, May 2005.

[10]   S. Kunjachan, B. Rychlik, G. Storm, F. Kiessling, and T. Lammers, "Multidrug Resistance: Physiological Principles and Nanomedical Solutions," *Adv. Drug Deliv. Rev.*, vol. 65, no. 0, pp. 1852–1865, Nov. 2013.

[11]   G. Szakács, A. Váradi, C. Özvegy-Laczka, and B. Sarkadi, "The role of ABC transporters in drug absorption, distribution, metabolism, excretion and toxicity (ADME–Tox)," *Drug Discov. Today*, vol. 13, no. 9–10, pp. 379–393, May 2008.

[12]   Z. Ni, Z. Bikadi, M. F. Rosenberg, and Q. Mao, "Structure and function of the human breast cancer resistance protein (BCRP/ABCG2)," *Curr. Drug Metab.*, vol. 11, no. 7, pp. 603–617, Sep. 2010.

[13]   R. L. Juliano and V. A. Ling, "Juliano, R. L. & Ling, V. A. A surface glycoprotein modulating drug permeability in Chinese hamster ovary cell

mutants. Biochem. Biophys. Acta. 455, 152-162," *Biochim. Biophys. Acta*, vol. 455, no. 1, pp. 152–62, Dec. 1976.

[14]   L. A. Doyle, W. Yang, L. V. Abruzzo, T. Krogmann, Y. Gao, A. K. Rishi, and D. D. Ross, "A multidrug resistance transporter from human MCF-7 breast cancer cells," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 95, no. 26, pp. 15665–15670, Dec. 1998.

[15]   R. Allikmets, L. M. Schriml, A. Hutchinson, V. Romano-Spica, and M. Dean, "A human placenta-specific ATP-binding cassette gene (ABCP) on chromosome 4q22 that is involved in multidrug resistance," *Cancer Res.*, vol. 58, no. 23, pp. 5337–5339, Dec. 1998.

[16]   K. Miyake, L. Mickley, T. Litman, Z. Zhan, R. Robey, B. Cristensen, M. Brangi, L. Greenberger, M. Dean, T. Fojo, and S. E. Bates, "Molecular cloning of cDNAs which are highly overexpressed in mitoxantrone-resistant cells: demonstration of homology to ABC transport genes," *Cancer Res.*, vol. 59, no. 1, pp. 8–13, Jan. 1999.

[17]   T. Nakanishi and D. D. Ross, "Breast cancer resistance protein (BCRP/ABCG2): its role in multidrug resistance and regulation of its gene expression," *Chin. J. Cancer*, vol. 31, no. 2, pp. 73–99, Feb. 2012.

[18]   S. K. Rabindran, H. He, M. Singh, E. Brown, K. I. Collins, T. Annable, and L. M. Greenberger, "Reversal of a novel multidrug resistance mechanism in human colon carcinoma cells by fumitremorgin C," *Cancer Res.*, vol. 58, no. 24, pp. 5850–5858, Dec. 1998.

[19]   J. D. Allen, A. van Loevezijn, J. M. Lakhai, M. van der Valk, O. van Tellingen, G. Reid, J. H. M. Schellens, G.-J. Koomen, and A. H. Schinkel, "Potent and specific inhibition of the breast cancer resistance protein multidrug transporter in vitro and in mouse intestine by a novel analogue of fumitremorgin C," *Mol. Cancer Ther.*, vol. 1, no. 6, pp. 417–425, Apr. 2002.

[20]   "Potent and selective inhibitors of breast cancer resistance protein (ABCG2) derived from the p-glycoprotein (ABCB1) modulator tariquidar. - PubMed - NCBI." [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19170519. [Accessed: 17-Nov-2015].

[21]   C. Ozvegy-Laczka, T. Hegedus, G. Várady, O. Ujhelly, J. D. Schuetz, A. Váradi, G. Kéri, L. Orfi, K. Német, and B. Sarkadi, "High-affinity interaction of tyrosine kinase inhibitors with the ABCG2 multidrug transporter," *Mol. Pharmacol.*, vol. 65, no. 6, pp. 1485–1495, Jun. 2004.

[22]   M. Wierdl, A. Wall, C. L. Morton, J. Sampath, M. K. Danks, J. D. Schuetz, and P. M. Potter, "Carboxylesterase-mediated sensitization of human tumor cells to CPT-11 cannot override ABCG2-mediated drug resistance," *Mol. Pharmacol.*, vol. 64, no. 2, pp. 279–288, Aug. 2003.

[23]   J. Xu, Y. Liu, Y. Yang, S. Bates, and J.-T. Zhang, "Characterization of oligomeric human half-ABC transporter ATP-binding cassette G2," *J. Biol. Chem.*, vol. 279, no. 19, pp. 19781–19789, May 2004.

[24]   W. Mo and J.-T. Zhang, "Human ABCG2: structure, function, and its role in multidrug resistance," *Int. J. Biochem. Mol. Biol.*, vol. 3, no. 1, pp. 1–27, Mar. 2011.

[25]   G. D. Leonard, "The Role of ABC Transporters in Clinical Practice," *The Oncologist*, vol. 8, no. 5, pp. 411–424, Oct. 2003.

[26]   P. D. W. Eckford and F. J. Sharom, "ABC Efflux Pump-Based Resistance to Chemotherapy Drugs," *Chem. Rev.*, vol. 109, no. 7, pp. 2989–3011, Jul. 2009.

[27]   P. A. Ashworth, J. E. Brumwell, A. J. Folkes, S. Hunjan, L. M. Maximen, M. J. Roe, H. Ryder, J. T. Sanderson, and S. Williams, "Anthranilic acid derivatives as multi drug resistance modulators," WO1998017648 A1, 30-Apr-1998.

[28]   M. Roe, A. Folkes, P. Ashworth, J. Brumwell, L. Chima, S. Hunjan, I. Pretswell, W. Dangerfield, H. Ryder, and P. Charlton, "Reversal of P-glycoprotein mediated multidrug resistance by novel anthranilamide derivatives," *Bioorg. Med. Chem. Lett.*, vol. 9, no. 4, pp. 595–600, Feb. 1999.

[29]   J. J. Starling, R. L. Shepard, J. Cao, K. L. Law, B. H. Norman, J. S. Kroin, W. J. Ehlhardt, T. M. Baughman, M. A. Winter, M. G. Bell, C. Shih, J. Gruber, W. F. Elmquist, and A. H. Dantzig, "Pharmacological characterization of LY335979: a potent cyclopropyldibenzosuberane modulator of P-glycoprotein," *Adv. Enzyme Regul.*, vol. 37, pp. 335–347, 1997.

[30]   A. H. Dantzig, R. L. Shepard, K. L. Law, L. Tabas, S. Pratt, J. S. Gillespie, S. N. Binkley, M. T. Kuhfeld, J. J. Starling, and S. A. Wrighton, "Selectivity of the multidrug resistance modulator, LY335979, for P-glycoprotein and effect on cytochrome P-450 activities," *J. Pharmacol. Exp. Ther.*, vol. 290, no. 2, pp. 854–862, Aug. 1999.

[31]   L. van Zuylen, K. Nooter, A. Sparreboom, and J. Verweij, "Development of multidrug-resistance convertors: sense or nonsense?," *Invest. New Drugs*, vol. 18, no. 3, pp. 205–220, Aug. 2000.

[32]   M. J. Newman, J. C. Rodarte, K. D. Benbatoul, S. J. Romano, C. Zhang, S. Krane, E. J. Moran, R. T. Uyeda, R. Dixon, E. S. Guns, and L. D. Mayer, "Discovery and characterization of OC144-093, a novel inhibitor of P-glycoprotein-mediated multidrug resistance," *Cancer Res.*, vol. 60, no. 11, pp. 2964–2972, Jun. 2000.

[33]   F. Hyafil, C. Vergely, P. Du Vignaud, and T. Grand-Perret, "In vitro and in vivo reversal of multidrug resistance by GF120918, an acridonecarboxamide derivative," *Cancer Res.*, vol. 53, no. 19, pp. 4595–4602, Oct. 1993.

[34]   S. P. Cole, G. Bhardwaj, J. H. Gerlach, J. E. Mackie, C. E. Grant, K. C. Almquist, A. J. Stewart, E. U. Kurz, A. M. Duncan, and R. G. Deeley, "Overexpression of a transporter gene in a multidrug-resistant human lung cancer cell line," *Science*, vol. 258, no. 5088, pp. 1650–1654, Dec. 1992.

[35]   C. Martin, G. Berridge, P. Mistry, C. Higgins, P. Charlton, and R. Callaghan, "The molecular interaction of the high affinity reversal agent XR9576 with P-glycoprotein," *Br. J. Pharmacol.*, vol. 128, no. 2, pp. 403–411, Sep. 1999.

[36]   T. Tsuruo, H. Iida, M. Yamashiro, S. Tsukagoshi, and Y. Sakurai, "Enhancement of vincristine- and adriamycin-induced cytotoxicity by verapamil in P388 leukemia and its sublines resistant to vincristine and adriamycin," *Biochem. Pharmacol.*, vol. 31, no. 19, pp. 3138–3140, Oct. 1982.

[37]   I. K. Pajeva, C. Globisch, and M. Wiese, "Comparison of the inward- and outward-open homology models and ligand binding of human P-glycoprotein," *FEBS J.*, vol. 276, no. 23, pp. 7016–7026, Dec. 2009.

