Identification of substrates of P-Glycoprotein using *in-silico* methods

Victor Osho

A thesis submitted in part fulfilment of the requirements for the award of an MPharm Honours Degree in Pharmacy

February 2013

Medway School of Pharmacy

The Universities of Kent and Greenwich

''I certify that this work has not been accepted in substance for any degree, and is not concurrently being submitted for any degree other than that of the MPharm Honours Degree in Pharmacy. I also declare that this work is my own except where otherwise identified by references and that I have not plagiarised another's work''.

Signature..... **Date.....**

Total word count = 8056 words. This excludes tables, figures, appendices and references.

ACKNOWLEDGMENTS

I would like to thank my supervisor, Dr Ghafourian for the opportunity to conduct this research and for her guidance and thorough attention during the research process. I would also like to thank other members of the Drug Delivery Group for their support. I would also like to recognise my parents, friends and family for their advice and words of encouragement.

ABSTRACT

The ABC transporter superfamily is one of the largest and abundant families of proteins. It is a large group of proteins that transport a range of substances in cell systems. The ABC transporter P-glycoprotein (ABCB1, P-gp), a polyspecific protein has demonstrated its function as a transporter of hydrophobic drugs as well as transporting lipids, steroids and metabolic products. As well as this, previous studies have shown that P-gp is over expressed in cancerous tissues and plays a role in multidrug resistance. In this study, *in-silico* methods were used to dock a data set of compounds to P-glycoprotein structures available in the Protein data bank. Binding sites were defined using co-crystallised ligand structures of P-gp and docking energies were calculated using MOE. Statistical models were built to gain a better understanding of how compounds may interact with P-gp. The protein was able to bind to structurally different compounds and results indicate that LogP is the most important factor for drug binding to P-glycoprotein.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 Introduction

1.1 Transporters

In every living cell, the transport of molecules is an important part of its survival. For molecules to enter, leave or cross cell membranes they usually require a protein to aid their movement. It is reported that the human genome has nearly 900 transporter genes (Anderle, et al, 2004) that encode proteins which are mainly responsible for transporting molecules, nutrients and drugs throughout the body. From this group, it is reported that approximately 350 are genes that encode intracellular transporters. These transporters can be separated into three classes relating to their binding or carrier potential; ATP powered pumps, channel transporters and translocators (Lodish, et al., 2000). ATP pumps are transporters that require energy from ATP hydrolysis to be able to transfer various molecules across a membrane against the concentration gradient. ATP is hydrolysed to ADP and P_i, which release energy enabling transport of the molecule across the membrane. Examples of ATP powered pumps include ATP-binding cassette transporter, Na+/K+-ATPase and Hydrogen potassium ATPase. Channel transporters create a passage through the membrane which they wish to transport molecules through. They are regularly involved in the transport of water and ions such as sodium and chloride which require transport across membranes by facilitated transport.

1.1 ABC Transporters

The ABC transport superfamily is one of the largest and abundant families of proteins. It is a large group of proteins that transport a range of substances that include amino acids, drugs and lipids as well as several others (Sharom, 2008). The ABC proteins can be located in several organ membranes in humans, and in prokaryotes they are found in the cytoplasmic membrane of bacteria (Sharom, 2008).

Part of their structure includes an ATP-binding domain that utilises ATP hydrolysis to transport compounds across cell membranes. A typical ABC transporter has four core domains; two membrane-associated domains and two ATP-binding domains (Higgins, 2001). The trans-membrane domains are situated across the membrane and function as the route for molecules to cross the membrane. The ATP-binding domains are located in the cytoplasm of the cell and are consequently hydrophilic in nature (Dean, 2002).

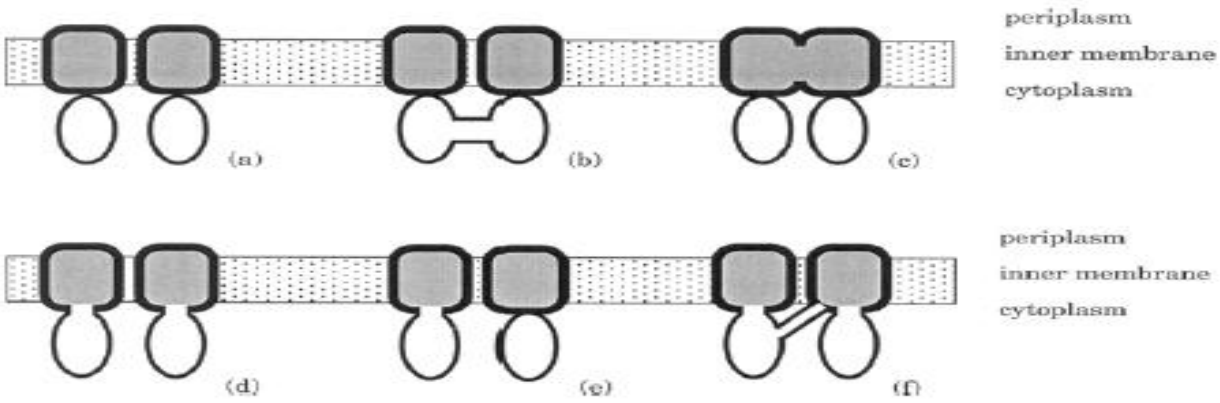


Figure 1. ABC transporter organisation. Domains of ABC transporters can be expressed by different combinations. A) encoded as four individual polypeptides B) combined ATP-binding domain C) combined membrane-associated domains D) membrane-associated domain fused to ATP-binding domain E) membrane-associated domain fused to ATP-binding domain F) All domains fused into one single polypeptide (Higgins, 2001)

ABC transporter proteins bind ATP and use the energy derived from this to transfer molecules across cell membranes. In eukaryotic cells, ABC transporters usually direct molecules from the cytoplasm to the outside of the cell (Dean, 2002) with the main function of transporting xenobiotic compounds out of the cell for transport to other areas of the body or for excretion. On the other hand, ABC transporters in prokaryotic cells can be either an importer or exporter of compounds. Bacterial importers are important for the cell survival and typically important substrates such as iron, inorganic ions as well as peptides and amino acids. Substances requiring removal from prokaryotic cells include cell wall components such as liposaccharides and toxins involved in pathogens e.g. haemolysin (Davidson et al, 2008). ABC proteins also play a role in the translation of mRNA and are involved in repairing DNA.

The ABC trans-membrane protein family is a large group and 48 ABC transporter genes have been recognised in the human genome (Ambudkar, et al., 2003). ABC genes are classified into subfamilies which are further divided into subgroups (Dean, 2002). The best studied groups include ABCB1 also known as MDR1 due to its ability to produce multiple drug resistance in cancer cells. The sulphonylurea receptor (SUR) subfamily is involved in regulating insulin secretion in β-cells of the pancreas (Dassa and Bouige, 2001). Others include the ABCC subfamily which encodes the cystic fibrosis transmembrane conductance regulator (CFTR) protein that plays a part in exocrine secretions of chloride (Dean, 2002; Dassa and Bouige, 2001). Despite playing a functional role in cells, mutations in up to 14 mammalian ABC transporters have been associated with

disease states (Borst and Elferink, 2002). These proteins as well as ABCG2 and ABCB1 are reported to be overexpressed in malignant cells thus causing these cells to be resistant to drug therapy, hence the multidrug resistance terminology.

1.1.2 Role of ABC Transporters in Multidrug Resistance

During cancer treatment, tumour cells can become resistant to chemotherapy due to increased excretion of drugs out of tumour cells or target proteins (Dean, 2002). Pathways such as these can lead to multidrug resistance thus contributing to the failure of chemotherapy in malignant diseases. Multidrug resistance is the term given to describe tumours developing resistance to two or more chemotherapeutic drugs. This is the net result of the over-expression of membrane transporters that actively remove toxic chemotherapeutic agents out of tumour cells (Sarkadi, et al., 2006). ABC transporters have been widely associated with resistance and the ABC genes ABCB1, ABCC1 and ABCG2 can be upregulated in cancerous cells.

For example, ABCG2 was first named as the Breast Cancer Resistant Protein (BCRP) when it was found in doxorubicin resistant cancer cells (Saito, et al., 2010). ABCG2 can be located in normal tissues and endothelial cells where it forms a barrier between blood supply and tissues. ABCG2 is also expressed in placental trophoblast cells, in the epithelium of small intestine and liver membrane as well as ducts and lobes of the breast (Saito, et al., 2010). The fact that there are high levels of expression of this protein in trophoblast cells suggests that BCRP is responsible for transporting compounds into blood supply and for removing toxic metabolites (Satio, et al., 2010).

The ABCC1 gene encodes for the multidrug resistance protein MRP1 (Dassa and Bouige, 2001). MRP1 is expressed in epithelial cells and in non-malignant cells it plays a role in protecting kidney tissues, bone marrow and the intestinal mucosa from xenobiotics as well as contributing to the removal of drugs from the cerebrospinal fluid (Schinkel and Jonker, 2012). Moreover, MRP1 confers drug resistance to a range of cancer drugs and transports conjugates of hydrophobic drugs as well as organic anions (Schinkel and Jonker, 2012).

P-glycoprotein is the transporter encoded by the ABCB1 gene. It was one of the first ABC transporters to be associated with resistance (Leslie, et al., 2005) and led to the discovery of other genes in the ABC transporter family involved in multidrug resistance. P-glycoprotein is one of the most widely studied ABC transporters because it transports a wide range of substrates including anticancer drugs. P-glycoprotein is highly expressed in cancerous tissues and it is reported to be involved in cancers of the liver, colon and kidney tissues (Schinkel and Jonker, 2012). Due to its diverse substrate specificity, scientists have sought to gain a better understanding of how substrates bind to it, in order to develop potential inhibitors that may improve the efficacy of anticancer therapy.

1.2 P-glycoprotein

P-glycoprotein (P-gp), a well studied glycoprotein was first discovered in 1976 by surface labelling studies in drug resistant ovary cells (Juliano and Ling, 1976). Since then, it has demonstrated its function as a transporter of hydrophobic drugs as well as transporting lipids, steroids and metabolic products. Encoded by the ABCB1 gene, it is also known to play a role in transporting compounds across the blood brain barrier and is involved in the uptake of the cardiac glycoside Digoxin in the kidneys (de Lannoy and Silverman, 1992). It is highly expressed in various cells of the body but is mainly presented in epithelial cells. In the blood brain barrier, P-gp protects the brain from toxic products and drugs that cross this threshold. P-gp substrates that are lipophilic can easily diffuse across endothelial cells and enter the brain. However a high proportion of P-gp surrounds this area of the brain preventing their accumulation and the role of P-gp is to distribute substrates back into blood circulation (Schinkel and Jonker, 2012). Similarly in cells of the liver, P-gp is responsible for the excretion of drugs from hepatocytes into the bile thus reducing the bioavailability of drugs exerting their effects in these cells. Inclusive of these two areas, P-gp can also be found to be expressed in the intestines, placenta, kidneys and adrenal glands excreting harmful metabolic products (Dean, 2002).

1.2.1 Structure of P-glycoprotein

As described previously, most ABC transporters consist of two trans-membrane domains and an ATP-binding domain that uses energy from ATP to transport products (Figure 1). P-glycoprotein is known as a full transporter and contains six transmembrane domains with an ATP-binding domain separated by a flexible linker region (Ambudkar, et al., 2003). The structure of human p-glycoprotein was first elucidated by electron microscopy (Rosenberg, et al., 1997) and image analysis. P-gp was reported as having a central core with an opening to the extracellular side of the membrane but is closed towards the cytoplasm.

In 2009, Aller et al reported a medium resolution (3.8-4.4Å) X-ray structure of P-glycoprotein that supported previous claims about the structure of P-gp and revealed tentative binding sites for drug compounds (Aller, et al., 2009). The study proposed the structure for mouse P-gp with 87% sequence identity to human p-glycoprotein (Figure 2). In addition to this, the structure of P-gp co-crystallised with the cyclic peptide inhibitors cyclic-tris-(R)-valineselenazole (QZ59-RRR) and cyclic-tris-(S)-valineselenazole (QZ59-SSS) was also determined, suggesting particular amino acid residues that are involved in drug binding (Figure 2).

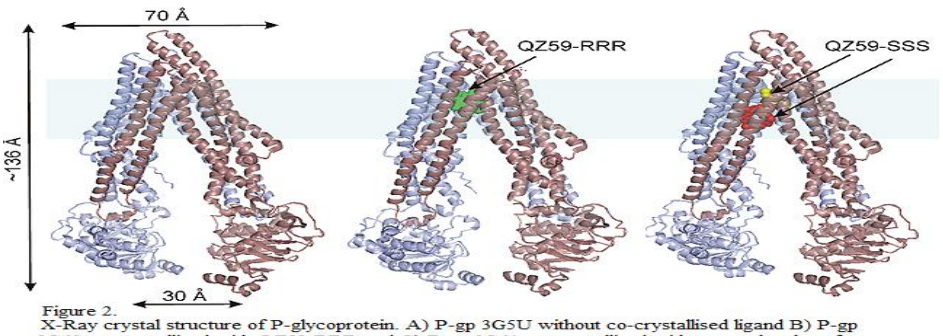


Figure 2.
X-Ray crystal structure of P-glycoprotein. A) P-gp 3G5U without co-crystallised ligand B) P-gp 3G60 co-crystallised with QZ59-RRR and C) P-gp 3G61 co-crystallised with two molecules of QZ59-SSS (Aller et al, 2009)

1.2.2 Binding sites of P-glycoprotein

The X-ray crystal structures proposed by Aller did give some useful information regarding the amino acid residues involved in substrate binding to P-gp. The crystal structure showed one molecule of QZ59-RRR bound to the middle site in the binding pocket (Figure 2B), and two molecules of QZ59-SSS bound at upper and lower sites which are overlapping the middle site (Figure 2C). This showed that P-gp can bind to two drug molecules at the same time and confirmed the diverse and polyspecific nature of P-glycoprotein (Gutmann, et al., 2009).

The binding pocket was said to include the transmembrane helices 1, 6, 7 and 12 which mainly consisted of hydrophobic and aromatic residues. These included Phenylalanine (Phe) and Tyrosine (Tyr) residues in addition to the aromatic and aliphatic residues Serine, Threonine and Glutamine (Ser, Thr, Gln). Despite these key attributes being made available, questions have been raised about the absence of ATP in the structure and the fact that the structures do not appear to undergo conformational changes upon drug binding (Gottesman, et al., 2009).

Substrates of P-gp mainly interact with the protein by hydrophobic interactions, π-π stacking and Van der Waals forces. The P-gp X-ray crystal structure also shows this as the cyclic peptide inhibitors bind to P-gp through hydrophobic aromatic side residues (Aller, et al., 2009). Studies have also demonstrated that P-glycoprotein is a flexible molecule that can alter its confirmation in order for substrate entry. These findings led to a proposed induced-fit mechanism for drug binding to P-gp, in which the substrate enters the large binding pocket and both drug and protein modify their shape to generate more favourable contacts unique to that substrate (Alonso, et al., 2006).

This mechanism is supported by the X-ray structure of P-gp, where each of the ligands bound to P-gp interact with the protein at different or the same overlapping amino acid residues (Figure 3).

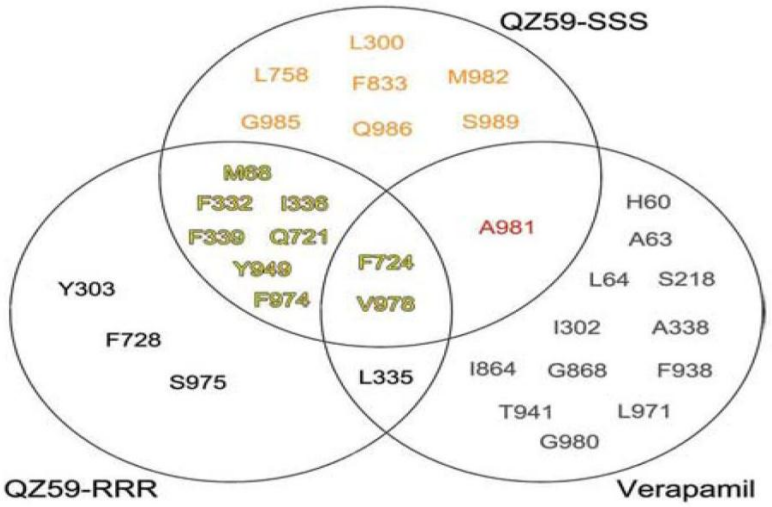


Figure 3. Venn diagram of amino acid residues involved in binding of cyclic peptide inhibitors and those predicted to be involved in Verapamil binding. (Aller, et al., 2009)

1.3 *In-Silico* Methods

Traditionally, drugs are usually discovered in biological assays and in time-consuming *in vivo* and *in vitro* testing. However, the use of computer modelling in drug discovery has rapidly been developed creating techniques and software that are able to analyse and predict information about biological, chemical and medical data. The term *'In-Silico'* refers to the computational approach of drug discovery which is complementary to *in vivo* and *in vitro* experiments (Ekins, et al., 2007). In a widely expanding field, *in-silico* techniques have been used to create virtual models that enable scientists to make predictions about biological activity and provide advances in medicine.

Various approaches can be considered an *in-silico* method, and one of the most well known is quantitative structure-activity relationships (QSAR). Since the 1960s, QSAR has been used to describe the mathematical relationship between the structure of a molecule and biological activity (Van de Waterbeemd and Rose, 2003). It has now been further developed with branches of 2D and 3D QSAR methods based on the type of molecular descriptors involved. In contrast to QSAR is virtual screening, a knowledge based method that requires structural information about the target or the compound being developed (Klebe, 2006). A sample of small molecules highlighted as candidate ligands are ranked in order of affinity for the target and this way of generating lead compounds has become an essential part of the pharmaceutical industry. Other *in-silico* methods involve pharmacophore modelling that uses 3D structure representations to describe how candidate ligands may bind to a target (Ekins, et al., 2007). Target based methods involve docking compounds to a target site and the use of scoring functions to score the binding affinity of the ligand to the target. It has gained popularity in recent times and has been involved in the discovery of inhibitors of HIV-1 integrase (Hayouka, et al., 2010).