[38]   H. J. Broxterman, H. M. Pinedo, C. M. Kuiper, L. C. Kaptein, G. J. Schuurhuis, and J. Lankelma, "Induction by verapamil of a rapid increase in ATP consumption in multidrug-resistant tumor cells," *FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol.*, vol. 2, no. 7, pp. 2278–2282, Apr. 1988.

[39]   F. Tiberghien and F. Loor, "Ranking of P-glycoprotein substrates and inhibitors by a calcein-AM fluorometry screening assay," *Anticancer. Drugs*, vol. 7, no. 5, pp. 568–578, Jul. 1996.

[40]   D. Boesch, C. Gavériaux, B. Jachez, A. Pourtier-Manzanedo, P. Bollinger, and F. Loor, "In vivo circumvention of P-glycoprotein-mediated multidrug resistance of tumor cells with SDZ PSC 833," *Cancer Res.*, vol. 51, no. 16, pp. 4226–4233, Aug. 1991.

[41]   A. Palmeira, E. Sousa, M. H. Vasconcelos, and M. M. Pinto, "Three decades of P-gp inhibitors: skimming through several generations and scaffolds," *Curr. Med. Chem.*, vol. 19, no. 13, pp. 1946–2025, 2012.

[42]   P. Mistry, A. J. Stewart, W. Dangerfield, S. Okiji, C. Liddle, D. Bootle, J. A. Plumb, D. Templeton, and P. Charlton, "In vitro and in vivo reversal of P-glycoprotein-mediated multidrug resistance by a novel potent modulator, XR9576," *Cancer Res.*, vol. 61, no. 2, pp. 749–758, Jan. 2001.

[43]   H. Thomas and H. M. Coley, "Overcoming multidrug resistance in cancer: an update on the clinical strategy of inhibiting p-glycoprotein," *Cancer Control J. Moffitt Cancer Cent.*, vol. 10, no. 2, pp. 159–165, Apr. 2003.

[44]   A. Tamaki, C. Ierano, G. Szakacs, R. W. Robey, and S. E. Bates, "The controversial role of ABC transporters in clinical oncology," *Essays Biochem.*, vol. 50, no. 1, pp. 209–232, Sep. 2011.

[45]   A. Pick, H. Müller, and M. Wiese, "Structure-activity relationships of new inhibitors of breast cancer resistance protein (ABCG2)," *Bioorg. Med. Chem.*, vol. 16, no. 17, pp. 8224–8236, Sep. 2008.

[46]   A. Seelig and E. Landwojtowicz, "Structure-activity relationship of P-glycoprotein substrates and modifiers," *Eur. J. Pharm. Sci. Off. J. Eur. Fed. Pharm. Sci.*, vol. 12, no. 1, pp. 31–40, Nov. 2000.

[47]   C. O. Puentes, P. Höcherl, M. Kühnle, S. Bauer, K. Bürger, G. Bernhardt, A. Buschauer, and B. König, "Solid phase synthesis of tariquidar-related modulators of ABC transporters preferring breast cancer resistance protein (ABCG2)," *Bioorg. Med. Chem. Lett.*, vol. 21, no. 12, pp. 3654–3657, Jun. 2011.

[48]   M. Kühnle, M. Egger, C. Müller, A. Mahringer, G. Bernhardt, G. Fricker, B. König, and A. Buschauer, "Potent and selective inhibitors of breast cancer resistance protein (ABCG2) derived from the p-glycoprotein (ABCB1) modulator tariquidar," *J. Med. Chem.*, vol. 52, no. 4, pp. 1190–1197, Feb. 2009.

[49]   P. Kannan, S. Telu, S. Shukla, S. V. Ambudkar, V. W. Pike, C. Halldin, M. M. Gottesman, R. B. Innis, and M. D. Hall, "The 'specific' P-glycoprotein inhibitor Tariquidar is also a substrate and an inhibitor for breast cancer

resistance protein (BCRP/ABCG2)," *ACS Chem. Neurosci.*, vol. 2, no. 2, pp. 82–89, Feb. 2011.

[50] I. K. Pajeva and M. Wiese, "Structure–Activity Relationships of Tariquidar Analogs as Multidrug Resistance Modulators," *AAPS J.*, vol. 11, no. 3, pp. 435–444, Jun. 2009.

[51] "Tariquidar Analogues: Synthesis by CuI-Catalysed N/O–Aryl Coupling and Inhibitory Activity against the ABCB1 Transporter - Egger - 2007 - European Journal of Organic Chemistry - Wiley Online Library." [Online]. Available:
http://onlinelibrary.wiley.com/doi/10.1002/ejoc.200700142/abstract. [Accessed: 12-Dec-2015].

[52] C. Hansch, D. Hoekman, A. Leo, L. Zhang, and P. Li, "The expanding role of quantitative structure-activity relationships (QSAR) in toxicology," *Toxicol. Lett.*, vol. 79, no. 1–3, pp. 45–53, Sep. 1995.

[53] "V05_QSAR - V05-QSAR.pdf." .

[54] "QuaSAR-Descriptor."   [Online].   Available: https://www.chemcomp.com/journal/descr.htm. [Accessed: 08-Jan-2016].

[55] P. Ertl, B. Rohde, and P. Selzer, "Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties," *J. Med. Chem.*, vol. 43, no. 20, pp. 3714–3717, Oct. 2000.

[56] M. Hubensack, C. Müller, P. Höcherl, S. Fellner, T. Spruss, G. Bernhardt, and A. Buschauer, "Effect of the ABCB1 modulators elacridar and tariquidar on the distribution of paclitaxel in nude mice," *J. Cancer Res. Clin. Oncol.*, vol. 134, no. 5, pp. 597–607, May 2008.

[57] "Benzanilide–Biphenyl Replacement: A Bioisosteric Approach to Quinoline Carboxamide-Type   ABCG2   Modulators."   [Online].   Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4027546/. [Accessed: 18-Nov-2015].

[58] C. W. Yap, "PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints," *J. Comput. Chem.*, vol. 32, no. 7, pp. 1466–1474, May 2011.

[59] "Journal of Cheminformatics | Full text | jCompoundMapper: An open source Java library and command-line tool for chemical fingerprints." [Online]. Available: http://www.jcheminf.com/content/3/1/3. [Accessed: 18-Nov-2015].

[60] N. Brown, "Chemoinformatics—an Introduction for Computer Scientists," *ACM Comput Surv*, vol. 41, no. 2, pp. 8:1–8:38, Feb. 2009.

[61] D. Rogers and M. Hahn, "Extended-Connectivity Fingerprints," *J. Chem. Inf. Model.*, vol. 50, no. 5, pp. 742–754, May 2010.

[62] I. S. Helland, "On the structure of partial least squares regression," *Commun. Stat. - Simul. Comput.*, Jun. 2007.

[63] "Abdi-PLS-pretty.pdf." .

[64] "Microsoft Word - 1 - 511_full.pdf." .

[65] I. Jabeen, P. Wetwitayaklung, F. Klepsch, Z. Parveen, P. Chiba, and G. F. Ecker, "Probing the stereoselectivity of P-glycoprotein—synthesis, biological

activity and ligand docking studies of a set of enantiopure benzopyrano[3,4-b][1,4]oxazines," *Chem Commun*, vol. 47, no. 9, pp. 2586–2588, Feb. 2011.

[66] A. Golbraikh and A. Tropsha, "Beware of q2!," *J. Mol. Graph. Model.*, vol. 20, no. 4, pp. 269–276, Jan. 2002.

[67] F. Klepsch, I. Jabeen, P. Chiba, and G. F. Ecker, "Pharmacoinformatic approaches to design natural product type ligands of ABC-transporters," *Curr. Pharm. Des.*, vol. 16, no. 15, pp. 1742–1752, May 2010.

[68] S. G. Aller, J. Yu, A. Ward, Y. Weng, S. Chittaboina, R. Zhuo, P. M. Harrell, Y. T. Trinh, Q. Zhang, I. L. Urbatsch, and G. Chang, "Structure of P-glycoprotein reveals a molecular basis for poly-specific drug binding," *Science*, vol. 323, no. 5922, pp. 1718–1722, Mar. 2009.

[69] J. Li, K. F. Jaimes, and S. G. Aller, "Refined structures of mouse P-glycoprotein," *Protein Sci. Publ. Protein Soc.*, vol. 23, no. 1, pp. 34–46, Jan. 2014.

[70] A. L. C. Bazzan, *Advances in Bioinformatics and Computational Biology: Third Brazilian Symposium on Bioinformatics, BSB 2008, Sao Paulo, Brazil, August 28-30, 2008, Proceedings.* Springer Science & Business Media, 2008.

[71] C. Oshiro, E. K. Bradley, J. Eksterowicz, E. Evensen, M. L. Lamb, J. K. Lanctot, S. Putta, R. Stanton, and P. D. J. Grootenhuis, "Performance of 3D-database molecular docking studies into homology models," *J. Med. Chem.*, vol. 47, no. 3, pp. 764–767, Jan. 2004.

[72] "Comparative protein structure modeling of genes and genomes. - PubMed - NCBI." [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/10940251. [Accessed: 07-Dec-2015].

[73] F. Klepsch, P. Chiba, and G. F. Ecker, "Exhaustive Sampling of Docking Poses Reveals Binding Hypotheses for Propafenone Type Inhibitors of P-Glycoprotein," *PLoS Comput. Biol.*, vol. 7, no. 5, May 2011.

[74] I. Jabeen, K. Pleban, U. Rinner, P. Chiba, and G. F. Ecker, "Structure–Activity Relationships, Ligand Efficiency, and Lipophilic Efficiency Profiles of Benzophenone-Type Inhibitors of the Multidrug Transporter P-Glycoprotein," *J. Med. Chem.*, vol. 55, no. 7, pp. 3261–3273, Apr. 2012.

[75] "Comparative Protein Structure Modeling Using Modeller." [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4186674/. [Accessed: 04-Jan-2016].

[76] "Laskowski et al., 1993." [Online]. Available: http://www.ebi.ac.uk/thornton-srv/software/PROCHECK/manual/paper1.html. [Accessed: 07-Dec-2015].

[77] "Stereochemistry of polypeptide chain configurations." [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0022283663800236. [Accessed: 04-Jan-2016].

[78] "CORINA Classic - Fast Generation of High-Quality 3D Molecular Models | Inspiring Chemical Discovery." [Online]. Available: https://www.molecular-networks.com/products/corina. [Accessed: 21-Nov-2015].

[79] "chemicalize.org." [Online]. Available: http://www.chemicalize.org/. [Accessed: 20-Nov-2015].

[80] R. A. Friesner, J. L. Banks, R. B. Murphy, T. A. Halgren, J. J. Klicic, D. T. Mainz, M. P. Repasky, E. H. Knoll, M. Shelley, J. K. Perry, D. E. Shaw, P.

Francis, and P. S. Shenkin, "Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy," *J. Med. Chem.*, vol. 47, no. 7, pp. 1739–1749, Mar. 2004.

[81]   "Dissertation_Peter_Hoecherl.pdf." .

[82]   V. B. Chen, W. B. Arendall, J. J. Headd, D. A. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson, and D. C. Richardson, "MolProbity: all-atom structure validation for macromolecular crystallography," *Acta Crystallogr. D Biol. Crystallogr.*, vol. 66, no. Pt 1, pp. 12–21, Jan. 2010.

[83]   "Glide User Manual - gli55_user_manual.pdf." .

[84]   "SPC_v13.pdf." .

[85]   "1026.full.pdf." .

[86]   A. Jurik, F. Klepsch, and B. Zdrazil, "Molecular Modeling and Simulation of Membrane Transport Proteins," in *Medicinal Chemistry and Drug Design*, D. Ekinci, Ed. InTech, 2012.

[87]   "RSC_CP_C0CP00151A 1..10 - 2010-PCCP-Scoring_functions.pdf." .

[88]   H. J. Böhm, "The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure," *J. Comput. Aided Mol. Des.*, vol. 8, no. 3, pp. 243–256, Jun. 1994.

[89]   H. J. Böhm, "Prediction of binding constants of protein ligands: a fast method for the prioritization of hits obtained from de novo design or 3D database search programs," *J. Comput. Aided Mol. Des.*, vol. 12, no. 4, pp. 309–323, Jul. 1998.

[90]   "Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein−Ligand Complexes - Journal of Medicinal Chemistry (ACS Publications)." [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/jm051256o. [Accessed: 22-Dec-2015].

[91]   P. S. Charifson, J. J. Corkery, M. A. Murcko, and W. P. Walters, "Consensus Scoring: A Method for Obtaining Improved Hit Rates from Docking Databases of Three-Dimensional Structures into Proteins," *J. Med. Chem.*, vol. 42, no. 25, pp. 5100–5109, Dec. 1999.

[92]   R. D. Smith, J. B. Dunbar, P. M.-U. Ung, E. X. Esposito, C.-Y. Yang, S. Wang, and H. A. Carlson, "CSAR Benchmark Exercise of 2010: Combined Evaluation Across All Submitted Scoring Functions," *J. Chem. Inf. Model.*, vol. 51, no. 9, pp. 2115–2131, Sep. 2011.

[93]   "Wiley: Computational Drug Design: A Guide for Computational and Medicinal Chemists - D. C. Young." [Online]. Available: http://eu.wiley.com/WileyCDA/WileyTitle/productCd-047012685X.html. [Accessed: 13-Dec-2015].

[94]   C. F. Higgins, "Multiple molecular mechanisms for multidrug resistance transporters," *Nature*, vol. 446, no. 7137, pp. 749–757, Apr. 2007.

[95]   C. A. McDevitt and R. Callaghan, "How can we best use structural information on P-glycoprotein to design inhibitors?," *Pharmacol. Ther.*, vol. 113, no. 2, pp. 429–441, Feb. 2007.

[96]   T. W. Loo and D. M. Clarke, "Cross-linking of human multidrug resistance P-glycoprotein by the substrate, tris-(2-maleimidoethyl)amine, is altered by

ATP hydrolysis. Evidence for rotation of a transmembrane helix," *J. Biol. Chem.*, vol. 276, no. 34, pp. 31800–31805, Aug. 2001.

[97] E. E. Chufan, K. Kapoor, H.-M. Sim, S. Singh, T. T. Talele, S. R. Durell, and S. V. Ambudkar, "Multiple transport-active binding sites are available for a single substrate on human P-glycoprotein (ABCB1)," *PloS One*, vol. 8, no. 12, p. e82463, 2013.

[98] T. W. Loo, M. C. Bartlett, and D. M. Clarke, "Simultaneous Binding of Two Different Drugs in the Binding Pocket of the Human Multidrug Resistance P-glycoprotein," *J. Biol. Chem.*, vol. 278, no. 41, pp. 39706–39710, Oct. 2003.

[99] A. Rothnie, J. Storm, R. McMahon, A. Taylor, I. D. Kerr, and R. Callaghan, "The coupling mechanism of P-glycoprotein involves residue L339 in the sixth membrane spanning segment," *FEBS Lett.*, vol. 579, no. 18, pp. 3984–3990, Jul. 2005.

[100] T. W. Loo and D. M. Clarke, "Defining the drug-binding site in the human multidrug resistance P-glycoprotein using a methanethiosulfonate analog of verapamil, MTS-verapamil," *J. Biol. Chem.*, vol. 276, no. 18, pp. 14972–14979, May 2001.

[101] T. W. Loo, M. C. Bartlett, and D. M. Clarke, "Permanent Activation of the Human P-glycoprotein by Covalent Modification of a Residue in the Drug-binding Site," *J. Biol. Chem.*, vol. 278, no. 23, pp. 20449–20452, Jun. 2003.

[102] D. M. C. T W Loo, "Identification of Residues in the Drug-binding Site of Human P-glycoprotein Using a Thiol-reactive Substrate," *J. Biol. Chem.*, vol. 272, no. 51, pp. 31945–8, 1997.

[103] T. W. Loo, M. C. Bartlett, and D. M. Clarke, "Methanethiosulfonate derivatives of rhodamine and verapamil activate human P-glycoprotein at different sites," *J. Biol. Chem.*, vol. 278, no. 50, pp. 50136–50141, Dec. 2003.

[104] T. W. Loo, M. C. Bartlett, and D. M. Clarke, "Transmembrane segment 7 of human P-glycoprotein forms part of the drug-binding pocket," *Biochem. J.*, vol. 399, no. 2, pp. 351–359, Oct. 2006.

[105] P. Höcherl, "New tariquidar-like ABCB1 modulators in cancer chemotherapy: Preclinical pharmacokinetic, pharmacodynamic investigations and computational studies," 19-Jul-2010. [Online]. Available: http://epub.uni-regensburg.de/15219/. [Accessed: 17-Nov-2015].

[106] "Protein contacts and ligand binding in the inward-facing model of human P-glycoprotein. - PubMed - NCBI." [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23564544. [Accessed: 21-Jan-2016].

[107] "Application of In Silico Methods to Study ABC Transporters Involved in Multidrug Resistance." [Online]. Available: http://ebooks.benthamscience.com/book/9781608051427/chapter/53291/. [Accessed: 21-Jan-2016].

[108] E. E. Chufan, H.-M. Sim, and S. V. Ambudkar, "Molecular basis of the polyspecificity of P-glycoprotein (ABCB1): recent biochemical and structural studies," *Adv. Cancer Res.*, vol. 125, pp. 71–96, 2015.

[109] T. W. Loo and D. M. Clarke, "Drug-stimulated ATPase activity of human P-glycoprotein requires movement between transmembrane segments 6 and 12," *J. Biol. Chem.*, vol. 272, no. 34, pp. 20986–20989, Aug. 1997.

[110] W. Sherman, T. Day, M. P. Jacobson, R. A. Friesner, and R. Farid, "Novel procedure for modeling ligand/receptor induced fit effects," *J. Med. Chem.*, vol. 49, no. 2, pp. 534–553, Jan. 2006.

Ich habe mich bemüht, sämtliche Inhaber der Bildrechte ausfindig zu machen und ihre Zustimmung zur Verwendung der Bilder in dieser Arbeit eingeholt. Sollte dennoch eine Urheberrechtsverletzung bekannt werden, ersuche ich um Meldung bei mir.

7. Appendix

Abstract

The ATP-driven drug efflux transporters P-glycoprotein (Pgp, ABCB1) and breast cancer resistance protein (BCRP, ABCG2) play an important role in multidrug resistance (MDR) of a wide range of malignancies. The MDR modulator tariquidar is amongst the most potent inhibitors of these ABC transporters. However, several experiments and measurements revealed that some of its analogs appear to be even more potent inhibitors for BCRP than of Pgp.