1.4 Molecular Docking

Molecular docking is a computational method used to estimate the binding energy of a ligand to a specific receptor (Huang, 2007). It is the process of building a model based on molecular properties of an individual compound or a library of compounds with a target structure usually a protein. In protein ligand docking, the docking program aims to find the preferred conformation of the ligand at a binding site of the target (Sousa., et al, 2006). The binding energy is then calculated for each conformation and is ranked and scored to give an estimation of the binding affinity between a compound and target. Docking has been successful in the discovery of novel ligands and inhibitors of enzymes. Docking has also been involved in producing inhibitors of aldose reductase (Iwata, 2001), carbonic anhydrase as well as HIV-1 integrase mentioned previously.

At present, there is a wide range of docking software available in the market with different scoring functions. The program AUTODOCK is one of the most cited docking programs and uses the Lamarckian genetic algorithm as well as a traditional genetic algorithm (Sousa, et al., 2006). GOLD is another program that is popular in the field and enables flexibility of the protein hydrogen bonds as well as the ligand being tested. Unlike AUTODOCK, docking scores in GOLD are ranked using a force field scoring function that includes the contributions of hydrophobic interactions, Van der Waals forces and number of hydrogen bonds (Cummings, et al., 2005). FlexX is another software package that permits protein flexibility and scores the final position of molecules using the empirical Böhm's

scoring function (Sousa, et al., 2006). In addition to these aforementioned programs, the Molecular Operating Environment (MOE) is a suite of applications that can be used for medicinal chemistry purposes. It includes a docking tool that searches for complimentary binding poses between a ligand and a rigid receptor which can be used to determine interactions between candidate ligands and targets.

1.4.1 Scoring Functions

Scoring functions are used to calculate the binding energy of poses generated after docking placement. A very accurate scoring function is desired to be able to successfully predict binding affinity, however due to the complexity and high computational cost involved, scoring functions make assumptions about molecular interactions based on experimental data from independent reactions (Lipkowitz and Boyd, 2002). In all scoring functions, a lower score indicates a more favourable pose while higher scores suggest that binding is less likely. Scoring functions are based on different calculation methods and can be divided into three categories: knowledge-based, force field and empirical based methods.

Knowledge-based functions use data from statistical analysis of structural complexes in the protein data bank, to estimate interatomic reactions occurring frequently between a ligand and the protein in specified intervals (Schulz-Gasch and Stahl, 2004). Typical examples of knowledge based scoring functions include Muegges's potential of Mean Force (PMF), DrugScore and the SMoG score (Sousa, Fernandes and Ramos, 2006).

GoldScore, Assisted Model Building and Energy Refinement (AMBER) and the Optimised Potentials for Liquid Simulations function (OPLS), are examples of force-field scoring functions. Force-field scores are calculated by measuring electrostatic and Van der Waals interactions (Schulz-Gasch and Stahl, 2004) but is limited by the exclusion of solvation and entropic properties (Sousa, et al., 2006). In contrast to these two scoring functions, empirical scores estimate free binding energy based on a sum of localised independent reactions (Lipkowitz and Boyd, 2002). In most cases, the constants in empirical formulas are derived from binding energies calculated in experiments of receptor-ligand complexes (Sousa, et al., 2006). An example of an empirical scoring function is the London dG scoring utilised in MOE (Figure 4).

$$\Delta G_{LdG} = c + E_{flex} + \sum_{h-bonds} c_{hb} f_{hb} + \sum_{metal-lig} c_m f_m + \sum_{atomsi} \Delta D_i$$

Figure 4. London dG Scoring Function (Corbeil et al, 2012)

The formula above calculates binding energy where E_{flex} represents the energy due to loss of flexibility of the ligand, f_{hb} and C_{hb} are measurements of hydrogen bonds while C_M and f_M measure energies related to metal ligation.

1.5 Data Mining

Data mining is the process of extracting information and establishing relationships from large sets of data. Methods involve computer-based statistics, pattern recognition and database technology (Hand, et al., 2001). Data mining methods can be classified into two categories: predictive and descriptive data mining (Kantardzic, 2011). Descriptive data mining tools aim to produce new information about a dataset and to establish patterns and relationships from a large data set. In contrast, predictive methods are for building models regarding the information available in the database which can then be used to make predictions or classify data (Kantardzic, 2011).

Classification and regression trees (CART) are predictive systems that can be used to classify data into pre-defined classes. Decision trees are built which split data into categories in response to a dependent variable. Classification trees are developed if the dependent variable is categorical (e.g. substrate/non-substrate) and regression trees are formed when continuous data is available (Deconinck, et al., 2005). Interaction trees are similar but the user can manually select a variable as the independent variable and then allow the software to grow the tree automatically. Support Vector Machine (SVM) is a machine learning method that uses a mathematical algorithm to classify data. It constructs a hyperplane in a high dimensional space which separate sets of linear data by the maximum margin (Lipkowitz, Cundari, 2007). In the case of non-linear data, SVM constructs a higher dimensional feature space, using Kernel functions such as Gaussians, polynomial and RBF kernels (Wang, 2005).

1.6 Research objectives

P-glycoprotein is a poly-specific protein and is able to recognise structurally diverse substrates. In this study, we aim to develop a better understanding of how compounds interact with P-gp. It is our aim to build a model using statistical methods to describe the common features of compounds that bind to P-glycoprotein. In addition to this, there are currently three X-ray structures of P-gp available in the protein data bank (3G5U, 3G60, 3G61). In these structures, there are at least two binding sites that have been suggested. It is our aim to optimise the molecular docking strategy to identify substrates of P-gp and suggest the most favourable binding site. By doing this, we can establish the type of interactions that occur between P-gp substrates in the drug binding pocket and provide information about the amino acid residues that are involved.

2 Methods

2.1 Data Set

Compounds used in the data set as part of this study, were selected from a previous research paper, Matsson et al, (2009) which had studied activity of ABC transporters. The compounds in the data set had been chosen as they represented a cross section of available and licensed oral drugs. The compounds in the data set are reported as inhibitors and non-inhibitors of P-gp based on the previous experimental results. Using this information, each compound was assigned a value of '1' if in the literature it was a substrate or inhibitor and assigned a value of '0' if it was a non-inhibitor.

We then proceeded to gather the simplified molecular-input line-entry system code (SMILES) and the registration number (RN) for each compound into Microsoft Excel. This information was collected from online databases ChemSpider and Pubchem. These databases were used as they are available to access for free and provide reliable information about the structures and properties of millions of compounds. While searching for the SMILES code for Ivermectin, we noticed that it had two isomers which had not been included in the data extracted from the literature. As a result, we decided to use both isomers and include this in our data set (Ivermectin A & B). In total, 123 compounds were included in our data set (Table 1), with 54 classified as substrates and 69 non-substrates. From the 123 compounds, 98 compounds were separated randomly into a training set and 25 in the validation set.

Table 1. Substrates and Non-Substrate of P-glycoprotein. 0=non substrate, 1=substrate

Compound	ABCB1P-gp	Compound	ABCB1-Pgp	Compound	ABCB1-Pgp
Chlorprotixene	1	Diltiazem	1	Colchicine	0
Cyclosporine-A	1	Taurolithocholic acid	1	Dehydroisoandrosterone-3-sulfate	0
Diethylstilbestrol	1	Haloperidol	1	Desipramine	0
Dipyridamole	1	Maprotiline	1	Digoxin	0
Flupentixol	1	Noscapine	1	Doxorubicin	0
GF120918	1	Prednisone	1	Erythromycin	0
Isradipine	1	Procyclidine	1	Estradiol-17β-	0
Ivermectin	1	Propafenone	1	Etoposide	0
Loperamide	1	Quinidine	1	Fexofenadine	0
Lopinavir	1	Quinine f	1	Flucloxacillin	0
MK571	1	Taurocholate	1	Hydrochlorothiazide	0
Quercetin	1	Tetracycline	1	Hydrocortisone f	0
Reserpine	1	Vinblastine	1	Indinavir	0
Ritonavir	1	Amodiaquine	0	Indomethacin	0
Saquinavir	1	Fumitremorgin	0	Mesalazine	0
Silymarin	1	Hoechst 33342	0	Methotrexate	0
Tamoxifen	1	Mitoxantrone	0	Metoprolol	0
Terfenadine	1	Naringenin	0	Nevirapine	0
Thioridazine	1	Omeprazole	0	Nicotine	0
Benzbromarone	0	Prazosin	0	Ofloxacin	0
Amiodarone	1	Progesterone	0	Phenobarbital	0
Apigenin	1	Bromosulfalein	0	Phenylethyl isothiocyanate	0
17β-estradiol	1	Lansoprazole	0	Phenytoin	0
Biochanin A	1	P-aminohippuric acid	0	Pravastatin	0
Chlorpromazine	1	Rifampicin	0	Prednisolone	0
Chrysin	1	1-methyl-4-phenylpyridinium	0	Probenecid	0
Ergocristine	1	4-Methylumbelliferoneglucuronide	0	Propranolol	0
Felodipine	1	Amantadine	0	Ranitidine	0
Gefitinib	1	Amiloride	0	Sotalol	0
Genistein	1	Amitriptyline	0	Sparfloxacin	0
Glibenclamide	1	Antipyrine	0	Sulfasalazine	0
Imatinib mesylate	1	Atropine	0	Sulfinpyrazone	0
Ketoconazole	1	Budesonide	0	Sulindac	0
Kol 43	1	Captopril	0	Testosterone	0
Medroxyprogesterone	1	Carbamazepine	0	Tinidazole	0
Mifepristone	1	Carnitine	0	Trimethoprim	0
Nicardipine	1	Cefamandole	0	Valproic acid	0
Nitrendipine	1	Chloroquine	0	Warfarin	0
Simvastatin	1	Chlorzoxazone	0	Vincristine	0
Tipranavir	1	Cholic acid	0	Zidovudine	0
Verapamil	1	Cimetidine	0		

2.2 Calculation of Molecular Properties

The SMILES codes collected from online databases was added to an Excel spreadsheet. From Excel, the chemical name and SMILES code of all compounds in the data set was copied and pasted into a notepad file. It was then saved as ".txt" file because this is the format that can be read by the software. This file was imported into Advanced Chemistry Development, Inc. ACD Labs/ Log D Suite Version 12.0 (ACD/Labs) software which is an application used to calculate molecular properties. Some of the compounds are in salt form, so before calculations the desalt function of the software was used to remove the ionic forms of some of the compounds and minimise their charges. Molecular properties were calculated for 123 compounds and examples of descriptors calculated include pKa, LogD, LogP and molar volume. The structure of each compound was saved by ACD Labs/ Log D (ACDlabs) in SDF file format. It was saved this way as this will allow the Molecular Operating Environment (MOE) software to read the structures formed in ACDlabs and produce a 3D version of the structures.

2.3 Preparation of compounds for Docking

Before docking could take place, the SDF file was imported into the software MOE. MOE is a suite of applications that can be used to manipulate and analyse a collection of compounds. For docking to work efficiently, it is essential that each structure is in a form suitable for it to be docked to a ligand. As a result, the software's 'Wash' application was used to clean the structures and neutralise the protonation state of each compound. This will neutralise all atoms and form the structure of the compound in its least charge-bearing state. The next step was to

lower the potential energy of the structures. This was completed using the "Energy minimize" function from the software. The compounds in the database were now ready to be computed and molecular descriptors were calculated. In total, MOE calculated 320 descriptors for the compounds in the dataset.

2.4 Protein-Ligand Docking

Docking of compounds in the dataset was carried out using the Dock application in MOE. The structure of the P-gp protein that the compounds in our dataset would be docked to was downloaded from the Protein Data Bank (PDB) online. The X-ray structure of mouse P-gp, 3G5U (QZ59-RRR bound) and 3G61 (QZ59-RRR bound), were used for docking. The docking site for P-gp 3G5U was determined using residues that have been shown to interact with cyclic-peptide inhibitors (Aller, et al., 2009). In total, 4 docking sites and their interacting residues were identified for P-gp 3G5U; Verapamil site, QZ59-RRR, QZ59-SSS upper and QZ59-SSS lower sites. The docking site of P-gp 3G61 was determined using the co-crystallised ligand QZ59-SSS already bound to the protein. The amino acid residues involved in each of the sites has been reported in Table 2. In the MOE software, the default Triangle Matcher was used as the placement method followed by forcefield refinement and London dG scoring was used for the docking runs. The top scoring conformation of each compound for each binding site was calculated as well as the root mean square deviation of each pose (RMSD). The maximum number of poses kept after the rescoring stage was 30 and duplicates were also removed

2.5 Statistical Analysis

Results collected from docking scores and molecular descriptor calculations were analysed using data mining tools on Statistica 11.0 software. Data mining tools that were used for this study include Classification and Regression Tree (CART), Support Vector Machine and Interactive Tree (IT). CART is a statistical method that is used to partition data based on continuous dependent variables (regression) or categorical predictor variables (classification). In this study, CART was used to determine the importance of variables such as docking scores and molecular descriptors for classification of substrates and non-substrates.

In all CART models that were developed, substrate/non-substrate property (ABCB1-Pgp) was the dependent variable and in each model other variables acted as independent variables. The dependent variable was categorical i.e. 1 (substrate), 0 (nonsubstrate), therefore categorical analysis was performed. Interactive tree are similar to CART, however a specific independent variable e.g. docking scores at QZ59-SSS (lower) site was chosen manually as the first splitting variable and then the tree was allowed to grow further using statistically selected variables.

Table 2. Amino acid residues used for the definition of binding site

Binding site	Amino acid Residues	P-gp model
Verapamil	H60, A63, L64, S218, I302, L335, A338, F724, I864, G868, F938, T941, L971, V978, G980, A981	3G5U
QZ59-SSS(upper)	M68, F332, I336, Y949, F974, V978, A981	3G5U
QZ59-SSS (lower)	L300, Y303, F339, Q721, F724, L758, F833, F974, V978, A981, M982, G985, Q986, S989	3G5U
QZ59-RRR	M68, Y303, F332, L335, I338, F339, Q721, F724, F728, Y949, F974, S975, V978	3G5U
QZ59-SSS	M68, F332, I336, Y949, F974, V978, A981 L300, Y303, F339, Q721, F724, L758, F833, F974, V978, A981, M982, G985, Q986, S989	3G61

3 Results

3.1 Docking results

Docking energies for each compound was calculated by docking each compound to the three binding sites of 3G5U and the cyclic peptide binding site of 3G61. The lowest score for each compound was recorded and the average was calculated. The average of the lowest docking energies is shown in the graph below for all 123 compounds (Figure 5). The average docking energy for substrate and non-substrate is also shown (Figure 6).

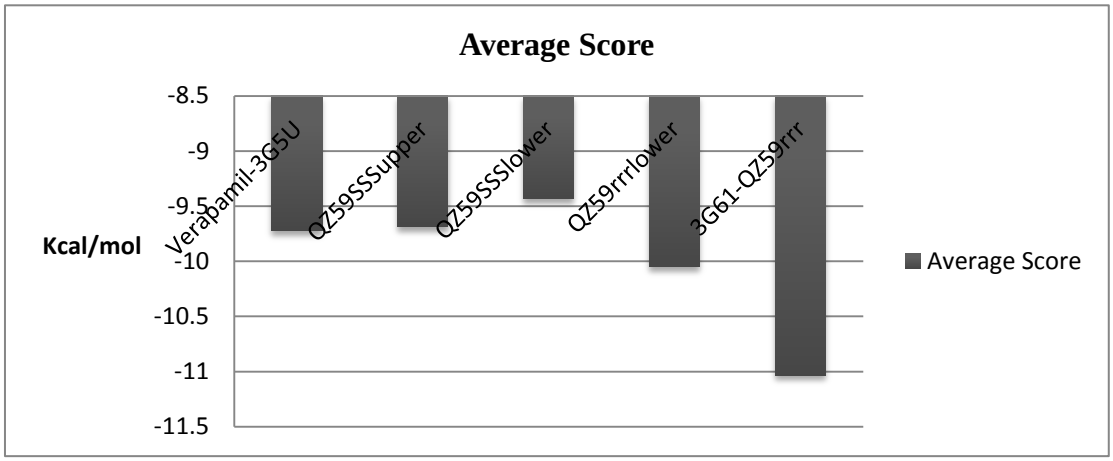


Figure 5. Average of the lowest docking scores

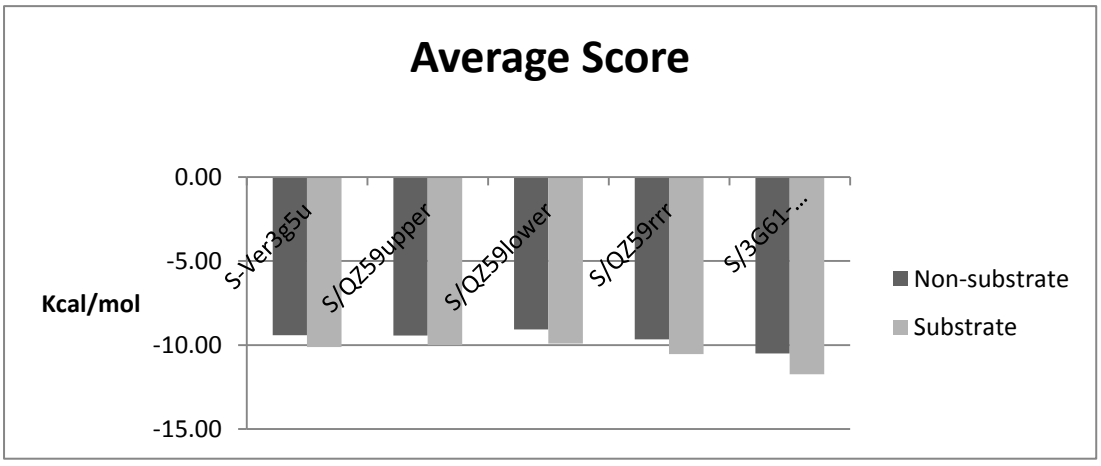


Figure 6. Average docking energy of substrate and non-substrates

Table 3. Docking energy of top three poses for all compounds

The table below shows the average value of the top three docking energies calculated for all compounds at each binding site.

Binding site	**Docking energy of top three scoring poses (kcal/mol)**			**Average (kcal/mol)**
Verapamil – 3G5U	-15.4	-12.97	-12.3	-13.6
QZ59SSSupper – 3G5U	-17.5	-15.3	-14.8	-15.9
QZ59SSSlower – 3G5U	-15.7	-15.7	-15.1	-15.5
QZ59RRRlower – 3G5U	-16.9	-15.6	-14.2	-15.6
QZ59SSS – 3G61	-17.6	-15.9	-15.7	-16.4

3.1.2 Docking Performance

After obtaining docking scores for each compound, docking energies were evaluated to determine which compounds had greater affinity for the binding sites of P-gp. The table on the following page shows the top ten compounds that had good docking performance at each binding site.