The aim of this work is to describe the structure-activity relationship of tariquidar analogs and to elucidate the molecular basis of differences encountered in the derivative's affinities towards Pgp and BCRP by using both ligand- and structure-based *in silico* models.

In ligand-based studies, 2D-QSAR studies were carried out to build models. Protein homology modeling was the method of choice for structure-based studies as ABC transporters are embedded in the membrane and thus difficult to crystallize. The newly available refined structure of mouse Pgp (PDB ID: 4M1M) was used in this work.

Two potential binding hypotheses that are able to distinguish the different orientation between the active and inactive ligands of Pgp were revealed. Moreover, two putative binding modes of tariquidar were hypothesised. Another finding is that the analysed compounds interacted differently with the new mouse Pgp compared to the old structure.

Zusammenfassung

Die ATP-getriebenen Efflux-Transporter P-Glykoprotein (Pgp, ABCB1) und das Brustkrebs-Resistenz-Protein, (eng. Breast Cancer Resistance Protein, BCRP, ABCG2) spielen eine wichtige Rolle in der Multidrug-Resistenz (MDR) von verschiedenen Arten von Krebszellen. Der MDR-Modulator Tariquidar gehört zu den hochwirksamen Inhibitoren dieser beiden ABC-Transporter. Jedoch zeigten Experimente, dass einige der modifizierten Tariquidar Analoga noch potentere Inhibitoren für BCRP als für Pgp zu sein scheinen.

Das Ziel der vorliegenden Diplomarbeit ist die Ermittlung der Struktur-Aktivität-Beziehungen von Tariquidar Analoga und deren Präferenz für Pgp oder BCRP mit Hilfe von *in silico* Modellen. Sowohl liganden-basierte als auch struktur-basierte computergestützte Ansätze wurden angewendet.

In ligand-basierten Studien wurden 2D-QSAR Studien durchgeführt, um die Modelle zu bauen. Proteinhomologiemodellierung war die Methode der Wahl für die struktur-basierten Studien, da ABC-Transporter in die Membran eingebettet sind und deshalb die Kristallisation eines solchen Proteins erschwert ist. Die neu verfügbare, verbesserte Struktur des Maus-Pgp wurde in dieser Arbeit verwendet.

Zwei potenzielle Bindungshypothesen, die in der Lage sind, die Orientierungen der aktiven und inaktiven Verbindungen zu unterscheiden, wurden gezeigt. Außerdem wurden zwei potenzielle Bindungshypothesen von Tariquidar postuliert. Die in dieser Arbeit gezeigten Ergebnisse sind unter anderem deswegen relevant, da sie den Ergebnissen anderer Studien, die allerdings die alte Mouse-Pgp-Struktur verwenden, widersprechen. Das deutet darauf hin, dass die Ergebnisse auf Unterschiede in den Strukturen zurückzuführen sind.

List of Abbreviations

ABC ATP-binding cassette

BCRP Breast cancer resistance protein

Cpd Compound

DOPE Discrete Optimized Protein Energy

FP Fingerprints

jCMapper JCompoundMapper

IC_{50} half maximal inhibitory concentration

MD Molecular dynamics

MDR Multidrug resistance

MID Maximum Intracluster Distance

MMFF94x Merck molecular force field 94x

MOE Molecular operating environment

Molpdf Molecular probability density function

MRP1 Multidrug-resistance associated proteins 1

NBD Nucleotide-binding domain

NMR Nuclear magnetic resonance

OPLS Optimized potentials for liquid simulations

pIC_{50} the negative log of the IC50 value in molar

PDB Protein Data Bank

Pgp P-glycoprotein

PLS	Partial least squares
PSI-Blast	Position-specific iterated BLAST
q^2	Predicted variance
r^2	Coefficient of determination
RMSD	Root-mean-square deviation
TMD	Transmembrane domain
TMH	Transmembrane helix
WEKA	Wakaito environment for knowledge analysis

Fingerprints descriptors for building Pgp and BCRP models

Model	WEKA		MOE		Descriptor.	Type of descriptor
	r^2	q^2	r^2	q^2		
Pgp	0.54	0.29	0.54	0.41	MACCSFP32	MACCS fingerprints
					MACCSFP53	
					MACCSFP99	
	0.91	**0.35**	0.91	**0.69**	HASH-7	ECFP fingerprints
					HASH-32	
					HASH-33	
					HASH-54	
					HASH-116	
					HASH-138	
					HASH-355	
					HASH-361	
					HASH-523	
					HASH-789	
					HASH-825	
					HASH-847	
					HASH-930	
BCRP	0.81	-0.01	0.43	0.07	HASH-32	ECFP fingerprints
					HASH-99	
					HASH-123	
					HASH-144	
					HASH-147	
					HASH-208	
					HASH-255	
					HASH-332	
					HASH-388	
					HASH-455	
					HASH-466	
					HASH-473	
					HASH-658	
					HASH-682	
					HASH-687	
					HASH-706	
					HASH-828	
					HASH-969	
					HASH-995	

SiteFinder output

LEU35 MET39 PHE42 PHE170 PHE174 LEU189 SER192 LEU195 GLY196 ALA199 ILE269 ALA272
PHE273 ILE276 TYR277 TYR280 PHE306 LEU309 ILE310 ALA312 PHE313 VAL315 GLY316 GLN317
GLN641 PHE644 MET865 TYR869 PHE894 PHE899 MET902 ALA903 GLN906

Scripts

Common scaffold clustering

```
#svl

#set main 'substructures'
#set author 'lars & freya'



function substructures mdb

// 21.01.2009

/////////////// EINGABE ////////////////////////
local mol_field = 'ligand_minimized';
// Ph scaffold
local Scaffold = 'n1cccc([#6])c1-a1aaa(aa1)[#6]Nc1ccc(ac1)C';
local subs = [Scaffold];
local subs_fields = ['Scaffold'];

/////////////////////////////////////////////

local db = dbv_DefaultView [];
local entries = db_Entries db;
local entry;


local sub;
db_EnsureField [db, 'Scaffold','molecule'];

for entry in entries loop

    Close[force:1];
    local lig = cat db_ReadFields [db,entry,mol_field];
    mol_Create lig;


    // Substructures
```

```
        local i=1;
        for sub in subs loop

                local aKeys = uniq cat sm_MatchAtoms[sub,Atoms[]];


                local data = [mol_Extract aKeys];
                db_Write [db,entry,tag [subs_fields(i),data] ];

        i = i+1;
        endloop

endloop

endfunction
```

R_clustering_centroids.R

```
# This script applies hierarchical clustering on the previously obtained
RMSD-matrix.
# It cuts the output dendrogram at a defined "niveau" (parameter of the
script) and delivers the cluster centroids.
# All files located in the defined _path_ will be considered.
# The output consists of .csv-files with information about
#  - amount of cluster
#  - clustersize
#  - centroids
# additional informations are located in the defined _directory of the
logfiles_
#
# Use as input: a square matrix with comma separated values
#
# Starting the script within R:
#     source ("/home/phi_employee/Scripts/R_clustering_centroids.R")

# Problem: the 'height' component of the output 'tree' can have ties (2
clusters could potentially be merged at the same distance). To cut the tree
we need a well sorted tree (no ties), that is why the trick of round(,6) was
applied to increase the precision in the distances:
# see: http://tolstoy.newcastle.edu.au/R/e4/help/08/05/12735.html
# or: http://www.mail-archive.com/r-help@r-project.org/msg21360.html

# if the package is not installed: type install.packages() and select...



library(clv)

### Definition of the niveau, where the cluster-tree should be cut
### corresponds to the maximal distance within a cluster in Angstrom,
### if the clustering algorithm is set to 'complete-linkage'
### TO CUSTOMIZE
niveau=3

### Definition of the paths ###
### TO CUSTOMIZE
path <- "/home/daria/data/R/"

### INITIAL CLUSTERING AFTER DOCKING (NIVEAU=1) ###

#matrixfiles <- "2_matrix/"
#clusterfiles <- "3_cluster/"
#centroidfiles <- "4_centroids/"
#logfiles <- "clusterlogs/"

### CLUSTERING AFTER MINIMISATION ###

matrixfiles <- "matrix/"
#clusterfiles <- "cluster/"
#centroidfiles <- "centroids/"
#logfiles <- "clusterlogs/"

### CLUSTERING WITHOUT SUBDIRECTORIES ###

#matrixfiles <- "/"
```

```
clusterfiles <- "matrix/"
centroidfiles <- "matrix/"
logfiles <- "matrix/"


### Assignment of the paths

pathmatrixfiles <- gsub(" ", "", paste(path,matrixfiles))
pathclusterfiles <- gsub(" ","",paste(path,clusterfiles))
pathcentroidfiles <- gsub(" ","",paste(path,centroidfiles))
pathlogfiles <- gsub(" ","",paste(path,logfiles))

### Load of data

files <- list.files(pathmatrixfiles, pattern="_matrix.csv")
print(files)

ende = length(files)

anzahl = 1

for ( anzahl in 1 : ende )
{

### Read the RMSD matrix, compute and save the clusters
### NB: the output of the hclust includes something called "height" which is
not an integer corresponding to the level in the tree BUT a distance
### corresponding to the latest merge between 2 clusters at that given level.
Check http://stat.ethz.ch/R-manual/R-patched/library/stats/html/hclust.html
### for more information!!
      matrixfile=gsub (" ", "", paste (pathmatrixfiles, files[anzahl]))
      inputmatrix=read.csv(matrixfile, header=T, sep=",", row.names=1)
      matrix=as.dist(inputmatrix)  ## properly format our input RMSD matrix
      cluster=hclust(matrix, method="complete")   ## matrix is the squared
RMSD distance matrix between all poses
      cluster$height<-round(cluster$height, 6)  ## precision (rounding at 6
numbers after the comma for the distances) DO NOT CHANGE, IT IS NOT A PARAMETER
YOU SHOULD TWEAK
      outputcluster = cutree(cluster, h=niveau)  ## Cuts the tree into several
groups by specifying the cut height
      clusterfile = gsub ("_matrix.csv","_cluster.csv", files[anzahl])
      clusterpath = gsub(" ","",paste (pathclusterfiles, clusterfile))
      write.table(outputcluster,    clusterpath,    quote=FALSE,    sep=";",
col.names="index_matrix;Cluster")  ## saves a table ID;cluster assignment in
the file ending with "_cluster.csv"

### Computation of the centroid matrix

      temp_centroids=cls.attrib(inputmatrix, outputcluster)  ## Return a list
of Means, centers of each cluster, numbers of objects in each cluster
      centroidmatrix=t(temp_centroids$cluster.center)       ##   extract   the
centers
      centroidmatrixfile   =   gsub   ("_cluster.csv","_centroidmatrix.csv",
clusterfile)
      centroidmatrixpath           =           gsub("           ","",paste
(pathlogfiles,centroidmatrixfile))
      write.table(centroidmatrix, centroidmatrixpath, quote=FALSE, sep=";",
col.names=TRUE, row.names=FALSE)  ## writes a table n_poses * n_clusters with
distances to center

### Fusion of the cluster assignment table and the distance to center table
```

```
    clust=read.table(clusterpath, sep=";", header=T)
      temp=read.csv(clusterpath)
      write.table(temp,clusterpath,          quote=FALSE,          sep=";",
col.names="index_matrix;ID;cluster")   ## adds a column with an ID to the
cluster assignment file
    cent=read.table(centroidmatrixpath, sep=";", header=T)
    clust_cent=cbind(clust,cent)     ## joins the tables from the cluster
assignment and the distances to cluster centers
    clustercentroidpath                      =                      gsub
("_centroidmatrix.csv","_clusters_and_centers.csv", centroidmatrixpath)
    write.table(clust_cent,  clustercentroidpath,  quote=FALSE,  sep=";",
col.names=TRUE, row.names=FALSE)

### Computation of the cluster informations

    clusterinfo = cls.scatt.diss.mx(inputmatrix, outputcluster)  ## compute
6 most popular intercluster distances and intracluster distances
    clusterstats                                                  =
cbind(t(clusterinfo$intracls.complete),t(clusterinfo$intracls.average),clus
terinfo$cluster.size)  ## extracts the intracluster "complete" and "average"
distances
    clusterdistancepath                       =                      gsub
("_centroidmatrix.csv","_cluster_dists.csv", centroidmatrixpath)
    write.table("complete;average;size", clusterdistancepath, quote=FALSE,
col.names=FALSE, row.names=FALSE)
    write.table(clusterstats,       clusterdistancepath,       append=TRUE,
quote=FALSE, sep=";", col.names=FALSE, row.names=FALSE)  ## writes a table
n_clusters*3 with complete intracluster distance,
## average intracluster distance and size of cluster into file finishing in
'_cluster_dists.csv'

### Cleanup variables

    rm(tmp)
    rm(fin)
    rm(out)

### Extrahieren und Ausgabe der Centroide

    i=1
    for (i in 1 : max(clust_cent[,2]))
    {
         cent=cbind(clust_cent[,1], clust_cent[,2], clust_cent[,i+2])
         tmp=cent[order(cent[,3]),]

         if (i==1) { fin=cbind(tmp[1,]) }
         else { fin=cbind(fin,tmp[1,]) }
         i=i+1
    }

    centroidpath  =  gsub  ("_centroidmatrix.csv","_centroids_tmp.csv",
centroidmatrixpath)
    write.table("index_matrix;Cluster;dist_to_centroid",    centroidpath,
quote=FALSE, row.names=FALSE, col.names=FALSE)
    write.table(t(fin),centroidpath, append=TRUE, quote=FALSE, sep=";",
col.names=FALSE, row.names=FALSE)
    cent_export=read.table(centroidpath, sep=";", header=T)
    cent_dists=read.table(clusterdistancepath, sep=";", header=T)

    finalcentroidfile   =   gsub   ("_matrix.csv","_centroids.csv",
files[anzahl])
```

```
        finalcentroidpath                =                gsub("                ","",paste
(pathcentroidfiles,finalcentroidfile))
        finalcentroids=cbind(cent_export,cent_dists)
        write.table(finalcentroids, finalcentroidpath, quote=FALSE, sep=";",
col.names=TRUE, row.names=FALSE)

        clust_stat_head=cbind("Datei","Niveau","nCluster","nPosen")
#       clust_stat_head="Datei;Niveau;nCluster;nPosen"
        clust_stat_data=cbind(files[anzahl],niveau,length(clusterinfo$cluster
.size),length(outputcluster))
        cluststatfullpath                                =                                gsub
("_centroidmatrix.csv","_clust_stats.csv",centroidmatrixpath)
        write.table(clust_stat_head, cluststatfullpath, quote=FALSE, sep=";",
col.names=FALSE, row.names=FALSE)
        write.table(clust_stat_data,        cluststatfullpath,        append=TRUE,
quote=FALSE, sep=";", col.names=FALSE, row.names=FALSE)

print(t(c(anzahl,"/",ende, files[anzahl])), quote=FALSE)

        anzahl=anzahl+1
}
```

rmsd-matrix

```svl
#svl

#set main 'rmsd_distanz_matrix'

function mol_RMSD;

function rmsd_distanz_matrix []

load '/home/daria/data/scripts/mol_rmsd.svl';

//////////////// EINGABE ////////////////////////
local db_field = 'carbonized';
local label_field = 'index';
////////////////////////////////////////////////

local db = db_KeyList[];
local entries = db_Entries db;


// indexierung

local ent_Aussen =1;
for length entries loop
    local mol_ref = cat db_ReadFields [db, entries(ent_Aussen),db_field];
    local    field_name    =    totok    db_ReadFields    [db,
entries(ent_Aussen),label_field];
    db_EnsureField [db, field_name,'float'];

    local ent_Innen=1;
    for length entries loop

        local    mol_target    =    cat    db_ReadFields
[db,entries(ent_Innen),db_field];

        local rms = mol_RMSD [mol_ref, mol_target];
        db_Write [db,entries(ent_Innen),tag [field_name,rms]];
    ent_Innen = ent_Innen+1;
    endloop

ent_Aussen=ent_Aussen+1;
endloop

endfunction


function rmsd_distanz_matrix_mol_in_MOE []

// RMSD-Matrix zu Molekuel in MOE


//////////////// EINGABE ////////////////////////
local db_field = 'carbonized';
local label_field = 'index';
////////////////////////////////////////////////

local db = db_KeyList[];
local entries = db_Entries db;
```

```
// indexierung

//local ent_Aussen =1;
//for length entries loop

db_EnsureField [db, label_field, 'float'];

local mol_ref = mol_Extract Atoms[];
local field_name = label_field;

    local ent_Innen=1;
    for length entries loop

        local      mol_target      =      cat      db_ReadFields
[db,entries(ent_Innen),db_field];

        local rms = mol_RMSD [mol_ref, mol_target];
        db_Write [db,entries(ent_Innen),tag [field_name,rms]];
    ent_Innen = ent_Innen+1;
    endloop

//ent_Aussen=ent_Aussen+1;
//endloop

endfunction
```