Table 4. Compounds with high docking performance

Compound	Verapamil	Compound	QZ59sss-upper	Compound	QZ59sss-lower	Compound	QZ59rrr	Compound	3G61
	Score		Score		Score		Score		Score
Ivermectin A	-15.407999	Ivermectin A	-17.49052	Ivermectin B	-15.702105	Cyclosporine-A	-16.93857	Cyclosporine-A	-17.6087
Bromosulfalein	-12.974476	Digoxin	-15.264123	Cyclosporine-A	-15.671002	Rifampicin	-15.607213	Rifampicin	-15.8589
Doxorubicin	-12.270447	Rifampicin	-14.776971	Ivermectin A	-15.053252	Ivermectin A	-14.21278	Bromosulfalein	-15.7715
Tetracycline	-12.25106	Ivermectin B	-14.182034	Digoxin	-13.845214	Taurocholate	-13.826485	Silymarin	-15.6766
Silymarin	-12.211308	Cyclosporine A	-13.913272	Erythromycin	-12.796752	Etoposide	-12.774714	Etoposide	-15.3638
Mitoxantrone	-12.069983	Silymarin	-12.740396	Rifampicin	-12.707092	Ivermectin B	-12.753691	Taurocholate	-15.2546
Methotrexate	-11.986519	Colchicine	-12.727611	Doxorubicin	-12.654102	Erythromycin	-12.736913	Doxorubicin	-14.8735
Tipranavir	-11.627663	Methotrexate	-12.722812	Bromosulfalein	-11.747195	Bromosulfalein	-12.696462	Vincristine	-14.4704
Estradiol-17β-glucuronide	-11.490866	Vincristine	-12.462447	Flupentixol	-11.41256	Digoxin	-12.67726	Reserpine	-14.4181
Budesonide	-11.293019	Bromosulfalein	-12.310234	Vinblastine	-11.387159	Silymarin	-12.592422	Digoxin	-14.2525

3.1.3 Docked compounds

Docking runs were performed for each binding site and docking energies calculated. Below are images of example compounds docked to different binding sites of 3G5U and 3G61.

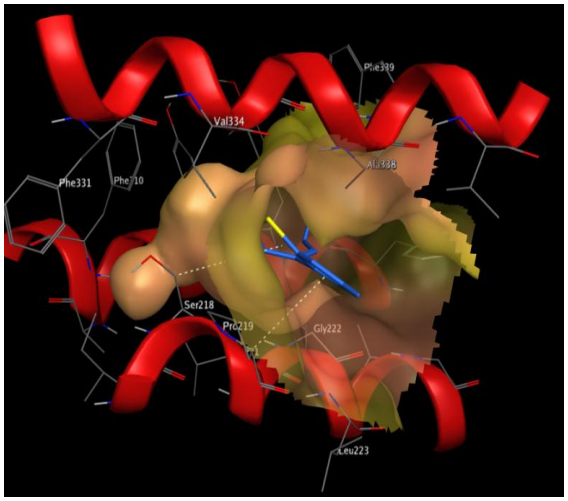


Figure 7. 17-β-estradiol (blue) docked to Verapamil binding site of 3G5U.

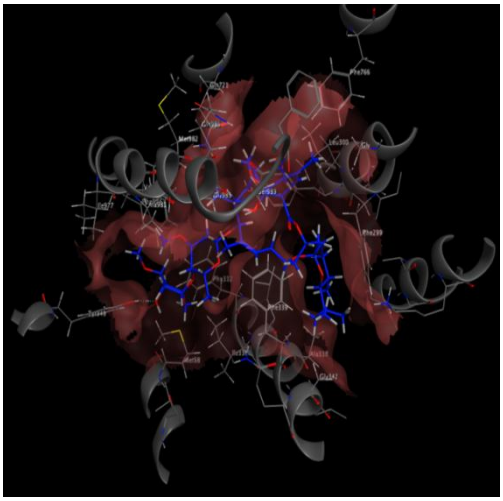


Figure 8. Ivermectin A (blue) docked to QZ59-SSS(lower) binding site of 3G5U.

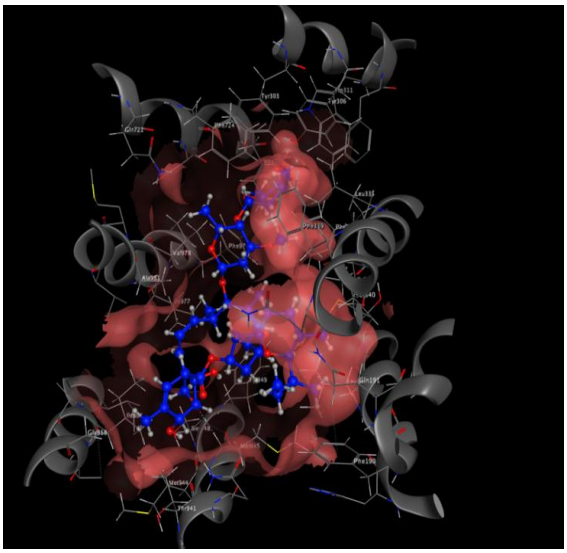


Figure 9. Ivermectin A (blue) docked to QZ59-SSS (upper) binding site of 3G5U

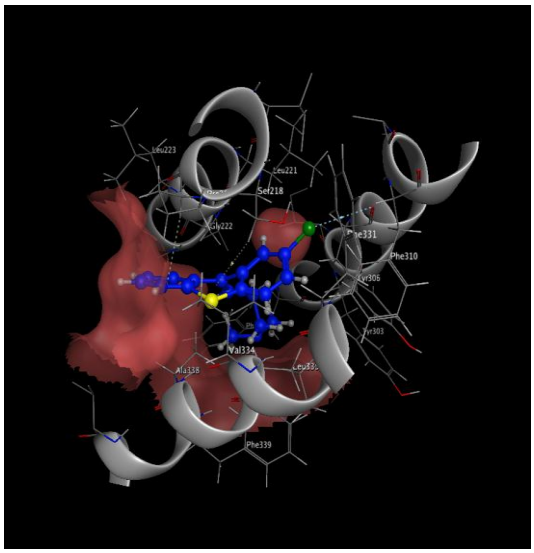


Figure 10. Chlorprotixene (blue) docked to binding site of 3G61

3.2 Statistical Models

3.2.1 Classification Trees (CART)

Classification trees were developed to make predictions about the classification of substrates and non-substrates. In all classification trees developed, substrate/non-substrate property (ABCB1-Pgp) was the dependent variable. Several models were developed because in each model particular variables were selected as the independent variable e.g. Docking scores. In order to control the splitting of the tree, the minimum number of cases was 49 and maximum cases were 1000.

CART 2 – Using all docking scores for each binding site as independent variables

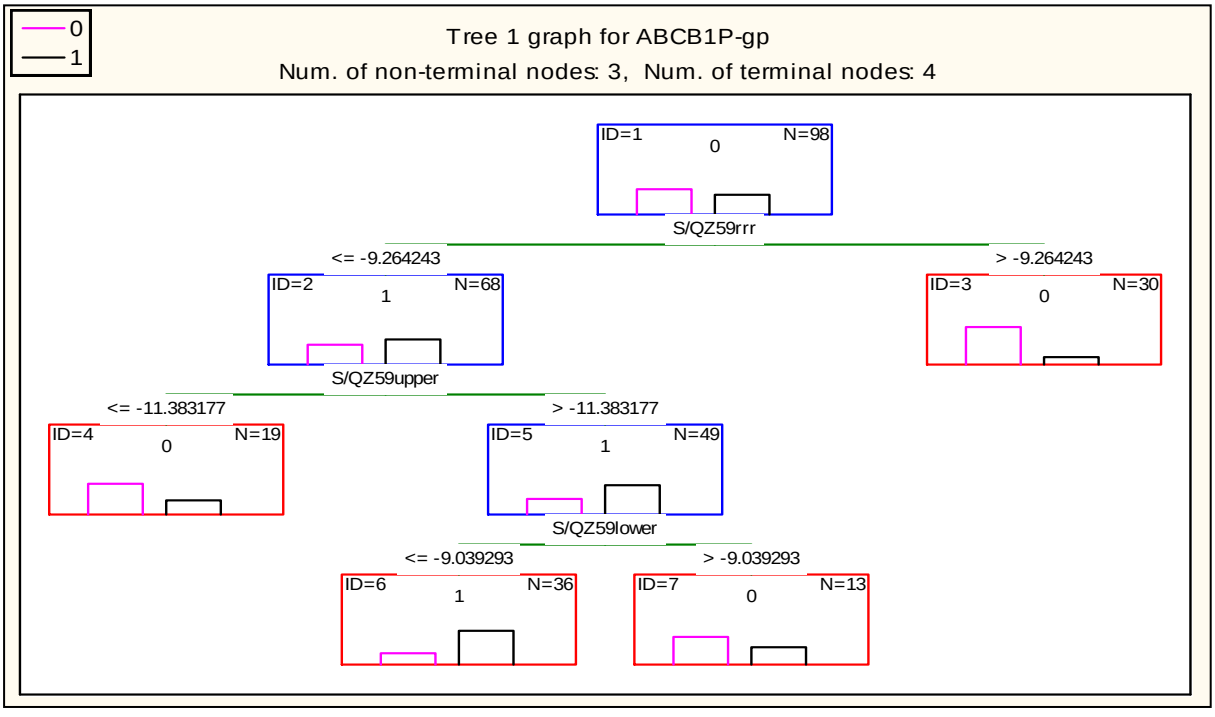


Figure 11 **Parameters:** 0 (Non-substrate), 1 (Substrate)
S/QZ59rrr – Docking score at the QZ59rrr binding site (3G5U)
S/QZ59upper – Docking score at the QZ59upper binding site (3G5U)
S/QZ59lower – Docking score at QZ59lower binding site (3G5U)

CART 3 – Using Docking score at 3G61-QZ59SSS site and RMSD as independent variables

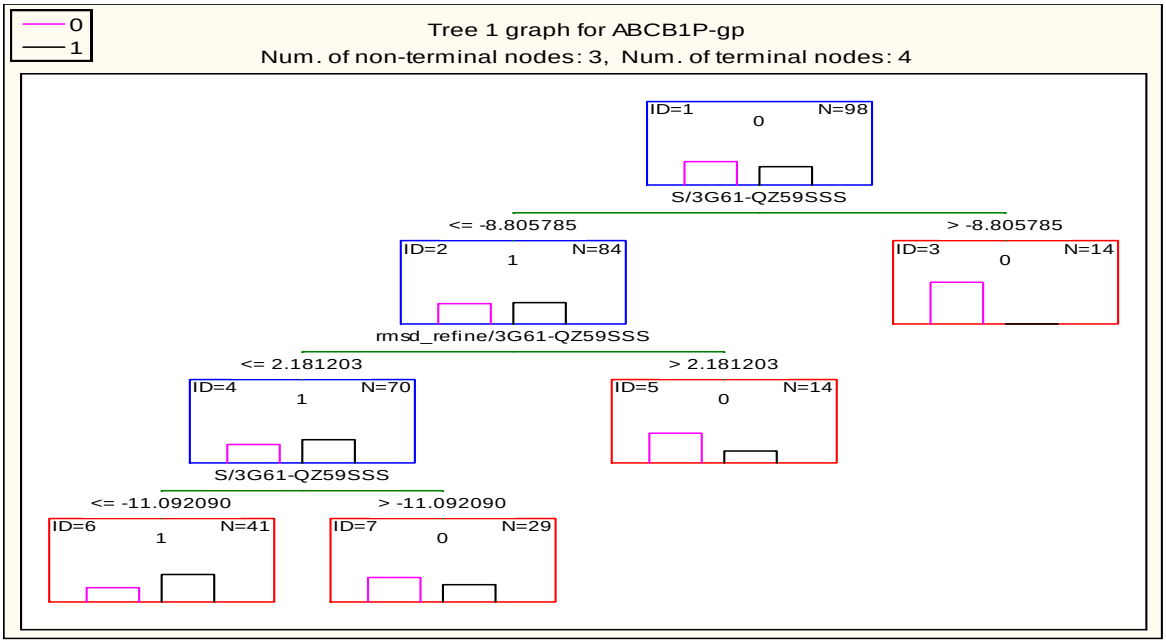


Figure 12
Parameters: 0 (Non-substrate) , 1 (Substrate) , S/3G61-QZ59RR – Docking score at the QZ59rrr site (3G61)
Rmsd_refine – The root mean square deviation of the pose

CART 4 – Docking scores of all binding sites and molecular descriptors

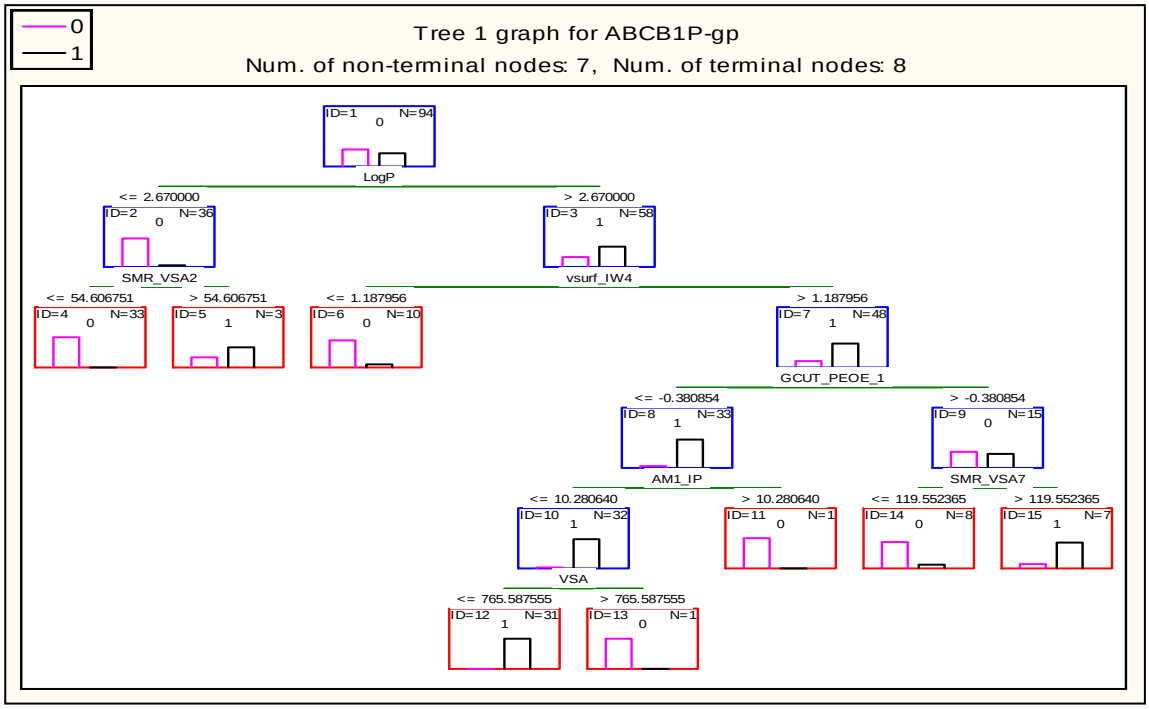


Figure 13
Parameters:
LogP - Log of the octanol/water partition coefficient,
SMR_VSA2 – Sum of van der Waals surface area (in $Å^2$) such that Molar Refractivity is in (0.26,0.35),
V surf_IW4 – Hydrophillic integy moment
GCUT_PEOE_1 - PEOE partial charge GCUT (1/3)
AM1_IP - The ionization potential (kcal/mol) calculated using the AM1 Hamiltonian [MOPAC]
SMR_VSA7 - Sum of van der Waals surface area (in $Å^2$) such that Molar Refractivity is > 0.56
VSA - Approximation to the sum of VDW surface areas

3.2.2 Interactive Trees

The aim of using interactive trees was to manually select one independent variable as the first splitting criteria and then allow the software to statistically select other important variables to grow the tree. In Figure 10, the first splitting criteria was the Docking score of all compounds at QZ59-RRR binding site (3G5U) and in Figure 11, Docking score at the QZ59lower binding site (3G5U) was the first splitting variable. Interactive tree models developed are shown below.

IT 1 – Docking score S/QZ59RRR dependent variable

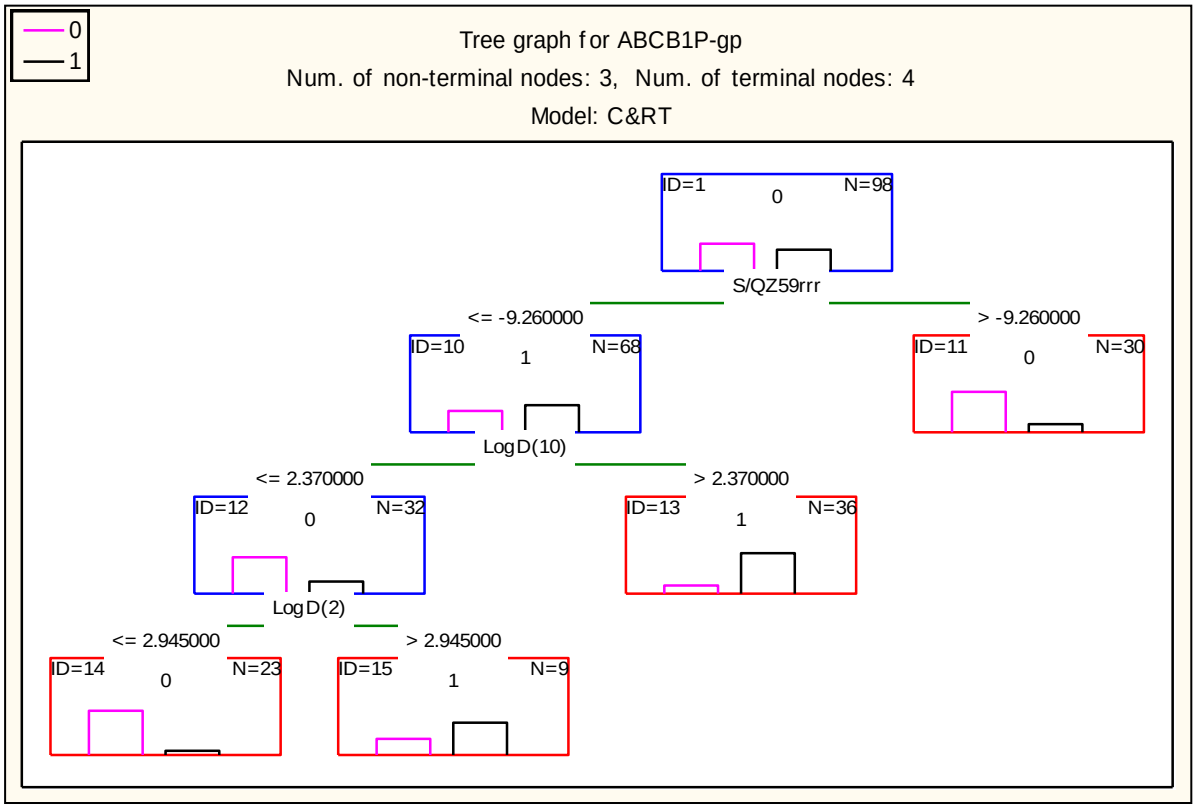


Figure 14
Parameters: S/QZ59rrr - Docking score at the QZ59-RRR binding site (3G5U) and is the Dependent variable
LogD – Log of the distribution coefficient

IT 2 – Docking score S/QZ59lower dependent variable

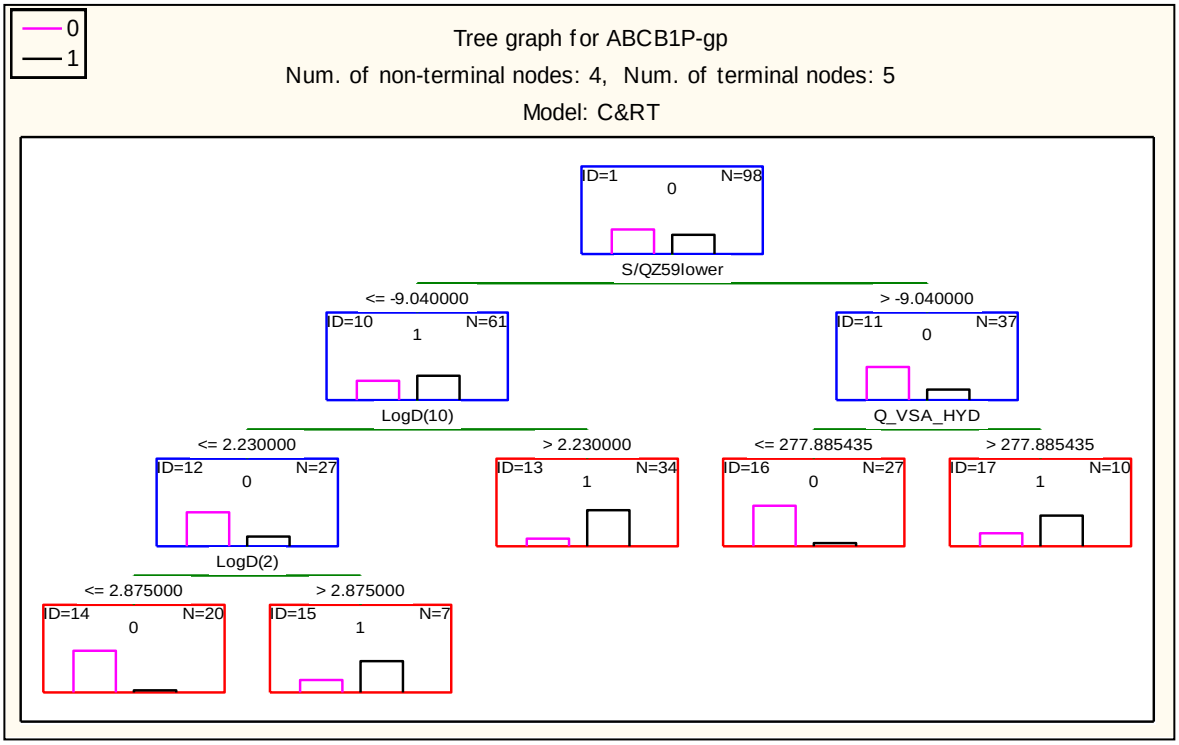


Figure 15
Parameters: S/QZ59rrr - Docking score at the QZ59-SSS (lower) binding site (3G5U) and is the Dependent variable
LogD – Log of the distribution coefficient
Q_VSA_HYD - Total hydrophobic van der Waals surface area.

3.2.3 Support Vector Machine Models

Further classification models were developed using Support Vector Machine. SVM Classification type 1 was used for all SVM calculations and Radial Basis Function (RBF) was selected as the kernel type. The independent variables were manually selected from the CART and Interactive trees developed.

Table 5. SVM Models

Model No.	Dependent	Independent	SVM Type	Kernel Type	Accuracy		
					Training	Test	Overall
SVM 1.	ABCB1-Pgp	S/3G61-QZ59RR S/QZ59upper S/QZ59RRR	Classification type 1 Capacity=7.00	Radial Basis function Gamma=0.33	64.3%	64%	64.2%
SVM 2.	ABCB1-Pgp	S/3G61-QZ59RR S/QZ59upper S/QZ59RRR S/Ver3g5u S/QZ59lower	Classification type 1 Capacity=10.0	Radial Basis function Gamma=0.2	69.4%	60%	67.5%
SVM 3.	ABCB1-Pgp	S/QZ59RRR LogD(2) LogD(10)	Classification type 1 Capacity=3.00	Radial Basis function Gamma=0.3	75.5%	76%	75.6%

3.3 Prediction accuracies of statistical models

The accuracy of predictions of all models developed is shown in Table 5 below. The accuracy of each model, Youden's J statistic and Matthews correlation coefficient (MCC) calculations were carried out to predict the accuracy of the models. The formulas for the calculations are also described below.

Table 6. Results of all statistical models

	Training Set					Validation Set				
Model	**Acc**	**SE**	**SP**	**Youden's J**	**MCC**	**Acc**	**SE**	**SP**	**Youden's J**	**MCC**
CART 1	0.64	0.88	0.45	0.34	0.36	0.68	1	0.43	0.43	0.50
CART 2	0.71	0.59	0.83	0.41	0.49	0.64	0.64	0.64	0.28	0.35
CART 3	0.69	0.64	0.73	0.38	0.38	0.6	0.36	0.79	0.15	0.17
CART 4	0.96	0.95	0.96	0.91	0.91	0.48	0.22	0.67	0.11	0.12
SVM 1	0.64	0.49	0.76	0.25	0.26	0.64	0.45	0.79	0.24	0.26
SVM 2	0.7	0.59	0.8	0.39	0.4	0.6	0.36	0.79	0.15	0.17
SVM 3	0.76	0.67	0.82	0.49	0.5	0.81	0.72	0.93	0.65	0.65
IT 1	0.83	0.83	0.83	0.66	0.66	0.81	0.78	0.83	0.61	0.61
IT2	0.85	0.93	0.79	0.72	0.71	0.81	0.67	0.92	0.58	0.61

Acc – Accuracy = $(TP + TN)/(TP + FP + TN + FN)$ **SE – Sensitivity**= $TP/\ (TP + FN)$

SP – Specificity= $TN/(TN + FP)$ **Youdens J** = $Sensitivity\ +\ Specificity$ – 1

MCC – Matthews correlation coefficient =

$$TP\ X\ TN - FP\ X\ FN/\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}$$

TP, TN, FP, and FN are True Positive, True Negatives, False Positive and False Negative respectively.

3.3.1 Accuracy vs. Matthews correlation coefficient

The accuracy and Matthews correlation coefficient calculations were compared for each model for the validation set. The result for the performance for each model is shown graphically below.

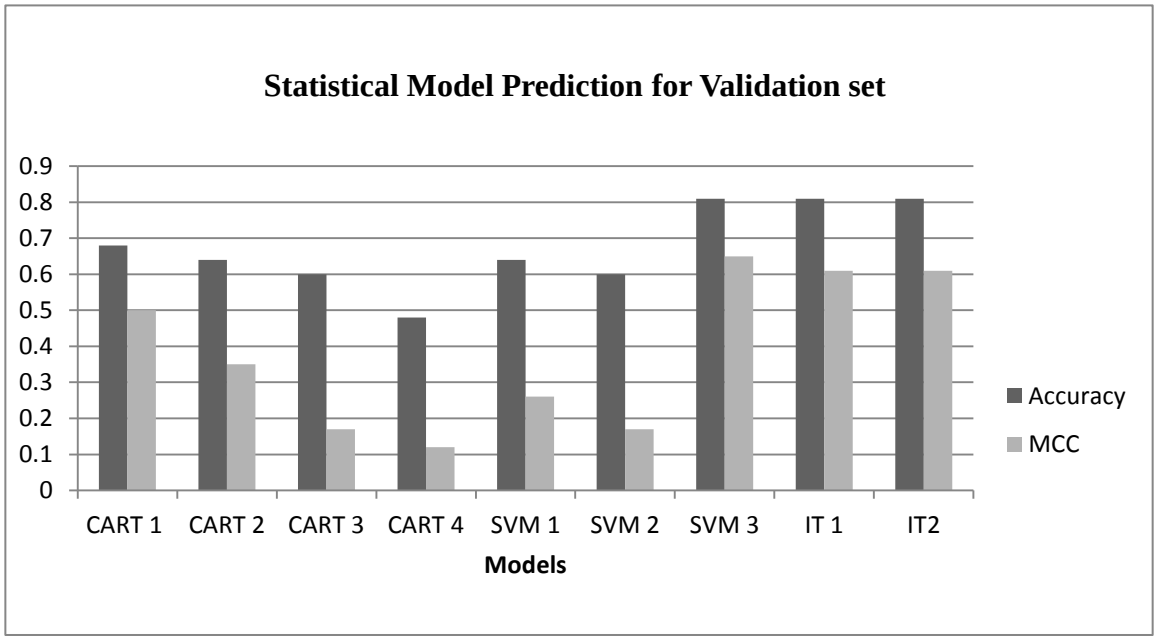


Figure 16. Graph showing performance of models for validation set

Overall the best model was SVM 3, due to a better performance of Youden's J and MCC for the validation set. As well as this IT 1 and IT2 models also showed good performance for both training and validation set.

4. Discussion

P-glycoprotein is a member of a super family of transporters expressed in cells of the liver, kidney and in the blood brain barrier. For it to function, ATP is required to transport molecules across cell membranes. The major role of this protein is to export drugs and metabolites out of cells and it has also been associated with multi-drug resistance in cancerous cells (Leslie, et al., 2009). The first objective of this study was to identify substrates and inhibitors of P-glycoprotein by evaluating their docking energies. At present, various studies have proposed a range of locations within the internal cavity of P-gp that compounds may bind to (Aller, et al., 2009; Gutmann, et al., 2009). To dock the compounds in our data set to P-gp, the binding sites of the protein had to be determined. Amino acid residues were selected from literature that had been demonstrated as binding sites of cyclic peptide inhibitors and Verapamil (Aller et al., 2009; Table 2) and compounds in the data set were docked to these areas.

Docking runs were carried out for each compound in the data set using MOE which calculated the docking energy for each pose generated. As shown in Figure 5, the average docking energy at the 3G61-QZ59SSS site was lowest (-11.1 kcal/mol) whilst compounds docked less favourably to the 3G5U binding sites. Table 3 shows that the average docking energy of the top 3 scores was superior at the binding site defined by 3G61-QZ59-SSS but average docking energy of the top three scores at the Verapamil site was -13.6 kcal/mol. These results showed that docking performance for compounds in the data set was considerably better when docked to the co-crystallised ligand in P-gp 3G61. Better docking performance in 3G61 could be due to the changed confirmation

of the protein upon ligand binding in accordance with the induced fit binding theory (Alonso, et al., 2006).

By docking to 3G61 (Figure 17), compounds were able to interact with more residues compared with other sites thus improving their chances of binding to the protein.

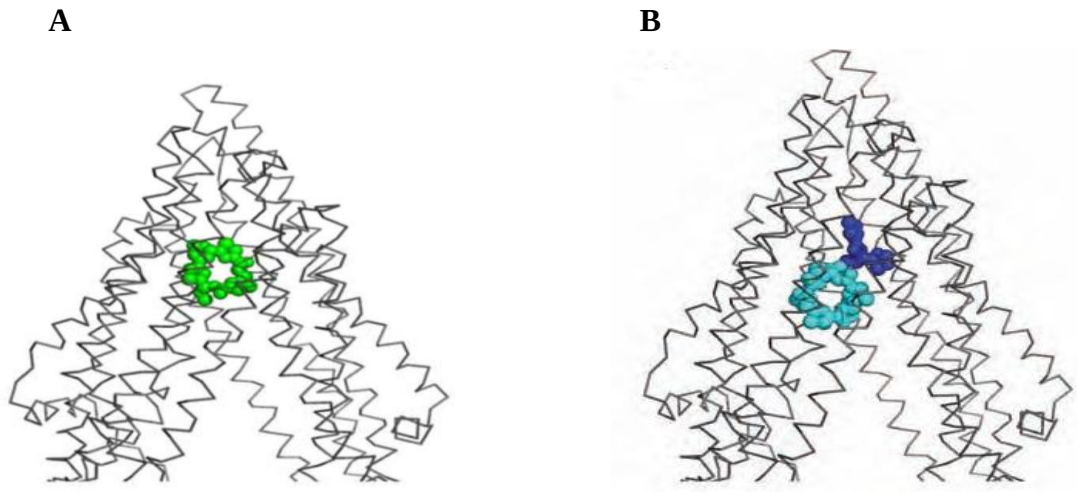


Figure 17 Location of QZ59 compounds in Pgp drug binding pocket. **A.** One QZ59-RRR molecule **B.** Two QZ59-SSS molecules. (Aller , et al., 2009)

A list of compounds with strong affinity for the binding sites of P-gp is shown in Table 4. Cyclosporin an immunosuppressant drug, was one of the compounds in the data set that had significantly better docking scores across all binding sites. As shown in Table 4, docking energy of Cyclosporin was lowest at 3G61 compared to other sites. As well as this, Rifampicin best docking energy at the QZ59-SSS upper site was -14.776971kcal/mol, and docking score at 3G61 was -15.8kcal/mol. Similarly, Bromosulfalein which had efficient docking performance across all binding sites had its lowest docking score when docked to 3G61 (Table 4).

This further suggests that docking to the ligand QZ59-SSS in P-gp-3G61 was the most favourable binding site for compounds in our data set compared with other binding sites that were used. As part of this study, it was also our aim to gain an understanding about the residues that play a major role in drug binding to P-gp. The differences in amino acid residues used at each binding site provides valuable information about the importance of some specific residues involved in drug binding to P-gp. The amino acid residues selected for each binding site were reported to be within the drug binding pocket of P-glycoprotein (Aller, et al., 2009). Some of the residues in particular binding sites appear to play a more important role in drug binding compared to others. An example of this is Tyr 303 a residue in TM helice 5 of P-gp, is part of the QZ59-RRR and QZ59-SSS (lower) site but not seen in the other two binding sites used. As well as this, Phe 332 is a residue in the QZ59-RRR and QZ59-SSS (upper) sites but not in the Verapamil binding site (Table 2). In addition to differences such as these, some residues appear to be involved in all binding sites used. For example, Phe 724 and Val 978 are part of each binding site and this suggests that they are important residues involved in the interaction of substrates with P-gp.

From Table 2, it can be seen that there are several other amino acid residues that are involved in binding compounds to the QZ59-RRR and QZ59-SSS sites. Examples of these residues include, Phe 339 (TM6), Gln 721 (TM7) which are in both QZ59 sites whilst residues such as Ala 981(TM12) and Leu 335 (TM6) are identifiable in both Verapamil and QZ59-RRR sites respectively. These observations suggest that different transmembrane helices are involved in drug binding and compounds use unique segments of the transmembrane domains to

bind to P-gp. It is also interesting to note the overlap of residues involved in drug-binding, which suggests that compounds in the data set may have similar or overlapping binding sites with other compounds (Ambudkar, et al., 2003).

To illustrate this, Rifampicin a bactericidal antibiotic was seen to demonstrate good docking performance across both QZ59 sites and when docked to P-gp 3G61 (Table 4). In contrast, when docked to the Verapamil binding site its best docking score was -10.88kcal/mol. As well as this, Erythromycin a macrolide antibiotic docked in a similar way to Rifampicin at the QZ59-SSS (lower) and QZ59-RRR sites but not as successfully to the Verapamil site (-11.14 kcal/mol). It could be suggested that the docking performance at these sites, is due to both compounds binding to the same or overlapping residues at these binding sites.

In addition to docking results observed, Figure 6 shows the difference between docking energies of substrates and non-substrates. The results here add further evidence to the reports that P-gp is a polyspecific protein which is capable of recognising different types of compounds as non-substrates were able to dock to P-glycoprotein. The results suggest that non-substrates in the data set are capable of binding to P-gp but have a weaker binding affinity. Despite these outcomes, limitations of the docking method discussed below should be considered when evaluating docking results.

The use of X-ray structures of 3G5U and 3G61in this study as docking targets may have had an influence on the accuracy of docking results produced. Apart from the fact that one of these structures belongs to the protein co-crystallised with a ligand and the other is free of a ligand, the resolutions of the crystal structures of P-gp models used is an important factor; 3.80Å and 4.35Å for

3G5U and 3G61 respectively. Generally, high resolution models of proteins are those considered as having a resolution lower or equal to 1.5Å, whilst low resolution models have values greater than 2.5Å (Davis, et al., 2003). In this case, the X-ray structures used in this study are of low resolution and therefore the accuracy of these structures is still uncertain. Structural models obtained at higher resolutions are more likely to produce better docking results (Mohan et al, 2005). This is because higher resolution models are developed using more experimental data whereas at lower resolutions, models are likely to be more subjective and include a greater number of errors (Davis, et al., 2003; Davis, et al., 2008). Consequently, docking results from this study may not be reproducible because we cannot be sure that the 3D structural models of P-gp are correct and validated.