Pgp-Database
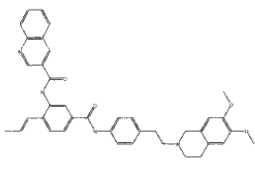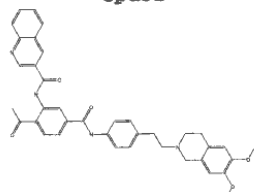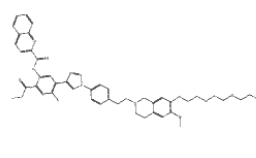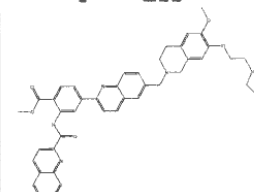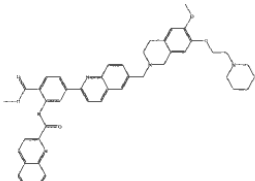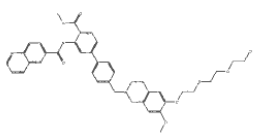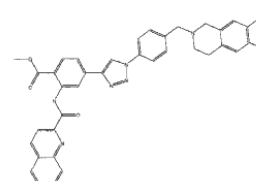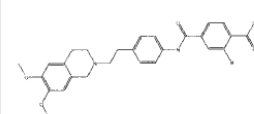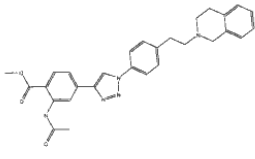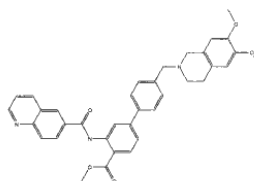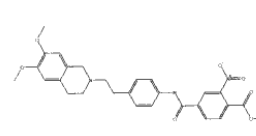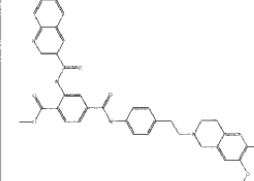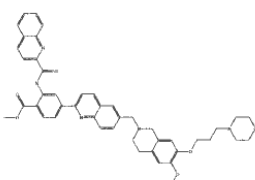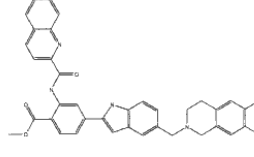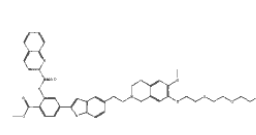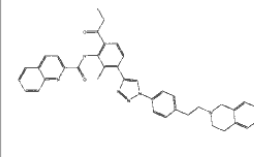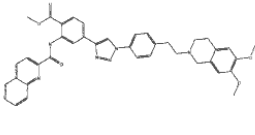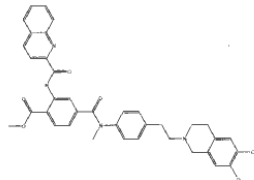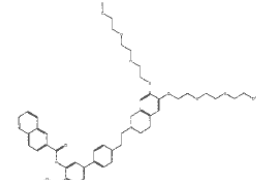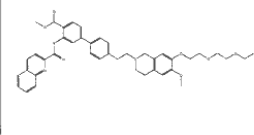
Cpd53	Cpd52	Cpd135_UR-Sts-19-3	Cpd130_QQS-7
pIC50_ABCB1 : 3.98 pIC50_ABCG2 : 5.8327	**pIC50_ABCB1** : 4.02 pIC50_ABCG2 : 6.2062	**pIC50_ABCB1** : 4.06 pIC50_ABCG2 : 7.2924	**pIC50_ABCB1** : 4.43 pIC50_ABCG2 : 5.8194
Cpd129_QQS-6	Cpd86	Cpd117_MB86-2	Cpd2
pIC50_ABCB1 : 4.73 pIC50_ABCG2 : 5.9329	**pIC50_ABCB1** : 4.74 pIC50_ABCG2 : 6.6253	**pIC50_ABCB1** : 4.79 pIC50_ABCG2 : 6.4310	**pIC50_ABCB1** : 4.90 pIC50_ABCG2 :
Cpd140_UR-St-5	Cpd84	Cpd1	Cpd42
pIC50_ABCB1 : 4.92 pIC50_ABCG2 : 5.6472	**pIC50_ABCB1** : 4.96 pIC50_ABCG2 : 5.8125	**pIC50_ABCB1** : 5.01 pIC50_ABCG2 :	**pIC50_ABCB1** : 5.02 pIC50_ABCG2 : 6.9245
Cpd132_QQS-10	Cpd108	Cpd109	Cpd134_UR-Sts-13.3-3
pIC50_ABCB1 : 5.06 pIC50_ABCG2 : 6.0438	**pIC50_ABCB1** : 5.10 pIC50_ABCG2 : 6.4342	**pIC50_ABCB1** : 5.13 pIC50_ABCG2 : 7.2291	**pIC50_ABCB1** : 5.18 pIC50_ABCG2 : 6.5003
Cpd113_MB19-5	Cpd50	Cpd92_UR-COP258	Cpd91
pIC50_ABCB1 : 5.19 pIC50_ABCG2 : 6.3965	**pIC50_ABCB1** : 5.22 pIC50_ABCG2 : 5.7212	**pIC50_ABCB1** : 5.22 pIC50_ABCG2 : 6.2644	**pIC50_ABCB1** : 5.26 pIC50_ABCG2 : 6.1192

Figure 35: Pgp database.

98
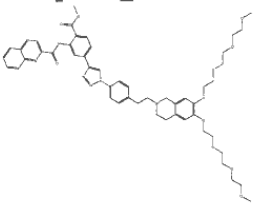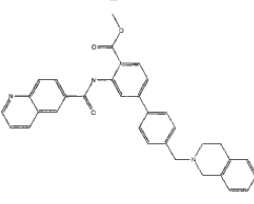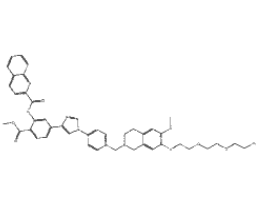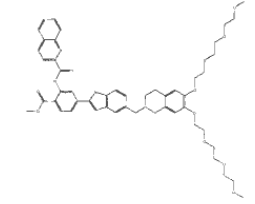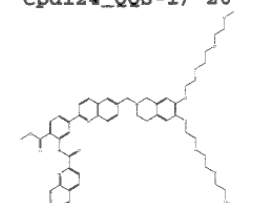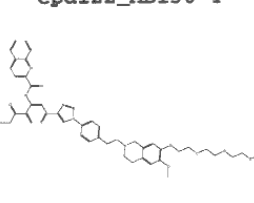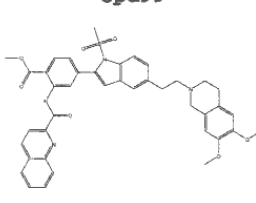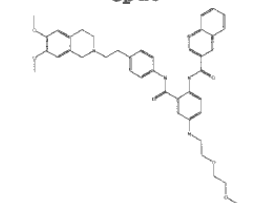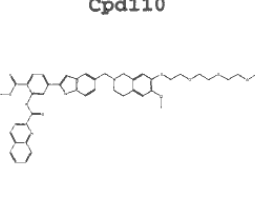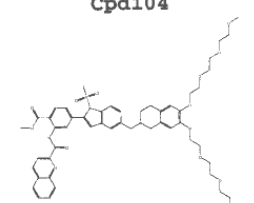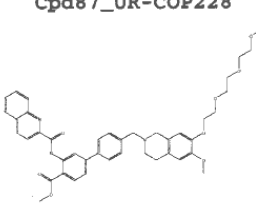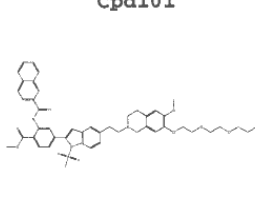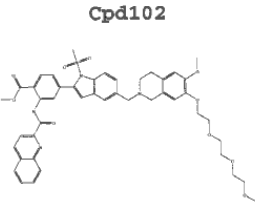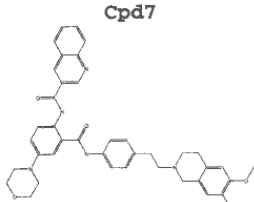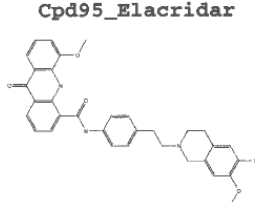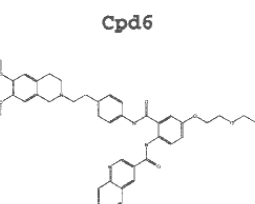
Cpd115_MB83-2	Cpd82	Cpd116_MB84-3	Cpd111
pIC50_ABCB1 : 5.31 pIC50_ABCG2 : 6.7747	pIC50_ABCB1 : 5.34 pIC50_ABCG2 : 6.1931	pIC50_ABCB1 : 5.42 pIC50_ABCG2 : 6.8633	pIC50_ABCB1 : 5.44 pIC50_ABCG2 : 7.3372
Cpd124_QQS-1/ 26	Cpd122_MB136-4	Cpd99	Cpd8
pIC50_ABCB1 : 5.56 pIC50_ABCG2 : 6.2708	pIC50_ABCB1 : 5.59 pIC50_ABCG2 : 7.2596	pIC50_ABCB1 : 5.68 pIC50_ABCG2 : 5.9263	pIC50_ABCB1 : 5.80 pIC50_ABCG2 :
Cpd110	Cpd104	Cpd87_UR-COP228	Cpd101
pIC50_ABCB1 : 5.87 pIC50_ABCG2 : 6.8539	pIC50_ABCB1 : 5.87 pIC50_ABCG2 : 6.0362	pIC50_ABCB1 : 5.91 pIC50_ABCG2 : 6.2284	pIC50_ABCB1 : 5.96 pIC50_ABCG2 : 6.1656
Cpd102	Cpd7	Cpd94_Tariquidar	Cpd95_Elacridar
pIC50_ABCB1 : 5.97 pIC50_ABCG2 : 6.1945	pIC50_ABCB1 : 6.22 pIC50_ABCG2 :	pIC50_ABCB1 : 6.65 pIC50_ABCG2 : 6.2790	pIC50_ABCB1 : 6.71 pIC50_ABCG2 : 6.8962
Cpd6	Cpd5		
pIC50_ABCB1 : 6.74 pIC50_ABCG2 :	pIC50_ABCB1 : 6.83 pIC50_ABCG2 :		

Figure 36: Pgp database (cont'd).