Additionally, the scoring method used to calculate binding energies of the compounds would have had a major impact on the docking results. During the docking process, the top 30 poses produced after placement were scored using the London dG scoring function. As an empirical scoring function, it calculates the binding energy of compounds based on the sum of independent reactions from experimental data (Lipkowitz and Boyd, 2002). The issue with this is that the experimental data used to derive the scoring function may not be consistent with the data set used in this study therefore inaccuracies in scoring are likely. Furthermore, the scoring function is also more inclined to favour larger compounds.

Due to the additive nature of the formula, larger compounds are more likely to have better docking energies than smaller compounds (Schulz-Gasch and Stahl, 2004; Lipkowitz and Boyd, 2002). This is reflected in the results produced as compounds with high docking performance (Table 4) e.g. Bromosulfalein, Ivermectin A, Rifampicin, (794.03g mol^{-1}, 875.09g mol^{-1} and 822.94g mol^{-1}) were part of the heaviest compounds in the data set. In contrast, smaller compounds such as Valproic Acid (144.21 g mol^{-1}) and Amantadine (151.25 g mol^{-1}) had average docking scores of -7.04kcal/mol and -6.37kcal/mol respectively (Appendix 3).

The lack of flexibility of the target protein used in docking should also be taken into consideration when assessing docking results. The main purpose of the dock application in MOE, is to calculate docking energies between a rigid protein target and flexible ligand. The inflexible nature of the protein during docking highlights the fact that *in-silico* methods do not totally represent what occurs in biological systems. For docking results to successfully guide our predictions of inhibitors and substrates of P-gp, it should take into account the flexible nature of the receptor. Previous studies have described the importance of protein flexibility in P-gp ligand interactions (Davis, et al. , 1999; Teague, 2003; Loo, et al. , 2003; Loo, et al. , 2009) and the induced-fit mechanism that drives this phenomenon.

Induced fit mechanism explains the fact that both drug and protein are flexible, and can modify their shape to generate more favourable contacts (Alonso, et al., 2006). Current evidence demonstrates that P-glycoprotein is able to accommodate a wide range of substrates due to the mobile nature of its transmembrane helices (Ambudkar, et al., 2003; Loo, et al., 2003). From this

hypothesis, it is possible that compounds in the data set may not be correctly identified as substrates or inhibitors of P-gp, because the docking process does not allow the protein to be mobile and therefore some compounds are not recognised as a substrate in the drug binding pocket. Further to this, the use of only specific binding sites in the drug binding pocket did not allow us to fully explore the diverse nature of P-glycoprotein. Using only the QZ59 and Verapamil binding sites meant that compounds could only dock to these areas. It is possible that variable docking energies may have been produced if compounds were docked to other residues of P-gp (Gutmann, et al., 2009).

The aim of docking the data set of 123 compounds was to find out if scoring functions can correctly identify the substrates and non-substrates of the data set. After obtaining docking results from each binding site, statistical models were built using data mining tools in Statistica to explore the classification accuracy of the docking scores. In each classification tree built, the classification of compounds as substrates/non-substrates was used as the dependent variable. The CART 2 model was built using docking scores from each binding site as independent variables. By examining CART 2, it shows us that docking scores at QZ59-RRR site were of importance and that if docking energy is lower than -9.26423kcal/mol, compounds are classified as substrates.

The tree was further split according to the docking score at QZ59-SSS (upper) site, with substrates classified as those having docking scores above -11.383177kcal/mol. The final node of this tree established that compounds with docking energy below -9.039293kcal/mol at QZ59-SSS (lower) site are substrates of P-glycoprotein. The disparity in the final two nodes may be related

to the differences in amino acid residues at each site and the fact that some compounds have overlapping binding sites (Ambudkar, et al., 2003) hence their binding to P-glycoprotein extends beyond the QZ59-SSS (upper) site.

The CART 3 model was built using docking scores and RMSD scores at the QZ59-SSS site of 3G61. This was performed because average docking scores for compounds in the data set was much better at this site. Consequently, the tree was split into three nodes with docking score being most important. The tree was then split according to RMSD scores and defined substrates as those having an RMSD below 2.18120kcal/mol as a substrate. Substrates meeting this criteria were also classified as a substrate if their docking score was lower than -11.092090kcal/mol in the final node of the tree.

A further tree was developed to detect the most important factor for P-gp binding, by combining docking scores and molecular descriptors calculated by ACD Labs. A total of 320 molecular descriptors were available. The model built using these variables (CART 4), selected LogP as the main descriptor required for P-gp binding. Substrates were classified as having a LogP above 2.67 and were further classified as substrates by various descriptors such as ionization potential and sum of Van der Waals surface area. Average LogP of compounds in the data set was 2.91 and 56% of compounds in the data set had a LogP value above 2.67. From this classification tree, it can be suggested that lipophilicity is an essential part of a compounds ability to bind to P-gp.

This is in agreement with previous studies that have described LogP as an important parameter in drug binding to P-gp (Wang, et al., 2003; Matsson, et al., 2009; Aller, et al., 2009). The significance of LogP for this data set is due to the

location of the residues in the binding sites. Amino acid residues in the binding sites were mostly located in the upper section of P-gp and this area purportedly contains hydrophobic and aromatic residues (Aller, et al., 2009). Therefore, it is likely that lipophilic substrates will bind to P-gp using these residues. It is also considerable that Van der Waals surface area factors, (SMR_VSA2, SMR_VSA7, and VSA) were also used to determine the classification of compounds.

The cyclic peptide inhibitors used to detect binding sites in P-glycoprotein, mainly interacted with residues by hydrophobic and Van der Waals interactions (Aller et al, 2009). Compounds in this data set seem to follow that trend according to CART 4 and it is has also been suggested that drug binding to P-gp is associated with compounds Van der Waals surface area (Litman, et al., 1997). Despite combining docking scores with molecular descriptors to build CART 4 model, molecular descriptors were selected as being of importance rather than the docking energy. From this outcome, it can be suggested that molecular descriptors are better predictors of a compounds class as a substrate or non-substrate.

After assessing results of the classification trees, Interaction trees were also developed (Figure 14 & 15). In each tree developed, one variable was manually selected as being of importance for the first splitting criteria and then the software statistically selected other variables to grow the tree. In both trees, the software selected LogD(2) and LogD(10) as other important attributes for classification of substrates and non-substrates. This is in accordance with suggestions that lipophilicity plays a major role in P-gp activity.

SVM models were then developed to further classify substrates and non-substrates of P-glycoprotein (Table 5). SVM 1 was built using docking scores from 3G61, QZ59 (upper) and QZ59-RRR sites of 3G5U. This model had an overall accuracy of 64.2% whereas SVM 2 had an improved accuracy of 67.5% when using all docking scores. SVM 3 model had an overall accuracy of 75.6% and this was based on docking score at the QZ59-RRR site of 3G5U and LogD(2) and LogD(10). LogD was used to develop this model because it was a valuable descriptor selected by the Interaction trees developed (Figure 14 & 15).

In the validation set, there are 5 substrates and 1 non substrate that had been misclassified by SVM 3 model; Apigenin, Imatinib, Prednisone, Progesterone, Quercetin, Taurocholate (Appendix 5). Out of this group, Progesterone a non-substrate was classified as a substrate by the SVM model. It is possible that it was classed as a substrate by the SVM model due to its high lipophilicity (LogP = 3.83), which has previously been discussed as an important property of P-gp substrates. Further complementing this are suggestions that Progesterone is a substrate by an induced fit mechanism (Loo, et al., 2003). However, differences in data preparation and method used for classification indicate that the identification of Progesterone as a substrate or non-substrate is still uncertain in the literature. Other compounds that were misclassified include Apigenin and Quercetin which are Flavanoids that have been classified as non-substrates by SVM. Similar to progesterone, misclassification could have occurred based on lipophilicity with both compounds exhibiting low LogP values (Appendix 3). Results from the Matsson study, suggest that flavanoids interact with the ATP-binding region of P-gp (Matsson, et al., 2009). Compounds did not interact with

this segment of P-gp during docking and so it could be suggested that these two compounds were misclassified by the model as a result of this. Overall, Table 6 shows a summary of accuracy performance by all models and this shows that the best performing model was SVM 3, due to a better performance of Youden's J and MCC for the validation set.

In addition to limitations discussed previously, accuracy of the models developed in this study could have been improved if a larger data set was used. A larger dataset would have provided more valuable information and this is necessary to build models of better quality (Chen, et al., 2011). Using the data set from the Matsson paper will have affected results as the data collected in this paper was based on human P-gp, whilst the structures used for docking are of mouse P-gp which has 87% sequence identity to human p-glycoprotein (Aller, et al., 2009). By forming a data set in this way reduces the reliability of the data set and therefore the class of compounds should be checked against other sources. Finally, docking energies were only calculated using one docking program. It would have been interesting to compare docking energies using other docking software such as Glide, Gold and FlexX.

Conclusions

In this study, a data set of 123 compounds was docked to P-gp structures available from the protein data bank. Docking results revealed that compounds had better docking performance at the QZ59-SSS binding site in P-gp 3G61 rather than the binding sites in 3G5U. The amino acid residues involved in all sites showed that some amino acid residues overlap and this suggests that compounds may have the same or overlapping residues in their binding sites. This also complements other studies, which suggest that P-gp is a polyspecific protein, in which a diverse range of compounds are able to bind to it. However, it is important to understand that the structures described by Aller et al, merely represent the authors view of the P-gp structure obtained in their study. For this reason, we should consider that the accuracy of this structure is still uncertain and coupled with the relatively low resolution of the structure, the use of these protein structures for docking is limited.

Despite this, results from this study do provide evidence about the amino acid residues that are important in drug binding e.g. Phe 724 and Val 978. In addition to this, docking results appear to be better when the protein structure conformation has already been changed in order to allow binding of a ligand as was observed from docking results of P-gp 3G61 co-crystallised with QZ59-SSS. Other significant results from this study also show that LogP is a major contributor to compounds availability and compounds with high LogP are more likely to be able to bind to P-glycoprotein.

Future work

The findings from this study do allow us to propose suggestions for future work. Amino acid residues that were part of all binding site sequences Phe 724 and Val 978 could possibly be a target for inhibitors of P-glycoprotein. These residues appear to be vital for P-gp interactions and inhibitors could aim to covalently bind to residues such as these, therefore disrupting P-gp function. The lack of high resolution models has severely limited work in this field but if higher resolution models of P-gp were made available, this would greatly improve the identification of binding sites within P-glycoprotein. Higher resolution models will also improve docking energies and allow us to visualise the interactions between P-gp and compounds. After combining docking scores and molecular descriptors to build models in this study, this could possibly be the way forward in improving identification of P-gp substrates and inhibitors. By using docking results, molecular descriptor calculations and results from other *in-silico* methods such as QSAR and pharmacophore modelling, this may provide better outcomes for identification of P-gp substrates and inhibitors. Although this may prove challenging, there are studies which suggest that combining methods together to build a global model can be beneficial (Li, et al., 2007; Pajeva, et al., 2009).

References

ACD Labs/ Log D Suite Version 12.0 Advanced Chemistry Development, Inc., Toronto, On, Canada, www.acdlabs.com, 2012.

Aller SG., Yu J., Ward., Weng., Chittaboina S., Zhuo., Harrell PM., Trinh YT., Zhang Q., Urnatsch IL., Chang G., 2009. Structure of P-glycoprotein reveals a molecular basis for poly-specific drug binding. *Science*; 323(5922): p1718-22.

Alonso H., Blizniyuk AA., Gready JE., 2006. Combining Docking and Molecular dynamic simulations in drug design. *Medicinal research reviews*; 26(5):p531-68.

Ambudkar SV., Kimchi-Sarfaty C., Sauna ZE., Gottesman MM., 2003. P-glycoprotein: from genomics to mechanism. *Oncogene*; 22(47):p7468-85.

Anderle P., Huang Y., Sadee W., 2004. Intestinal membrane transport of drugs and nutrients: Genomics of membrane transporters using expression microarrays. *European Journal of Pharmaceutical Science* , 21:p17–24.

Borst P., Elferink RO., 2002. Mammalian ABC transporters in health and disease. *Annual review of biochemistry*; 71:p537-92.

ChemSpider, 2012. *Royal Society of Chemistry. [online] Available at:* http://www.chemspider.com

Chen L., Li Y., Yu H., Zhang L., Hou T., 2011. Computational models for predicting substrates or inhibitors of P-glycoprotein. *Drug discovery today*;17(7-8):p343-51.

Corbeil CR., Williams CI., Labute P., 2012. Variability in docking success rates due to dataset preparation. *Journal of computer-aided molecular design*; 26(6):p775-86.

Cummings MD., DesJarlais RL., Gibbs AC., Mohan V., Jaeger EP., 2005. Comparison of Automated Docking Programs as Virtual Screening Tools. *Journal of medicinal chemistry* 48(4):p962-76.

Davidson AL., Dassa E., Orelle C., Chen J., 2008. Structure, Function, and Evolution of Bacterial ATP-Binding Cassette Systems. *Microbiology and molecular biology reviews*; 72(2):p317-64.

Davis AM., St-Gallay SA., Kleywegt GJ., 2008. Limitations and lessons in the use of X-ray structural information in drug design. *Drug Discovery today;*13(19-20):p831-41.

Davis AM., Teague SJ., 1999. Hydrogen Bonding, Hydrophobic Interactions, and Failure of the Rigid Receptor Hypothesis. *Angewandte Chemie (International ed. In English)*; 38:p736-749.

Davis AM., Teague SJ., Kleywegt GJ., 2003. Application and Limitations of X-ray Crystallographic Data in Structure-Based Ligand and Drug Design. *Angewandte Chemie (International ed. In English);* 23;42(24):p2718-36.

de Lannoy IA., Silverman M., 1992. The MDR1 gene product, P-glycoprotein, mediates the transport of the cardiac glycoside, digoxin. *Biochemical and biophysical research communications*; 189(1):p551-7.

Dean M.,, 2002. The Human ATP-Binding Cassette (ABC) Transporter Superfamily. 2002 Nov 18. In: Dean M. *The Human ATP-Binding Cassette (ABC) Transporter Superfamily* [Internet] Available from: http://www.ncbi.nlm.nih.gov/books/NBK31

Deconinck E., Hancock T., Coomans D., Massart DL., Heyden YV., 2005. Classification of drugs in absorption classes using the classification and regression trees (CART) methodology. *Journal of pharmaceutical and biomedical analysis*; 39(1-2):p91-103.

Ekins S., Mestres J., Testa B., 2007. In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling. *British journal of pharmacology*; 152(1):p9-20.

Gottesman MM., Ambudkar, SV., Xia D., 2009. Structure of a multidrug transporter, *Nature: Biotechnolgy*; *27*:546-547.

Gutmann DA., Ward A., Urbatsch IL., Chang G., van Veen HW., 2009. Understanding polyspecificity of multidrug ABC transporters: closing in on the gaps in ABCB1. *Trends in biochemical sciences*; 35(1):p36-42.

Hand D., Mannila H., Smyth P., 2001. *Principles of Data Mining*. Massachusetts: The MIT Press.

Hayouka Z., Hurevich M., Levin A., Benyamini H., Iosub A., Maes M., Shaley DE., Loyter A., Gilon C., Friedler A., 2010. Cyclic peptide inhibitors of HIV-1 integrase derived from the LEDGF/p75 protein. *Bioorganice & medicinal chemistry*; 18(23):p8388-95.

Higgins., 2001. ABC transporters: physiology, structure and mechanism – an overview. *Research in Microbiology*; 152(3-4):p205-10.

Huang., 2007. *Drug Discovery Research: New Frontiers in the Post-Genomic Era*. New Jersey: Wiley-Blackwell.

Iwata Y., Arisawa M., Hamada R., Kita Y., Mizutani MY., Tomioka N., Itai A., Miyamoto S., 2001. Discovery of novel aldose reductase inhibitors using a protein structure-based approach: 3D-database search followed by design and synthesis. *Journal of medicinal chemistry*; 44(11):p1718-28.

Juliano RL., Ling. V., 1976. A surface glycoprotein modulating drug permeability in Chinese hamster ovary cell mutants. *Biochim et Biophysica Acta*;455(1): p152- 162.

Klebe G., 2006. Virtual ligand screening: strategies, perspectives and limitations. *Drug discovery today*; 11(13-14):580-94.

Kantardzic M., 2011**.** *Data Mining: Concepts, Models, Methods, and Algorithms*. New Jersey: John Wiley & Sons, Inc

Leslie EM., Deeley RG., Cole SP., 2009. Multidrug resistance proteins: role of P-glycoprotein, MRP1, MRP2, and BCRP (ABCG2) in tissue defense. *Toxicology and applied pharmacology*; 204(3):p216-37.

Li WX., Li L., Eksterowicz J., Ling XB., Cardozo M., 2007. Significance analysis and multiple pharmacophore models for differentiating P-glycoprotein substrates. *Journal of Chemical Information and Modelling*; 47:p2429–2438.

Lipkowitz KB., Boyd DB., 2002. *Reviews in computational chemistry,Volume 18*. New Jersey: John Wiley & Sons, Inc.

Lipkowitz KB., Cundari TR., 2007. *Reviews in computational chemistry,Volume 23.* New Jersey: John Wiley & Sons, Inc.

Litman T., Zeuthen T., Skovsgaard T., Stein WD., 1997. Structure-activity relationships of P-glycoprotein interacting drugs: kinetic characterization of their effects on ATPase activity. *Biochimica et biophysica acta*; 1361(2):p159-68.

Lodish H., Berk A., Zipursky SL., Matsudaira P., Baltimore D., Darnell J., 2000. Molecular Cell Biology. 4th edition. New York: W. H. Freeman; 2000 Section 15.2, Overview of Membrane Transport Proteins

Loo TW., Bartlett MC., Clarke DM., 2003. Substrate-induced conformational changes in the transmembrane segments of human P-glycoprotein. Direct evidence for the substrate-induced fit mechanism for drug binding. *The Journal of biological chemistry*; 278(16):p13603-6.

Loo TW., Bartlett MC., Clarke DM., 2009. Identification of Residues in the Drug Translocation Pathway of the Human Multidrug Resistance P-glycoprotein by Arginine Mutagenesis. *The Journal of biological chemistry*; 284(36):p24074-87.

Loo TW., Clarke DM., 2008. Mutational Analysis of ABC proteins. *Archives of biochemistry and biophysics*; 476(1):p51-64.