99

BCRP-Database
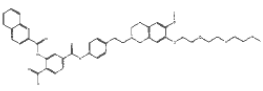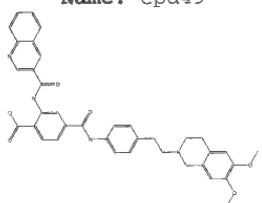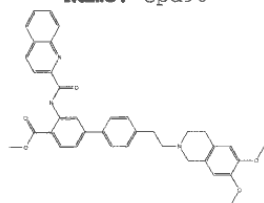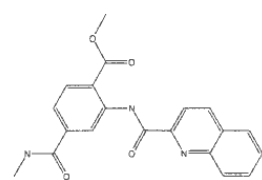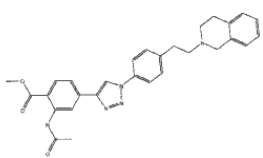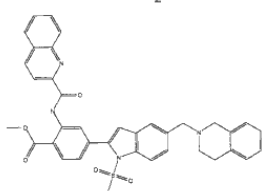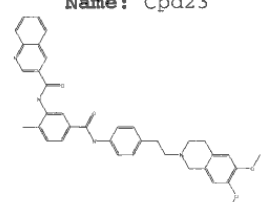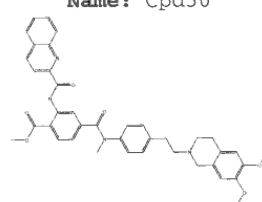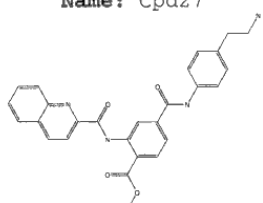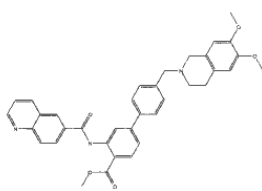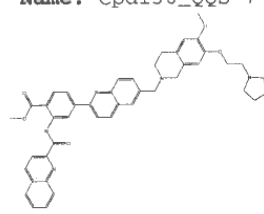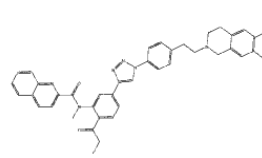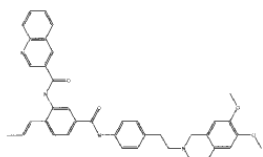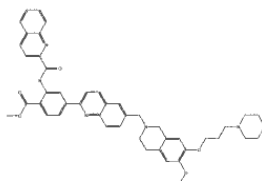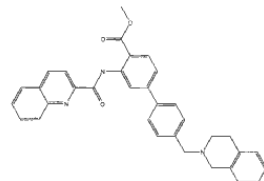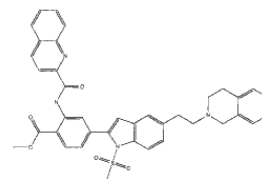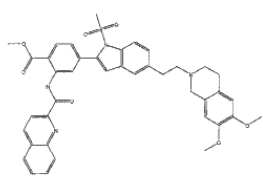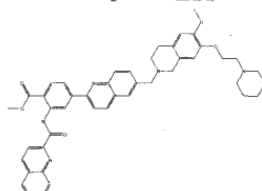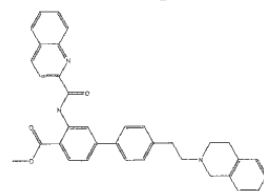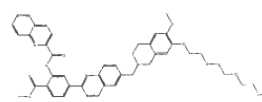
Name: Cpd69 **pIC50_ABCG2** : 4.20 pIC50_ABCB1 :	**Name:** Cpd49 **pIC50_ABCG2** : 5.46 pIC50_ABCB1 :	**Name:** Cpd90 **pIC50_ABCG2** : 5.49 pIC50_ABCB1 :	**Name:** Cpd56 **pIC50_ABCG2** : 5.62 pIC50_ABCB1 :
Name: Cpd140_UR-St-5 **pIC50_ABCG2** : 5.64 pIC50_ABCB1 : 4.9226	**Name:** Cpd98 **pIC50_ABCG2** : 5.65 pIC50_ABCB1 :	**Name:** Cpd23 **pIC50_ABCG2** : 5.70 pIC50_ABCB1 :	**Name:** Cpd50 **pIC50_ABCG2** : 5.72 pIC50_ABCB1 : 5.2218
Name: Cpd27 **pIC50_ABCG2** : 5.77 pIC50_ABCB1 :	**Name:** Cpd84 **pIC50_ABCG2** : 5.81 pIC50_ABCB1 : 4.9626	**Name:** Cpd130_QQS-7 **pIC50_ABCG2** : 5.81 pIC50_ABCB1 : 4.4372	**Name:** Cpd136_UR-PB40 **pIC50_ABCG2** : 5.82 pIC50_ABCB1 :
Name: Cpd53 **pIC50_ABCG2** : 5.83 pIC50_ABCB1 : 3.9854	**Name:** Cpd128_QQS-5 **pIC50_ABCG2** : 5.83 pIC50_ABCB1 :	**Name:** Cpd81 **pIC50_ABCG2** : 5.83 pIC50_ABCB1 :	**Name:** Cpd97 **pIC50_ABCG2** : 5.89 pIC50_ABCB1 :
Name: Cpd99 **pIC50_ABCG2** : 5.92 pIC50_ABCB1 : 5.6887	**Name:** Cpd129_QQS-6 **pIC50_ABCG2** : 5.93 pIC50_ABCB1 : 4.7385	**Name:** Cpd89 **pIC50_ABCG2** : 5.95 pIC50_ABCB1 :	**Name:** Cpd125_QQS-2/ **pIC50_ABCG2** : 5.98 pIC50_ABCB1 :

Figure 37: BCRP database.

100
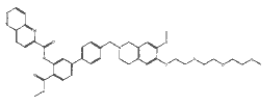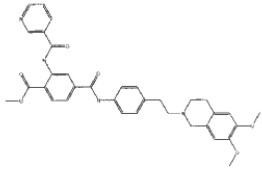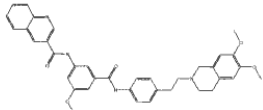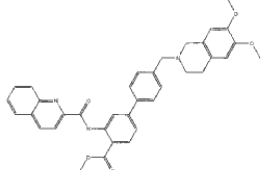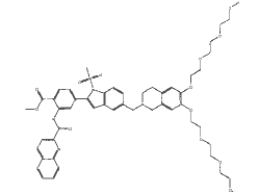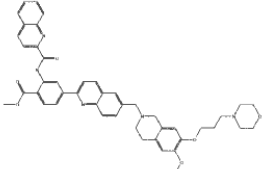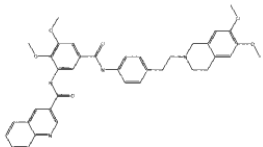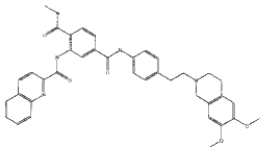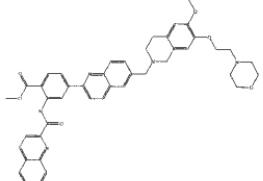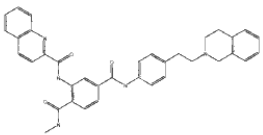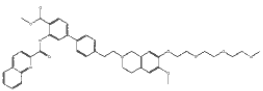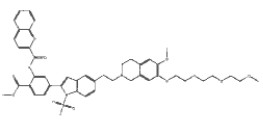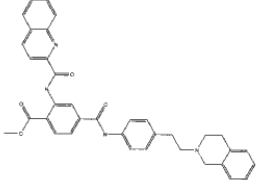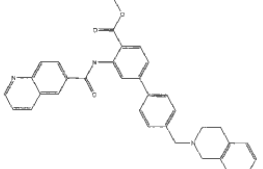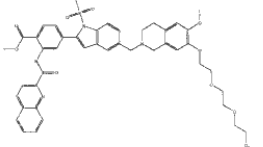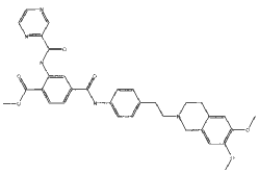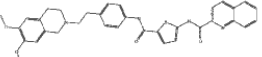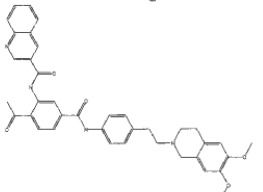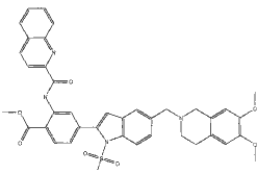
Name: Cpd85	Name: Cpd19	Name: Cpd22	Name: Cpd83
pIC50_ABCG2 : 5.98 pIC50_ABCB1 :	**pIC50_ABCG2** : 5.99 pIC50_ABCB1 :	**pIC50_ABCG2** : 6.01 pIC50_ABCB1 :	**pIC50_ABCG2** : 6.02 pIC50_ABCB1 :
Name: Cpd104	Name: Cpd132_QQS-10	Name: Cpd21	Name: Cpd51
pIC50_ABCG2 : 6.03 pIC50_ABCB1 : 5.8784	**pIC50_ABCG2** : 6.04 pIC50_ABCB1 : 5.0670	**pIC50_ABCG2** : 6.06 pIC50_ABCB1 :	**pIC50_ABCG2** : 6.06 pIC50_ABCB1 :
Name: Cpd127_QQS-4	Name: Cpd88	Name: Cpd46	Name: Cpd91
pIC50_ABCG2 : 6.07 pIC50_ABCB1 :	**pIC50_ABCG2** : 6.07 pIC50_ABCB1 :	**pIC50_ABCG2** : 6.09 pIC50_ABCB1 :	**pIC50_ABCG2** : 6.11 pIC50_ABCB1 : 5.2660
Name: Cpd101	Name: Cpd47	Name: Cpd82	Name: Cpd102
pIC50_ABCG2 : 6.16 pIC50_ABCB1 : 5.9682	**pIC50_ABCG2** : 6.17 pIC50_ABCB1 :	**pIC50_ABCG2** : 6.19 pIC50_ABCB1 : 5.3478	**pIC50_ABCG2** : 6.19 pIC50_ABCB1 : 5.9714
Name: Cpd18	Name: Cpd62	Name: Cpd52	Name: Cpd100
pIC50_ABCG2 : 6.19 pIC50_ABCB1	**pIC50_ABCG2** : 6.20 pIC50_ABCB1	**pIC50_ABCG2** : 6.20 pIC50_ABCB1 : 4.0209	**pIC50_ABCG2** : 6.21 pIC50_ABCB1

Figure 38: BCRP database (cont'd).

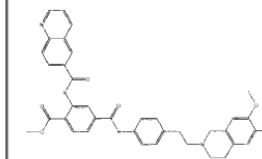101

Name: Cpd126_QQS-3/ **pIC50_ABCG2** : 6.22 pIC50_ABCB1 :	**Name:** Cpd87_UR-COP22 **pIC50_ABCG2** : 6.22 pIC50_ABCB1 : 5.9101	**Name:** Cpd131_QQS-9 **pIC50_ABCG2** : 6.23 pIC50_ABCB1 :	**Name:** Cpd17 **pIC50_ABCG2** : 6.25 pIC50_ABCB1 :
Name: Cpd92_UR-COP25 **pIC50_ABCG2** : 6.26 pIC50_ABCB1 : 5.2226	**Name:** Cpd124_QQS-1/ **pIC50_ABCG2** : 6.27 pIC50_ABCB1 : 5.5672	**Name:** Cpd94_Tariquid **pIC50_ABCG2** : 6.27 pIC50_ABCB1 : 6.6517	**Name:** Cpd114_MB81-2 **pIC50_ABCG2** : 6.37 pIC50_ABCB1 :
Name: Cpd113_MB19-5 **pIC50_ABCG2** : 6.39 pIC50_ABCB1 : 5.1972	**Name:** Cpd74 **pIC50_ABCG2** : 6.40 pIC50_ABCB1 :	**Name:** Cpd106 **pIC50_ABCG2** : 6.41 pIC50_ABCB1 :	**Name:** Cpd117_MB86-2 **pIC50_ABCG2** : 6.43 pIC50_ABCB1 : 4.7966
Name: Cpd108 **pIC50_ABCG2** : 6.43 pIC50_ABCB1 : 5.1020	**Name:** Cpd20 **pIC50_ABCG2** : 6.49 pIC50_ABCB1 :	**Name:** Cpd105 **pIC50_ABCG2** : 6.50 pIC50_ABCB1 :	**Name:** Cpd134_UR-Sts- **pIC50_ABCG2** : 6.50 pIC50_ABCB1 : 5.1805
Name: Cpd45 **pIC50_ABCG2** : 6.52 pIC50_ABCB1 :	**Name:** Cpd138_UR-St-2 **pIC50_ABCG2** : 6.57 pIC50_ABCB1 :	**Name:** Cpd139_UR-St-3 **pIC50_ABCG2** : 6.62 pIC50_ABCB1 :	**Name:** Cpd86 **pIC50_ABCG2** : 6.62 pIC50_ABCB1 : 4.7447

Figure 39: BCRP database (cont'd).

102
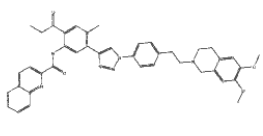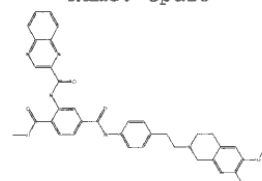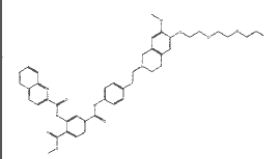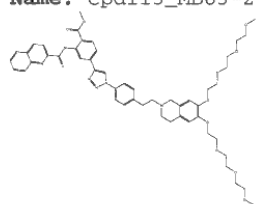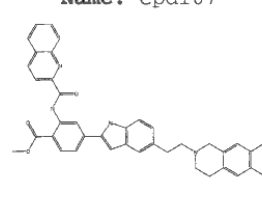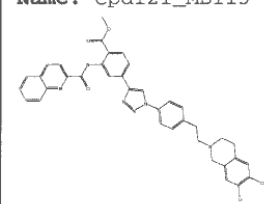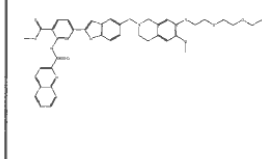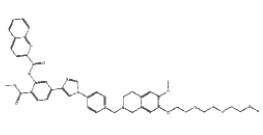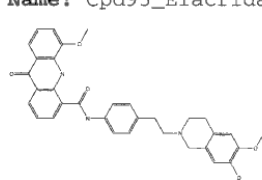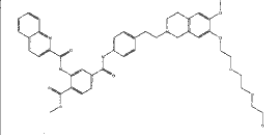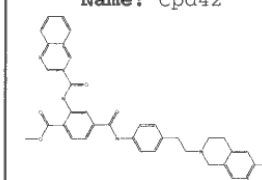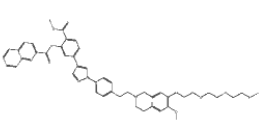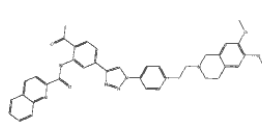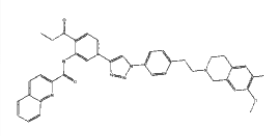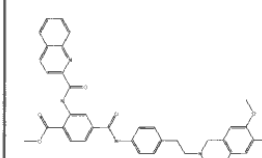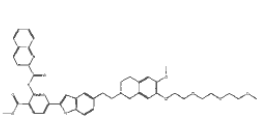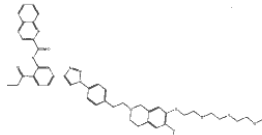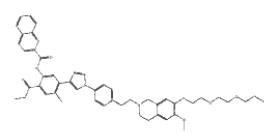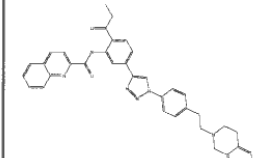
Name: Cpd133_UR-Sts-	Name: Cpd15	Name: Cpd73_UR-COP77	Name: Cpd72_UR-ME19-
pIC50_ABCG2 : 6.69 pIC50_ABCB1 :	**pIC50_ABCG2** : 6.73 pIC50_ABCB1 :	**pIC50_ABCG2** : 6.73 pIC50_ABCB1 :	**pIC50_ABCG2** : 6.76 pIC50_ABCB1 :
Name: Cpd115_MB83-2	Name: Cpd107	Name: Cpd121_MB113-3	Name: Cpd110
pIC50_ABCG2 : 6.77 pIC50_ABCB1 : 5.3192	**pIC50_ABCG2** : 6.79 pIC50_ABCB1 :	**pIC50_ABCG2** : 6.83 pIC50_ABCB1 :	**pIC50_ABCG2** : 6.85 pIC50_ABCB1 : 5.8726
Name: Cpd116_MB84-3	Name: Cpd95_Elacrida	Name: Cpd80	Name: Cpd42
pIC50_ABCG2 : 6.86 pIC50_ABCB1 : 5.4225	**pIC50_ABCG2** : 6.89 pIC50_ABCB1 : 6.7144	**pIC50_ABCG2** : 6.92 pIC50_ABCB1 :	**pIC50_ABCG2** : 6.92 pIC50_ABCB1 : 5.0246
Name: Cpd118_MB95-2	Name: Cpd137_UR-St-1	Name: Cpd120_MB108-4	Name: Cpd71_UR-ME22-
pIC50_ABCG2 : 6.96 pIC50_ABCB1 :	**pIC50_ABCG2** : 7.03 pIC50_ABCB1 :	**pIC50_ABCG2** : 7.19 pIC50_ABCB1 :	**pIC50_ABCG2** : 7.22 pIC50_ABCB1 :
Name: Cpd109	Name: Cpd122_MB136-4	Name: Cpd135_UR-Sts-	Name: Cpd119_MB107-6
pIC50_ABCG2 : 7.22 pIC50_ABCB1 : 5.1367	**pIC50_ABCG2** : 7.25 pIC50_ABCB1 : 5.5901	**pIC50_ABCG2** : 7.29 pIC50_ABCB1 : 4.0633	**pIC50_ABCG2** : 7.31 pIC50_ABCB1 :

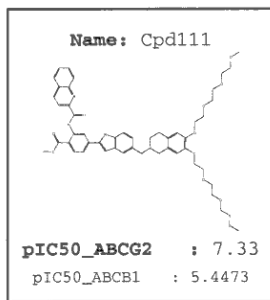Figure 40: BCRP database (cont'd).

103

Name: Cpd111

pIC50_ABCG2 : 7.33
pIC50_ABCB1 : 5.4473

Figure 41: BCRP database (cont'd).