Matsson P., Pedersen JM., Norinder U., Bergström CA., Artursson P., 2009. Identification of novel specific and general inhibitors of the tree major human ATP-binding cassette transports P-gp, BCRP and MRP2 among registered drugs. *Pharmaceutical Research*; 26(8): p1816-31.

Mohan V., Gibbs AC., Cummings MD., Jaeger EP., DesJarlais RL., 2005. Docking: Successes and challenges. *Current Pharmaceutical Design*; 11:p323-333.

Molecular Operating Environment (MOE), 2012.10; Chemical Computing Group Inc., 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, **2012**

National Center for Biotechnology Information. PubChem Compound Database, 2012. [online] Available at: http://pubchem.ncbi.nlm.nih.gov/

Pajeva IK., Globisch C., Wiese M., 2009. Combined pharmacophore modeling, docking, and 3D QSAR studies of ABCB1 and ABCC1 transporter inhibitors. *ChemMedchem*; 4(11):p1883–1896.

Rosenberg MF., Callaghan R., Ford RC., Higgins CF., 1997. Structure of the Multidrug Resistance P-glycoprotein to 2.5 nm Resolution Determined by Electron Microscopy and Image Analysis. *The Journal of biological chemistr*y; 272(16):p10685-94.

Sarkadi B., Homolya L., Szakacs G., Varadi A., 2006. Human Multidrug Resistance ABCB and ABCG Transporters: Participation in a Chemoimmunity Defense System. *Physiological reviews*; 86(4): p1179-236.

Schinkel AH., Jonker JW., 2012. Mammalian drug efflux transporters of the ATP binding cassette (ABC) family: an overview. *Advance drug delivery reviews;* 55(1):p3-29

Schulz-Gasch T., Stahl M., 2004. Scoring functions for protein–ligand interactions: a critical perspective. *Drug Discovery Today: Technologies*; p231-239.

Sharom FJ., 2008. ABC multidrug transporters: structure, function and role in chemoresistance. *Pharmacogenomics*; 9:p105–127.

Sousa SF., Fernandes PA., Ramos MJ., 2006. Protein–Ligand Docking: Current Status and Future Challenges. *Proteins*; 65(1):p15-26.

Teague SJ., 2003. Implications of protein flexibility for drug discovery. *Nature Reviews: Drug Discovery*; 2:p527-541.

Van de Waterbeemd H., Rose S., 2003. *The Practice of Medicinal Chemistry Second edition.* Oxford: Academic Press.

Wang L., 2005. *Support Vector Machines: Theory and Applications*. New York: Springer

Wang RB., Kuo CL., Lien LL., Lien EJ., 2003. Structure-activity relationship: analyses of p-glycoprotein substrates and inhibitors. *Journal of clinical pharmacy and therapeutics*; 28(3):p203-228

APPENDIX

Appendix 1: SMILES CODE

Compound	SMILES CODE
Chlorprotixene	Clc2cc1C(\c3c(Sc1cc2)cccc3)=C/CCN(C)C
Cyclosporine-A	O=C1N(C)[C@H](C(=O)N[C@H](C(=O)N(C)CC(=O)N(C)[C@H](C(=O)N[C@H](C(=O)N(C)[C@H](C(=O)N[C@H](C(=O)N[C@@H](C(=O)N([C@H](C(=O)N(C)[C@H](C(=O)N(C)[C@H]1C(C)C)CC(C)C)CC(C)C)C)C)CC(C)C)C(C)C)CC(C)C)CC)[C@H](O)[C@H](C)C/C=C/C
Diethylstilbestrol	Oc2ccc(/C(=C(/c1ccc(O)cc1)CC)CC)cc2
Dipyridamole	n3c(nc2c(nc(nc2N1CCCCC1)N(CCO)CCO)c3N4CCCCC4)N(CCO)CCO
Flupentixol	FC(F)(F)c2cc1C(\c3c(Sc1cc2)cccc3)=C/CCN4CCN(CCO)CC4
GF120918	O=C2c1cccc(c1Nc3c2cccc3OC)C(=O)Nc4ccc(cc4)CCN6Cc5c(cc(OC)c(OC)c5)CC6
Isradipine	O=C(OC)\C3=C(\N\C(=C(\C(=O)OC(C)C)C3c1cccc2nonc12)C)C
Ivermectin A	CO[C@H]1C[C@@H](O[C@@H](C)[C@@H]1O)O[C@@H]2[C@@H](OC)C[C@@H](O[C@H]2C)O[C@@H]3C(C)=CC[C@@H]6C[C@H](OC(=O)[C@@H]4C=C(C)[C@@H](O)[C@H]5OCC(=CC=C[C@@H]3C)[C@@]45O)C[C@@]7(O6)CC[C@H](C)[C@H](O7)[C@@H](C)CC
Ivermectin B	CO[C@H]1C[C@@H](O[C@@H](C)[C@@H]1O)O[C@@H]2[C@@H](OC)C[C@@H](O[C@H]2C)O[C@@H]3C(C)=CC[C@@H]6C[C@H](OC(=O)[C@@H]4C=C(C)[C@@H](O)[C@H]5OCC(=CC=C[C@@H]3C)[C@@]45O)C[C@@]7(O6)CC[C@H](C)[C@H](O7)C(C)C
Loperamide	Clc1ccc(cc1)C4(O)CCN(CCC(c2ccccc2)(c3ccccc3)C(=O)N(C)C)CC4
Lopinavir	O=C(N[C@@H](Cc1ccccc1)[C@@H](O)C[C@@H](NC(=O)[C@@H](N2C(=O)NCCC2)C(C)C)Cc3ccccc3)COc4c(cccc4C)C
MK571	CN(C)C(=O)CCSC(SCCC(=O)O)c1cccc(c1)/C=C/c2ccc3ccc(Cl)cc3n2
Quercetin	O=C1c3c(O/C(=C1/O)c2ccc(O)c(O)c2)cc(O)cc3O
Reserpine	O=C(OC)[C@H]6[C@H]4C[C@@H]3c2nc1cc(OC)ccc1c2CCN3C[C@H]4C[C@@H](OC(=O)c5cc(OC)c(OC)c(OC)c5)[C@@H]6OC
Ritonavir	CC(C)c4nc(CN(C)C(=O)N[C@@H](C(C)C)C(=O)N[C@@H](Cc1ccccc1)C[C@H](O)[C@H](Cc2ccccc2)NC(=O)OCc3cncs3)cs4
Saquinavir	O=C(N)C[C@H](NC(=O)c1nc2c(cc1)cccc2)C(=O)N[C@@H](Cc3ccccc3)[C@H](O)CN5[C@H](C(=O)NC(C)(C)C)C[C@@H]4CCCC[C@@H]4C5
Silymarin	O=C4c5c(O)cc(O)cc5O[C@H](c2ccc1O[C@H]([C@@H](Oc1c2)c3ccc(O)c(OC)c3)CO)[C@H]4O
Tamoxifen	O(c1ccc(cc1)/C(c2ccccc2)=C(\c3ccccc3)CC)CCN(C)C

Terfenadine	OC(c1ccccc1)(c2ccccc2)C4CCN(CCCC(O)c3ccc(cc3)C(C)(C)C)CC4
Thioridazine	S(c2cc1N(c3c(Sc1cc2)cccc3)CCC4N(C)CCCC4)C
Benzbromarone	Brc1cc(cc(Br)c1O)C(=O)c2c3ccccc3oc2CC
Amiodarone	Ic1cc(cc(I)c1OCCN(CC)CC)C(=O)c2c3ccccc3oc2CCCC
Apigenin	O=C\1c3c(O/C(=C/1)c2ccc(O)cc2)cc(O)cc3O
17ß-estradiol	C[C@]12CC[C@H]3[C@H]([C@@H]1CC[C@@H]2O)CCC4=C3C=CC(=C4)O
Biochanin A	O=C\1c3c(O/C=C/1c2ccc(OC)cc2)cc(O)cc3O
Chlorpromazine	CN(C)CCCN1c2ccccc2Sc3c1cc(cc3)Cl
Chrysin	O=C\1c3c(O/C(=C/1)c2ccccc2)cc(O)cc3O
Ergocristine	O=C3N1CCC[C@H]1[C@]2(O)O[C@](C(=O)N2[C@H]3Cc4ccccc4)(NC(=O)[C@@H]8/C=C7/c5cccc6c5c(cn6)C[C@H]7N(C)C8)C(C)C
Felodipine	O=C(OCC)\C1=C(\N/C(=C(/C(=O)OC)C1c2cccc(Cl)c2Cl)C)C
Gefitinib	COc1cc2c(cc1OCCCN3CCOCC3)c(ncn2)Nc4ccc(c(c4)Cl)F
Genistein	Oc1ccc(cc1)C\3=C\Oc2cc(O)cc(O)c2C/3=O
Glibenclamide	COc1ccc(Cl)cc1C(=O)NCCc2ccc(cc2)S(=O)(=O)NC(=O)Nc3ccccc3
Imatinib	Cc3ccc(cc3Nc1nc(ccn1)c2cccnc2)NC(=O)c4ccc(cc4)CN5CCN(C)CC5
Ketoconazole	O=C(N5CCN(c4ccc(OC[C@@H]1O[C@](OC1)(c2ccc(Cl)cc2Cl)Cn3ccnc3)cc4)CC5)C
Kol 43	O=C(OC(C)(C)C)CC[C@@H]1NC(=O)[C@H]4N(C1=O)[C@H](c3c(c2ccc(OC)cc2n3)C4)CC(C)C
Medroxyprogesterone	O=C4\C=C2/[C@]([C@H]1CC[C@@]3([C@@](O)(C(=O)C)CC[C@H]3[C@@H]1C[C@@H]2C)C)(C)CC4
Mifepristone	O=C5\C=C4/C(=C3/[C@@H](c1ccc(N(C)C)cc1)C[C@]2([C@@H](CC[C@]2(C#CC)O)[C@@H]3CC4)C)CC5
Nicardipine	O=C(OCCN(Cc1ccccc1)C)\C2=C(\N/C(=C(/C(=O)OC)C2c3cccc([N+]([O-])=O)c3)C)C
Nitrendipine	O=C(OCC)\C1=C(\N/C(=C(/C(=O)OC)C1c2cccc([N+]([O-])=O)c2)C)C
Simvastatin	O=C(O[C@@H]1[C@H]3C(=C/[C@H](C)C1)\C=C/[C@@H]([C@@H]3CC[C@H]2OC(=O)C[C@H](O)C2)C)C(C)(C)CC
Tipranavir	CCC[C@]1(CC(/O)=C(\C(=O)O1)[C@H](CC)c3cccc(NS(=O)(=O)c2ccc(cn2)C(F)(F)F)c3)CCc4ccccc4
Verapamil	N#CC(c1cc(OC)c(OC)cc1)(CCCN(CCc2ccc(OC)c(OC)c2)C)C(C)C
Diltiazem	O=C2N(c3c(S[C@@H](c1ccc(OC)cc1)[C@H]2OC(=O)C)cccc3)CCN(C)C
Taurolithocholic acid	C[C@H](CCC(=O)NCCS(=O)(=O)O)[C@H]1CC[C@@H]2[C@@]1(CC[C@H]3[C@H]2CC[C@H]4[C@@]3(CC[C@H](C4)O)C)C
Haloperidol	c1cc(ccc1C(=O)CCCN2CCC(CC2)(c3ccc(cc3)Cl)O)F

Maprotiline	c1ccc3c(c1)C4c2ccccc2C3(CC4)CCCNC
Noscapine	O=C2OC(c1ccc(OC)c(OC)c12)C5N(C)CCc4c5c(OC)c3OCOc3c4
Prednisone	O=C(CO)[C@@]3(O)CC[C@H]2[C@@H]4CC\C1=C\C(=O)\C=C/[C@]1(C)[C@H]4C(=O)C[C@@]23C
Procyclidine	OC(c1ccccc1)(CCN2CCCC2)C3CCCCC3
Propafenone	O=C(c1ccccc1OCC(O)CNCCC)CCc2ccccc2
Quinidine	O(c4cc1c(nccc1[C@H](O)[C@@H]2N3CC[C@@H](C2)[C@@H](/C=C)C3)cc4)C
Quinine	O(c4cc1c(nccc1[C@@H](O)[C@H]2N3CC[C@@H](C2)[C@@H](/C=C)C3)cc4)C
Taurocholate	C[C@H](CCC(=O)NCCS(=O)(=O)O)[C@H]1CC[C@@H]2[C@@]1([C@H](C[C@H]3[C@H]2[C@@H](C[C@H]4[C@@]3(CC[C@H](C4)O)C)O)O)C
Tetracycline	CN(C)[C@@H]2C(\O)=C(\C(N)=O)C(=O)[C@@]3(O)C(/O)=C4/C(=O)c1c(cccc1O)[C@@](C)(O)C4CC23
Vinblastine	O=C(OC)[C@]4(c2c(c1ccccc1n2)CCN3C[C@](O)(CC)C[C@H](C3)C4)c5c(OC)cc6c(c5)[C@@]89[C@@H](N6C)[C@@](O)(C(=O)OC)[C@H](OC(=O)C)[C@@]7(/C=C\CN([C@@H]78)CC9)CC
Amodiaquine	Clc1cc2nccc(c2cc1)Nc3cc(c(O)cc3)CN(CC)CC
Fumitremorgin C	O=C4N5[C@H](C(=O)N3[C@H](c2c(c1ccc(OC)cc1n2)C[C@H]34)\C=C(/C)C)CCC5
Hoechst 33342	CCOc1ccc(cc1)c2[nH]c3cc(ccc3n2)c4[nH]c5cc(ccc5n4)N6CCN(CC6)C
Mitoxantrone	O=C2c1c(c(NCCNCCO)ccc1NCCNCCO)C(=O)c3c2c(O)ccc3O
Naringenin	O=C2c3c(O[C@H](c1ccc(O)cc1)C2)cc(O)cc3O
Omeprazole	O=S(c2nc1ccc(OC)cc1n2)Cc3ncc(c(OC)c3C)C
Prazosin	O=C(N3CCN(c2nc1cc(OC)c(OC)cc1c(n2)N)CC3)c4occc4
Progesterone	O=C4\C=C2/[C@]([C@H]1CC[C@@]3([C@@H](C(=O)C)CC[C@H]3[C@@H]1CC2)C)(C)CC4
Bromosulfalein	c1cc(c(cc1/C(=C/2\C=CC(=O)C(=C2)S(=O)(=O)[O-])/c3c(c(c(c(c3Br)Br)Br)Br)C(=O)O)S(=O)(=O)[O-])O
Lansoprazole	FC(F)(F)COc1c(c(ncc1)CS(=O)c3nc2ccccc2n3)C
P-aminohippuric acid	O=C(c1ccc(N)cc1)NCC(=O)O
Rifampicin	CN1CCN(CC1)/N=C/c2c(O)c3c5C(=O)C4(C)O/C=C/C(OC)C(C)C(C(C)C(O)C(C)C(O)C(C)\C=C\C=C(\C)C(=O)Nc2c(O)c3c(O)c(C)c5O4)C(=O)OC
1-methyl-4-phenylpyridinium	c2cc(c1cc[n+](cc1)C)ccc2

4-Methylumbelliferone glucuronide	O=C/2Oc1cc(O)ccc1\C(=C\2)C
Amantadine	C1C2CC3CC1CC(C2)(C3)N
Amiloride	Clc1nc(C(=O)\N=C(/N)N)c(nc1N)N
Amitriptyline	c3cc2c(/C(c1c(cccc1)CC2)=C\CCN(C)C)cc3
Antipyrine	O=C2\C=C(/N(N2c1ccccc1)C)C
Atropine	CN3[C@H]1CC[C@@H]3C[C@@H](C1)OC(=O)C(CO)c2ccccc2
Budesonide	O=C\1\C=C/[C@]2(/C(=C/1)CC[C@H]3[C@H]4[C@](C[C@H](O)[C@H]23)([C@@]5(OC(O[C@@H]5C4)CCC)C(=O)CO)C)C
Captopril	O=C(O)[C@H]1N(C(=O)[C@H](C)CS)CCC1
Carbamazepine	c1ccc2c(c1)C=Cc3ccccc3N2C(=O)N
Carnitine	[O-]C(=O)C[C@@H](O)C[N+](C)(C)C
Cefamandole	O=C2N1/C(=C(\CS[C@@H]1[C@@H]2NC(=O)[C@H](O)c3ccccc3)CSc4nnnn4C)C(=O)O
Chloroquine	Clc1cc2nccc(c2cc1)NC(C)CCCN(CC)CC
Chlorzoxazone	Clc2cc1c(OC(=O)N1)cc2
Cholic acid	C[C@H](CCC(=O)O)[C@H]1CC[C@@H]2[C@@]1([C@H](C[C@H]3[C@H]2[C@@H](C[C@H]4[C@@]3(CC[C@H](C4)O)C)O)O)C
Cimetidine	N#CN\C(=N/C)NCCSCc1ncnc1C
Colchicine	O=C(N[C@@H]3C\1=C\C(=O)C(\OC)=C/C=C/1c2c(cc(OC)c(OC)c2OC)CC3)C
Dehydroisoandrosterone-3-sulfate	O=S(=O)(O)O[C@@H]4C/C3=C/C[C@@H]2[C@H](CC[C@@]1(C(=O)CC[C@H]12)C)[C@@]3(C)CC4
Desipramine	c1cc3c(cc1)CCc2c(cccc2)N3CCCNC
Digoxin	O=C\1OC/C(=C/1)[C@H]2CC[C@@]8(O)[C@]2(C)[C@H](O)C[C@H]7[C@H]8CC[C@H]6[C@]7(C)CC[C@H](O[C@@H]5O[C@H](C)[C@@H](O[C@@H]4O[C@@H]([C@@H](O[C@@H]3O[C@@H]([C@@H](O)[C@@H](O)C3)C)[C@@H](O)C4)C)[C@@H](O)C5)C6
Doxorubicin	C[C@H]1[C@H]([C@H](C[C@@H](O1)O[C@H]2C[C@@](Cc3c2c(c4c(c3O)C(=O)c5cccc(c5C4=O)OC)O)(C(=O)CO)O)N)O
Erythromycin	CC[C@@H]1[C@@]([C@@H]([C@H](C(=O)[C@@H](C[C@@]([C@@H]([C@H]([C@@H]([C@H](C(=O)O1)C)O[C@H]2C[C@@]([C@H](O2)C)O)(C)OC)C)O[C@H]3[C@@H]([C@H](C[C@H](O3)C)N(C)C)O)(C)O)C)O)(C)O
Estradiol-17ß-glucuronide	O=C(O)[C@@H]5OC(O[C@H]4CC[C@@H]2[C@]4(C)CC[C@@H]1c3ccc(O)cc3CC[C@H]12)[C@@H](O)[C@H](O)[C@H]5O

Etoposide	C[C@@H]1OC[C@@H]2[C@@H](O1)[C@@H]([C@H]([C@@H](O2)O[C@@H]3c4cc5c(cc4[C@H]([C@@H]6[C@@H]3COC6=O)c7cc(c(c(c7)OC)O)OC)OCO5)O)O
Fexofenadine	O=C(O)C(c1ccc(cc1)C(O)CCCN2CCC(CC2)C(O)(c3ccccc3)c4ccccc4)(C)C
Flucloxacillin	O=C(O)[C@@H]3N4C(=O)[C@@H](NC(=O)c2c(onc2c1c(F)cccc1Cl)C)[C@H]4SC3(C)C
Hydrochlorothiazide	O=S(=O)(c1c(Cl)cc2c(c1)S(=O)(=O)NCN2)N
Hydrocortisone	O=C4\C=C2/[C@]([C@H]1[C@@H](O)C[C@@]3([C@@](O)(C(=O)CO)CC[C@H]3[C@@H]1CC2)C)(C)CC4
Indinavir	CC(C)(C)NC(=O)[C@@H]1CN(CCN1C[C@H](C[C@@H](Cc2ccccc2)C(=O)N[C@H]3c4ccccc4C[C@H]3O)O)Cc5cccnc5
Indomethacin	Cc1c(c2cc(ccc2n1C(=O)c3ccc(cc3)Cl)OC)CC(=O)O
Mesalazine	O=C(O)c1cc(ccc1O)N
Methotrexate	O=C(O)[C@@H](NC(=O)c1ccc(cc1)N(C)Cc2nc3c(nc2)nc(nc3N)N)CCC(=O)O
Metoprolol	O(c1ccc(cc1)CCOC)CC(O)CNC(C)C
Nevirapine	O=C2Nc1c(ccnc1N(c3ncccc23)C4CC4)C
Nicotine	n1cc(ccc1)[C@H]2N(C)CCC2
Ofloxacin	Fc4cc1c2N(/C=C(\C1=O)C(=O)O)C(COc2c4N3CCN(C)CC3)C
Phenobarbital	CCC1(C(=NC(=O)N=C1O)O)c2ccccc2
Phenylethyl isothiocyanate	S=C=N/CCc1ccccc1
Phenytoin	O=C2NC(=O)NC2(c1ccccc1)c3ccccc3
Pravastatin	O=C(O)C[C@H](O)C[C@H](O)CC[C@H]2[C@H](/C=C\C1=C\[C@@H](O)C[C@H](OC(=O)[C@@H](C)CC)[C@@H]12)C
Prednisolone	O=C\1\C=C/[C@]4(/C(=C/1)CC[C@@H]2[C@@H]4[C@@H](O)C[C@@]3([C@@](O)(C(=O)CO)CC[C@@H]23)C)C
Probenecid	O=S(=O)(N(CCC)CCC)c1ccc(C(=O)O)cc1
Propranolol	CC(C)NCC(COc1cccc2c1cccc2)O
Ranitidine	[O-][N+](=O)\C=C(\NC)NCCSCc1oc(cc1)CN(C)C
Sotalol	O=S(=O)(Nc1ccc(cc1)C(O)CNC(C)C)C
Sparfloxacin	C[C@@H]1CN(C[C@@H](N1)C)c2c(c(c3c(c2F)n(cc(c3=O)C(=O)O)C4CC4)N)F
Sulfasalazine	O=S(=O)(Nc1ccccn1)c3ccc(/N=N/c2cc(C(O)=O)c(O)cc2)cc3

Sulfinpyrazone	O=C2N(c1ccccc1)N(C(=O)C2CCS(=O)c3ccccc3)c4ccccc4
Sulindac	O=S(c1ccc(cc1)\C=C3/c2ccc(F)cc2\C(=C3C)CC(=O)O)C
Testosterone	O=C4\C=C2/[C@]([C@H]1CC[C@@]3([C@@H](O)CC[C@H]3[C@@H]1CC2)C)(C)CC4
Tinidazole	[O-][N+](=O)c1cnc(n1CCS(=O)(=O)CC)C
Trimethoprim	COc1cc(cc(c1OC)OC)Cc2cnc(nc2N)N
Valproic acid	O=C(O)C(CCC)CCC
Warfarin	CC(=O)CC(C\1=C(/O)c2ccccc2OC/1=O)c3ccccc3
Vincristine	O=C(OC)[C@]4(c2c(c1ccccc1n2)CCN3C[C@](O)(CC)C[C@@H](C3)C4)c5c(OC)cc6c(c5)[C@@]89[C@@H](N6C=O)[C@@](O)(C(=O)OC)[C@H](OC(=O)C)[C@@]7(/C=C\CN([C@@H]78)CC9)CC
Zidovudine	O=C/1NC(=O)N(\C=C\1C)[C@@H]2O[C@@H]([C@@H](\N=[N+]=[N-])C2)CO

Appendix 2: Highest docking score for each compound

Compound	ABCB1P-gp	S-Ver3g5u	S/QZ59upper	S/QZ59lower	S/QZ59rrr	S/3G61
Chlorprotixene	1	-9.3004074	-8.459938	-8.1764832	-8.803051	-9.436208
Cyclosporine-A	1	-10.156612	-13.913272	-15.671002	-16.93857	-17.60867
Diethylstilbestrol	1	-9.4258194	-7.717423	-7.9587054	-10.264479	-9.587139
Dipyridamole	1	-11.091798	-10.233459	-10.369469	-11.163453	-12.34627
Flupentixol	1	-10.632503	-10.661702	-11.41256	-10.590546	-11.23787
GF120918	1	-9.8905039	-10.713839	-9.8388119	-10.92973	-11.64353
Isradipine	1	-10.450793	-11.525162	-10.018229	-9.8677025	-10.29302
Ivermectin A	1	-15.407999	-17.49052	-15.053252	-14.21278	-13.68392
Ivermectin B	1	-8.5880575	-14.182034	-15.702105	-12.753691	-11.51765
Loperamide	1	-9.830308	-9.6608973	-10.113488	-10.485672	-12.6576
Lopinavir	1	-10.369528	-9.2258024	-10.088408	-10.676792	-12.42869
MK571	1	-10.548156	-12.268476	-9.7600927	-9.9758444	-12.00426
Quercetin	1	-9.6049185	-9.846427	-9.4942179	-10.662757	-9.802873
Reserpine	1	-9.6042442	-11.332067	-10.40928	-11.331938	-14.4181
Ritonavir	1	-11.058732	-12.277213	-9.7571392	-10.353203	-12.82965
Saquinavir	1	-11.114758	-10.143492	-10.06812	-9.9175653	-13.85331
Silymarin	1	-12.211308	-12.740396	-11.329776	-12.592422	-15.67663
Tamoxifen	1	-9.6038904	-8.5444183	-9.0636072	-10.022164	-11.78149
Terfenadine A44	1	-10.828145	-8.8642521	-10.5563	-11.039285	-11.73056

Thioridazine	1	-8.9083242	-7.8792377	-10.050879	-9.5215521	-10.70159
Benzbromarone	0	-10.346081	-8.221384	-9.9977951	-10.500174	-11.18317
Amiodarone	1	-10.70286	-10.739605	-9.6271276	-10.519768	-12.65702
Apigenin	1	-10.61247	-9.2287788	-8.6095076	-10.622262	-9.70258
17β-estradiol	1	-9.3679514	-7.8578453	-9.462574	-10.319477	-8.940121
Biochanin A	1	-9.2910938	-10.263983	-8.1855145	-8.2347078	-9.657901
Chlorpromazine	1	-8.3345718	-7.7475543	-7.8980999	-8.9394903	-10.38405
Chrysin	1	-10.554909	-8.3061218	-10.490614	-10.57549	-8.911892
Ergocristine	1	-10.259492	-10.541076	-10.302964	-11.130157	-12.58193
Felodipine	1	-9.6793785	-7.7453742	-9.5227003	-9.5240564	-10.0143
Gefitinib	1	-10.119772	-10.34021	-9.0785007	-11.483757	-11.43558
Genistein	1	-9.0585032	-8.3349791	-10.125244	-9.279067	-9.529384
Glibenclamide	1	-10.019725	-11.332147	-8.7031746	-10.400661	-11.69834
Imatinib	1	-10.306933	-10.95615	-10.700336	-10.292026	-10.45726
Ketoconazole	1	-10.214561	-11.057756	-9.1053286	-9.9653654	-11.57038
Kol 43	1	-8.9050388	-10.648838	-9.1451588	-10.571228	-12.40026
Medroxyprogesterone	1	-9.6898108	-7.7250443	-9.3627224	-9.9533339	-11.0971
Mifepristone	1	-10.505825	-9.1948147	-10.158919	-9.4235382	-13.46323
Nicardipine	1	-10.663759	-9.9746799	-9.4923162	-10.301338	-13.60907
Nitrendipine	1	-9.9102268	-8.5903606	-9.1928492	-9.6273088	-10.03407
Simvastatin	1	-10.496209	-8.8446598	-9.1530542	-9.7082663	-11.98095
Tipranavir	1	-11.627663	-10.544059	-10.162716	-11.806179	-14.14789
Verapamil	1	-10.519194	-10.790701	-9.2990332	-8.6394806	-12.54895
Diltiazem	1	-9.7515516	-9.5250664	-7.7090964	-9.7192259	-12.28151
Taurolithocholic acid	1	-10.897123	-9.1023989	-10.131459	-11.243021	-12.83619

Haloperidol	1	-9.2387438	-10.970337	-8.5671577	-9.6470385	-11.13091
Maprotiline	1	-9.5215988	-8.9605398	-8.6476765	-9.4146843	-9.887517
Noscapine	1	-10.431744	-10.722205	-10.501363	-10.413963	-12.00903
Prednisone	1	-10.309956	-10.344711	-9.5839205	-10.487357	-10.65359
Procyclidine	1	-8.8565769	-8.4229498	-8.4684896	-8.7602282	-9.858139
Propafenone	1	-9.1153469	-9.4176445	-8.7731705	-10.269821	-11.47182
Quinidine	1	-9.539341	-7.6209884	-9.5094252	-9.761198	-9.330069
Quinine A50	1	-9.3600197	-8.193058	-8.6307859	-10.090897	-11.55817
Taurocholate	1	-10.252878	-9.6550837	-10.120589	-13.826485	-15.2546
Tetracycline	1	-12.25106	-9.2137432	-9.9086618	-12.249457	-11.59131
Vinblastine	1	-7.6626611	-9.919591	-11.387159	-9.285367	-13.51052
Amodiaquine	0	-10.609269	-9.0921354	-8.8429432	-10.655272	-10.72421
Fumitremorgin C	0	-10.10892	-9.4868574	-9.3236475	-10.058537	-10.67109
Hoechst 33342	0	-9.6359091	-11.045726	-9.3159275	-10.280766	-11.00911
Mitoxantrone	0	-12.069983	-11.434206	-10.387532	-10.386574	-13.87651
Naringenin	0	-10.699261	-9.1101408	-10.72489	-10.711188	-8.341228
Omeprazole	0	-9.6031446	-10.293558	-10.02894	-8.8998051	-9.922818
Prazosin	0	-8.8828516	-10.613285	-9.3761845	-9.1080608	-11.63355
Progesterone	0	-9.2005606	-8.8879795	-8.9872503	-10.076205	-10.39482
Bromosulfalein	0	-12.974476	-12.310234	-11.747195	-12.696462	-15.77145
Lansoprazole	0	-9.9817438	-9.4636879	-8.6808853	-9.5217371	-10.55191
P-aminohippuric acid	0	-7.5560284	-7.3900452	-7.3223267	-7.7994919	-8.224644
Rifampicin	0	-10.881661	-14.776971	-12.707092	-15.607213	-15.85894
1-methyl-4-phenylpyridinium	0	-6.934773	-7.7081928	-6.4839993	-6.8460197	-7.963677
4-Methylumbelliferoneglucuronide	0	-7.9916229	-7.8331814	-8.9206944	-7.9753709	-6.826306

Amantadine	0	-6.3526926	-5.3320203	-6.3266449	-6.3345504	-7.503057
Amiloride	0	-8.8389435	-7.5056643	-7.6710935	-8.2251282	-8.684454
Amitriptyline A98	0	-9.2818766	-8.9655581	-9.3389559	-9.2117443	-9.565305
Antipyrine	0	-7.3381066	-6.6651206	-7.3842626	-7.571815	-7.122243
Atropine	0	-9.0077009	-8.051362	-8.3661737	-8.9017839	-9.776143
Budesonide	0	-11.293019	-11.60719	-9.1883268	-9.7388487	-11.08708
Captopril	0	-7.6405721	-8.0399799	-8.0298433	-8.188179	-7.657503
Carbamazepine	0	-8.5274229	-7.2034917	-8.4886007	-9.3757973	-9.013274
Carnitine	0	-7.2166624	-7.0989256	-6.9449291	-6.4096179	-8.174127
Cefamandole	0	-10.145037	-10.903329	-10.494161	-12.38217	-12.55593
Chloroquine	0	-9.532608	-9.0205898	-8.1266356	-9.2494192	-9.54703
Chlorzoxazone	0	-7.3827214	-6.3034315	-6.025938	-7.3652225	-7.332795
Cholic acid	0	-10.776716	-8.9253407	-8.5485239	-10.711271	-12.42511
Cimetidine	0	-7.7259474	-7.2465606	-8.2314672	-8.4074793	-8.449764
Colchicine	0	-10.714324	-12.727611	-10.133555	-11.653122	-11.96499
Dehydroisoandrosterone-3-sulfate	0	-9.270402	-10.271486	-8.6834316	-9.9046402	-11.77644
Desipramine	0	-8.9099407	-9.0193939	-8.8133793	-8.9883986	-9.752567
Digoxin	0	-11.16558	-15.264123	-13.845214	-12.67726	-14.25247
Doxorubicin	0	-12.270447	-9.7047529	-12.654102	-12.111775	-14.87355
Erythromycin	0	-11.146794	-12.131126	-12.796752	-12.736913	-13.13491
Estradiol-17β-glucuronide A66	0	-11.490866	-10.448287	-10.062901	-11.392282	-14.00624
Etoposide	0	-11.064533	-11.451082	-10.737698	-12.774714	-15.36379
Fexofenadine	0	-10.728499	-12.205088	-9.9871464	-11.987885	-10.3285
Flucloxacillin	0	-11.07047	-9.6187134	-9.8800745	-10.685433	-11.20973
Hydrochlorothiazide	0	-8.4825487	-8.925005	-8.2919388	-8.507925	-9.363761

Hydrocortisone	0	-10.304188	-9.6608438	-9.8306179	-10.031824	-10.60378
Indinavir	0	-9.6872206	-11.956147	-9.8693228	-11.337888	-12.38338
Indomethacin	0	-9.4931955	-12.066354	-8.9237766	-9.7628975	-12.64803
Mesalazine	0	-6.9549108	-7.6400247	-8.1257553	-7.9664588	-7.541798
Methotrexate	0	-11.986519	-12.722812	-10.812212	-10.057438	-13.23796
Metoprolol	0	-7.9435692	-9.3319874	-7.7133036	-8.3539362	-10.60786
Nevirapine	0	-9.0224047	-8.4566927	-7.8702278	-8.9860678	-7.827874
Nicotine	0	-6.9170685	-7.0697546	-6.1248832	-6.8440375	-7.391009
Ofloxacin	0	-8.5635834	-10.142664	-8.5361586	-10.691373	-10.93162
Phenobarbital	0	-9.6933594	-7.4914031	-9.5400419	-8.8618202	-8.699678
Phenylethyl isothiocyanate	0	-6.7740951	-5.9904637	-6.1920362	-6.5109282	-7.488194
Phenytoin	0	-8.8452339	-7.7078223	-8.1711397	-8.8362265	-8.585069
Pravastatin	0	-10.683473	-10.081711	-8.7615652	-11.556952	-12.16104
Prednisolone	0	-10.106754	-9.0637264	-9.5830917	-10.937309	-10.8012
Probenecid	0	-8.9605207	-9.3834229	-8.4948893	-8.909256	-8.614807
Propranolol	0	-9.3896971	-9.2584066	-8.6361856	-8.5617714	-10.65507
Ranitidine	0	-8.7418537	-8.3862219	-8.4139299	-8.5178547	-9.984197
Sotalol	0	-8.5546455	-8.9202785	-9.2301331	-8.9183006	-9.847155
Sparfloxacin	0	-9.6356382	-10.27851	-9.5493898	-10.364615	-11.56888
Sulfasalazine	0	-10.049615	-10.035481	-9.5956984	-10.255516	-11.54249
Sulfinpyrazone	0	-10.12289	-9.5853405	-9.0000858	-10.790933	-11.81816
Sulindac	0	-9.1808004	-11.152394	-9.6613617	-9.4821892	-11.10014
Testosterone	0	-9.7777252	-7.3668432	-8.6111145	-9.8584127	-9.95339
Tinidazole	0	-8.1798	-8.2702465	-8.0159121	-8.2341499	-9.185381
Trimethoprim	0	-8.3030396	-10.153893	-9.2529078	-9.0906744	-9.998123

Valproic acid	0	-7.5599384	-6.473103	-6.8606076	-6.737505	-7.569929
Warfarin A37	0	-10.079456	-8.9441195	-9.6533222	-9.8448915	-10.05809
Vincristine	0	-9.6095648	-12.462447	-10.760735	-10.002497	-14.47043
Zidovudine	0	-9.2348881	-8.7530203	-8.329999	-10.314642	-8.797813

Appendix 3: LogP, LogD and MW of compounds

Substance	ABCB1P-gp	Group	S/3G61-QZ59RR	LogD(2)	LogD(5.5)	LogD(6.5)	LogD(7.4)	LogD(10)	LogP	MW
17ß-estradiol	1	Train	-8.94012	4.15	4.15	4.15	4.14	3.96	4.15	272.38
4-Methylumbelliferoneglucuronide	0	Train	-6.82631	2.43	2.43	2.42	2.33	0.46	2.43	176.17
Amantadine	0	Train	-7.50306	-0.66	-0.66	-0.63	-0.47	1.62	2.44	151.25
Amiloride	0	Train	-8.68445	0.76	1.22	1.22	1.22	1.22	1.22	229.63
Amiodarone	1	Train	-12.657	4.71	4.78	5.14	5.87	7.72	7.81	645.31
Amitriptyline	0	Train	-9.5653	1.31	1.41	1.87	2.64	4.35	4.41	277.4
Amodiaquine	0	Train	-10.7242	-0.96	0.0236	0.31	0.95	2.36	3.13	355.86
Atropine	0	Train	-9.77614	-1.72	-1.7	-1.57	-1.09	1.09	1.38	289.37
Benzbromarone	0	Train	-11.1832	6.65	5.75	4.82	4.05	3.5	6.65	424.08
Biochanin A	1	Train	-9.6579	3.34	3.28	2.93	2.11	-0.77	3.34	284.26
Bromosulfalein	0	Train	-15.7715	-3.67	-4.04	-4.04	-4.06	-4.89	1.46	794.03
Budesonide	0	Train	-11.0871	3.2	3.2	3.2	3.2	3.2	3.2	430.53
Captopril	0	Train	-7.6575	1.98	0.13	-0.81	-1.46	-2.14	1.99	217.29
Carbamazepine	0	Train	-9.01327	1.89	1.89	1.89	1.89	1.89	1.89	236.27
Cefamandole	0	Train	-12.5559	-0.17	-2.87	-3.55	-3.75	-3.8	-0.0443	462.5
Chloroquine	0	Train	-9.54703	0.31	0.65	1.16	1.59	3.81	4.41	319.87
Chlorpromazine	1	Train	-10.384	2.08	2.15	2.51	3.24	5.09	5.18	318.86
Chlorprotixene	1	Train	-9.43621	2.11	2.24	2.77	3.57	5.17	5.21	315.86

Chlorzoxazone	0	Train	-7.33279	1.82	1.82	1.81	1.78	0.39	1.82	169.57
Cholic acid f	0	Train	-12.4251	2.88	2.06	1.13	0.27	-0.85	2.88	408.57
Chrysin	1	Train	-8.91189	3.13	3.07	2.73	1.92	-0.98	3.13	254.24
Colchicine	0	Train	-11.965	1.07	1.07	1.07	1.07	1.07	1.07	399.44
Cyclosporine-A	1	Train	-17.6087	2.79	2.79	2.79	2.79	2.79	2.79	1202.61
Dehydroisoandrosterone-3-sulfate	0	Train	-11.7764	0.022	0.022	0.022	0.022	0.022	3.52	368.49
Desipramine	0	Train	-9.75257	0.72	0.88	0.95	1.27	3.48	3.97	266.38
Digoxin	0	Train	-14.2525	1.29	1.29	1.29	1.29	1.29	1.29	780.94
Dipyridamole	1	Train	-12.3463	0.51	2.61	3.19	3.33	3.35	3.35	504.63
Erythromycin	0	Train	-13.1349	-1.19	-0.54	0.35	1.16	1.9	1.91	733.93
Estradiol-17ß-glucuronide	0	Train	-14.0062	3.72	0.99	0.3	0.0959	-0.11	3.81	448.51
Etoposide	0	Train	-15.3638	0.28	0.28	0.27	0.27	-0.058	0.28	588.56
Felodipine	1	Train	-10.0143	3.98	4.76	4.76	4.76	4.76	4.76	384.25
Fexofenadine	0	Train	-10.3285	0.64	1.21	1.23	1.23	0.65	3.73	501.66
Flupentixol	1	Train	-11.2379	-0.21	1.93	2.85	3.44	3.67	3.67	434.52
Fumitremorgin C	0	Train	-10.6711	3.34	3.34	3.34	3.34	3.34	3.34	379.45
Gefitinib	1	Train	-11.4356	-1.4	0.97	2.07	2.56	2.7	2.7	446.9
Genistein	1	Train	-9.52938	3.11	3.06	2.73	1.93	-1.39	3.11	270.24
GF120918	1	Train	-11.6435	1.33	2.21	3.14	3.9	4.43	4.43	563.64
Glibenclamide	1	Train	-11.6983	3.23	1.96	1.4	1.26	1.23	3.23	487.96
Haloperidol	1	Train	-11.1309	0.66	1.23	2.11	2.93	3.75	3.76	375.86
Hydrochlorothiazide	0	Train	-9.36376	-0.0211	-0.0213	-0.023	-0.0362	-1.52	-0.0211	297.74
Hydrocortisone	0	Train	-10.6038	1.76	1.76	1.76	1.76	1.76	1.76	362.46
Indinavir	0	Train	-12.3834	-0.65	3.06	3.38	3.43	3.43	3.44	613.79
Indomethacin	0	Train	-12.648	4.25	2.7	1.74	0.98	0.5	4.25	357.79

Isradipine	1	Train	-10.293	3.08	3.73	3.73	3.73	3.73	3.73	371.39
Ivermectin B	1	Train	-11.5176	5.18	5.18	5.18	5.18	5.17	5.18	861.07
Ketoconazole	1	Train	-11.5704	0.59	2.66	3.51	3.93	4.04	4.04	531.43
Kol 43	1	Train	-12.4003	4.75	4.75	4.75	4.75	4.75	4.75	469.57
Lansoprazole	0	Train	-10.5519	0.41	2.57	2.58	2.58	2.27	2.58	369.36
Loperamide	1	Train	-12.6576	1.05	1.92	2.85	3.61	4.14	4.15	477.04
Lopinavir	1	Train	-12.4287	5.41	5.42	5.42	5.42	5.42	5.42	628.8
Maprotiline	1	Train	-9.88752	1.26	1.27	1.3	1.51	3.65	4.36	277.4
Medroxyprogesterone	1	Train	-11.0971	3.58	3.58	3.58	3.58	3.58	3.58	344.49
Methotrexate	0	Train	-13.238	-2.95	-3.79	-4.71	-5.1	-5.19	-0.45	454.44
Metoprolol	0	Train	-10.6079	-1.47	-1.42	-1.14	-0.47	1.5	1.63	267.36
Mifepristone	1	Train	-13.4632	3.74	5.91	6.15	6.19	6.19	6.19	429.59
Mitoxantrone	0	Train	-13.8765	-3.14	-2.52	-2.26	-1.58	-2.1	1.55	444.48
MK571	1	Train	-12.0043	3.01	2.16	1.19	0.37	-0.34	3.41	515.09
Naringenin	0	Train	-8.34123	2.63	2.62	2.56	2.23	-1.5	2.63	272.25
Nevirapine	0	Train	-7.82787	0.57	2.61	2.64	2.64	2.64	2.64	266.3
Nicotine	0	Train	-7.39101	-3.47	-2.22	-1.46	-0.62	0.55	0.57	162.23
Nitrendipine	1	Train	-10.0341	2.98	3.81	3.81	3.81	3.81	3.81	360.36
Noscapine	1	Train	-12.009	-0.69	1.51	2.17	2.35	2.38	2.38	413.42
Omeprazole	0	Train	-9.92282	-0.5	2.29	2.35	2.35	1.51	2.36	345.42
P-aminohippuric acid	0	Train	-8.22464	-0.71	-2.26	-3.17	-3.69	-3.87	-0.12	194.19
Phenobarbital	0	Train	-8.69968	0.5	-1.12	-2.58	-3.75	-3.99	0.51	232.24
Phenylethyl isothiocyanate	0	Train	-7.48819	3.47	3.47	3.47	3.47	3.47	3.47	163.24
Phenytoin	0	Train	-8.58507	1.42	1.42	1.42	1.38	-0.0121	1.42	252.27
Pravastatin	0	Train	-12.161	2.21	0.9	-0.0679	-0.88	-1.54	2.21	424.53

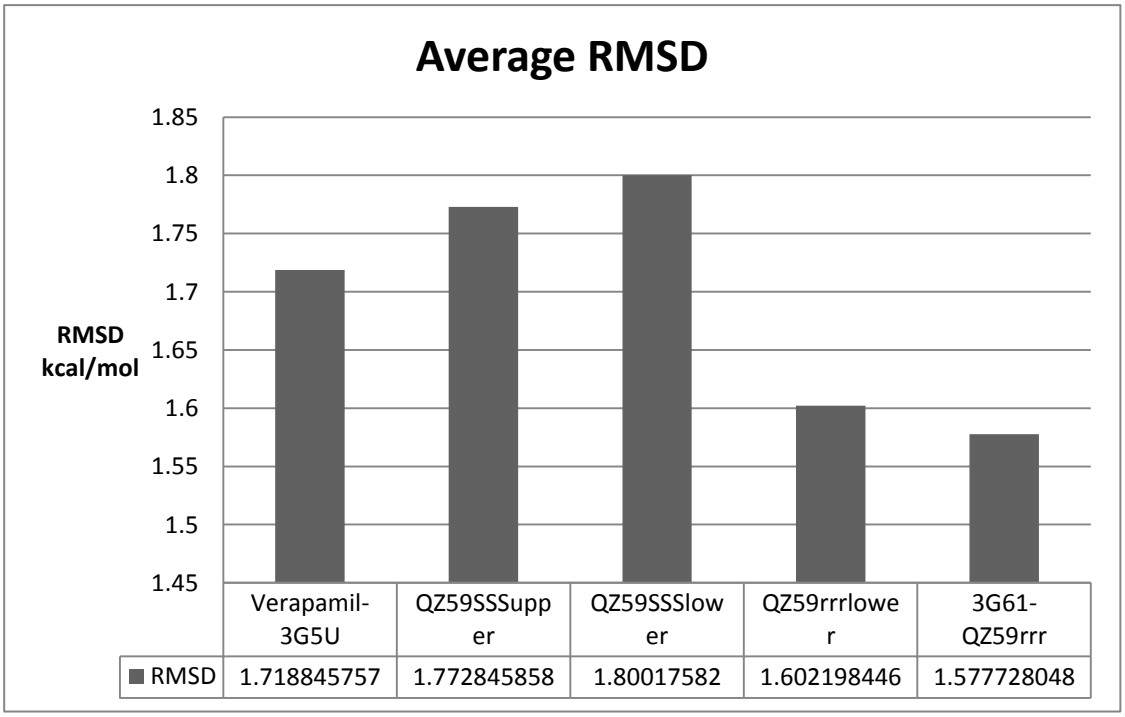
Prazosin	0	Train	-11.6336	-0.36	1.09	1.83	2.08	2.14	2.14	383.4
Probenecid	0	Train	-8.61481	2.5	0.72	-0.13	-0.53	-0.64	2.51	285.36
Procyclidine	1	Train	-9.85814	0.76	0.77	0.82	1.08	3.27	3.86	287.44
Propafenone	1	Train	-11.4718	0.25	0.31	0.66	1.37	3.25	3.35	341.44
Propranolol	0	Train	-10.6551	-0.2	-0.15	0.12	0.79	2.77	2.9	259.34
Quinidine	1	Train	-9.33007	-1.27	-0.26	0.22	0.98	2.75	2.82	324.42
Quinine	1	Train	-11.5582	-1.27	-0.26	0.22	0.98	2.75	2.82	324.42
Ranitidine	0	Train	-9.9842	-3.63	-2.73	-1.91	-1.07	-0.078	-0.0681	314.4
Reserpine	1	Train	-14.4181	1.35	2.71	3.63	4.21	4.45	4.45	608.68
Rifampicin	0	Train	-15.8589	-0.89	0.0531	0.12	-0.12	-1.79	2.39	822.94
Ritonavir	1	Train	-12.8296	1.41	2.33	2.33	2.33	2.32	2.33	720.94
Saquinavir	1	Train	-13.8533	1.89	4.22	4.87	5.05	5.04	5.08	670.84
Silymarin	1	Train	-15.6766	4.23	4.22	4.13	3.73	-0.0292	4.23	482.44
Simvastatin	1	Train	-11.981	4.72	4.72	4.72	4.72	4.72	4.72	418.57
Sotalol	0	Train	-9.84715	-2.86	-2.78	-2.39	-1.68	-1.25	0.24	272.36
Sulfasalazine	0	Train	-11.5425	2.74	0.0689	-0.079	-0.0992	-0.63	3.05	398.39
Sulfinpyrazone	0	Train	-11.8182	1.82	-0.42	-0.59	-0.61	-0.61	1.89	404.48
Sulindac	0	Train	-11.1001	2.55	1.29	0.32	-0.49	-1.19	2.55	356.41
Taurolithocholic acid	1	Train	-12.8362	3.19	0.58	0.52	0.51	0.51	4.01	483.7
Terfenadine	1	Train	-11.7306	2.52	2.57	2.9	3.6	5.51	5.62	471.67
Testosterone	0	Train	-9.95339	3.18	3.18	3.18	3.18	3.18	3.18	288.42
Tetracycline	1	Train	-11.5913	-2.48	-1.91	-1.95	-2.22	-3.98	0.62	444.43
Thioridazine	1	Train	-10.7016	2.8	2.82	2.98	3.49	5.65	5.9	370.57
Tipranavir	1	Train	-14.1479	6.91	5.74	4.28	2.9	2.42	6.92	602.66
Trimethoprim	0	Train	-9.99812	-1.9	-0.8	0.0441	0.47	0.59	0.59	290.32

Valproic acid	0	Train	-7.56993	2.58	1.81	0.89	0.0231	-1.16	2.58	144.21
Verapamil	1	Train	-12.5489	0.92	1.08	1.64	2.46	3.99	4.02	454.6
Vinblastine	1	Train	-13.5105	1.76	3.47	4.65	5.43	5.91	5.92	810.97
Vincristine	0	Train	-14.4704	1.65	3.15	4.39	5.2	5.72	5.75	824.96
Warfarin f	0	Train	-10.0581	3.13	2.09	1.14	0.33	-0.37	3.13	308.33
1-methyl-4-phenylpyridinium	0	Validation	-7.96368	-0.29	-0.29	-0.29	-0.29	-0.29	-0.29	170.23
Antipyrine	0	Validation	-7.12224	0.42	0.44	0.44	0.44	0.44	0.44	188.23
Apigenin	1	Validation	-9.70258	2.13	2.07	1.73	0.93	-2.11	2.13	270.24
Carnitine	0	Validation	-8.17413	-4.7	-4.13	-4.13	-4.13	-4.13	-4.73	162.21
Cimetidine	0	Validation	-8.44976	-2.86	-1.91	-1.05	-0.39	-0.0664	- 0.0651	252.34
Diethylstilbestrol	1	Validation	-9.58714	5.33	5.33	5.33	5.33	5.09	5.33	268.35
Diltiazem	1	Validation	-12.2815	1.63	1.85	2.52	3.36	4.7	4.73	414.52
Doxorubicin	0	Validation	-14.8735	-2.86	-2.56	-1.91	-1.47	-3.57	0.24	543.52
Ergocristine	1	Validation	-12.5819	5.14	6.79	7.66	8.1	7.68	8.24	609.71
Flucloxacillin	0	Validation	-11.2097	2.73	-0.16	-0.71	-0.84	-0.87	2.89	453.87
Hoechst 33342	0	Validation	-11.0091	-2.06	0.0772	1.55	2.44	2.76	2.96	452.55
Imatinib	1	Validation	-10.4573	-1.59	0.84	1.77	2.49	2.89	2.89	493.6
Ivermectin A	1	Validation	-13.6839	5.69	5.69	5.69	5.69	5.68	5.69	875.09
Mesalazine	0	Validation	-7.5418	-1.66	-1.88	-2.26	-2.39	-2.41	0.74	153.14
Nicardipine	1	Validation	-13.6091	1.23	3.1	4.03	4.64	4.89	4.89	479.52
Ofloxacin	0	Validation	-10.9316	-1.24	-0.16	-0.0777	-0.39	-1.84	1.86	361.37
Prednisolone f	0	Validation	-10.8012	1.63	1.63	1.63	1.63	1.63	1.63	360.44
Prednisone f	1	Validation	-10.6536	1.57	1.57	1.57	1.57	1.56	1.57	358.43
Progesterone	0	Validation	-10.3948	3.83	3.83	3.83	3.83	3.83	3.83	314.46
Quercetin	1	Validation	-9.80287	1.99	1.91	1.51	0.62	-3.3	1.99	302.24

Sparfloxacin	0	Validation	-11.5689	-0.5	-0.00256	0.62	0.83	-0.54	2.6	392.4
Tamoxifen f	1	Validation	-11.7815	2.03	2.29	2.99	3.83	5.11	5.13	371.51
Taurocholate	1	Validation	-15.2546	-0.57	-3.18	-3.24	-3.25	-3.25	0.25	515.7
Tinidazole	0	Validation	-9.18538	-0.76	-0.29	-0.29	-0.29	-0.29	-0.29	247.27
Zidovudine	0	Validation	-8.79781	0.0522	0.0522	0.0518	0.0492	-0.52	0.0522	267.24

Appendix 4: Average Root Mean Square Deviation (RMSD

RMSD – The root mean square deviation of the pose, in Å, from the original ligand.



Appendix 5: Predicted vs Observed class for validation set of SVM 3 Model

Compound	Observed	Predicted	Classification confusion matrix
1-methyl-4-phenylpyridinium	0	0	TN
Antipyrine	0	0	TN
Apigenin	1	0	FN
Carnitine	0	0	TN
Cimetidine	0	0	TN
Diethylstilbestrol	1	1	TP
Diltiazem	1	1	TP
Doxorubicin	0	0	TN
Ergocristine	1	1	TP
Flucloxacillin	0	0	TN
Hoechst 33342	0	0	TN
Imatinib	1	0	FN
Ivermectin A	1	1	TP
Mesalazine	0	0	TN
Nicardipine	1	1	TP
Ofloxacin	0	0	TN
Prednisolone	0	0	TN
Prednisone	1	0	FN
Progesterone	0	1	FP
Quercetin	1	0	FN
Sparfloxacin	0	0	TN
Tamoxifen	1	1	TP
Taurocholate	1	0	FN
Tinidazole	0	0	TN
Zidovudine	0	0	TN

TN – True Negative

FN – False Negative

TP – True Positive

FP - False Positive

1 – Non-substrate

0 - Substrate