



universität  
wien

# MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Data-based and structure-based analysis  
of bile salt export pump inhibition“

verfasst von / submitted by

Anna Magdalena Ambros BSc

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

Master of Science (MSc)

Wien, 2021

Studienkennzahl lt. Studienblatt /  
degree programme code as it appears on  
the student record sheet:

UA 066 606

Studienrichtung lt. Studienblatt /  
degree programme as it appears on  
the student record sheet:

Masterstudium Drug Discovery  
and Development

Betreut von / Supervisor:

Univ.-Prof. Mag. Dr. Gerhard Ecker



## Acknowledgements

Throughout the whole time I worked on my master's thesis, I received a great deal of advice, support and guidance. I genuinely enjoyed working in the Pharmacoinformatics Research Group, implementing knowledge I obtained during my master, and gaining a great deal of expertise and hands-on *in silico* experience. I could not have achieved everything I have if it were not for my family, friends, supervisor and research fellows.

First, I would like to thank Prof. Dr. Gerhard Ecker for his outstanding supervision and his scientific and personal guidance. I admire his approach of mentoring students and all the time and heart he puts into his research group. I could not have hoped for any better supervisor and will always be grateful for his support.

I would like to thank the whole Pharmacoinformatics Research Group for their advice, their support and the good working atmosphere. Due to the current circumstances, I did not get the chance to meet everyone in person, however I am thankful for all the outstanding scientists and kind humans I got the chance to work with. I would like to give my sincere thanks to Dr. Claire Colas, Dr. Melanie Grandits and Karin Grillberger BSc for giving me input on scientific issues, answering my questions and brightening up the ordinary workday.

My best and warmest thanks go to my family and friends. I would like to express my sincere gratitude to my parents, who were always there for me and encouraged me to keep going during difficult times. I would have never come this far without them and I will never forget their endless support. I would also like to thank my siblings, my grandparents, my former and current fellow students, my best friends and all those, who accompanied me on my way.

Lastly, I would like to thank all the professors, lecturers and organizers of the Drug Discovery and Development Master for providing such interesting and diverse lectures and training opportunities and for guiding students on their way of becoming prospective scientists.



## Table of Contents

1.	Introduction.....	1
1.1.	ABC transporter family.....	1
1.2.	Enterohepatic circulation of bile acids.....	3
1.2.1.	Bile acids – structure, synthesis and physiological function .....	3
1.2.2.	The role of transporters in enterohepatic circulation.....	4
1.3.	Bile salt export pump .....	5
1.4.	Drug-induced liver injury.....	6
2.	Aim of the thesis .....	7
3.	Material and Methods .....	8
3.1.	Data-based approach .....	8
3.1.1.	KNIME Analytics Platform .....	8
3.1.1.1.	Data collection.....	8
3.1.1.2.	Physicochemical property characterization .....	9
3.1.1.3.	Matched-Molecular-Pairs (MMPs) Workflow.....	9
3.1.1.4.	TGD-, GpiDAPH3- and MACCS-Fingerprint Clustering Workflow... ..	10
3.2.	Structure-based approach.....	11
3.2.1.	Cryo-EM structure BSEP.....	11
3.2.2.	Protein and ligand preparation .....	12
3.2.3.	Binding site detection.....	12
3.2.3.1.	Sequence alignment of P-glycoprotein and BSEP.....	12
3.2.3.2.	Binding site calculation .....	13
3.2.4.	Structure-based 3D-Pharmacophore creation and screening .....	13
3.2.5.	Docking studies.....	14
3.2.5.1.	Glide Docking.....	14
3.2.5.2.	Induced-Fit Docking (IFD) .....	15
3.2.5.3.	Structural Interaction Fingerprint (SIFt) analysis .....	15
4.	Results and Discussion .....	16
4.1.	Data-based approach .....	17
4.1.1.	Data collection and physiochemical characterization .....	17
4.1.2.	MMP analysis and fingerprint clustering .....	19
4.2.	Structure-based approach.....	21
4.2.1.	Binding site detection and sequence alignment .....	21
4.2.2.	Pharmacophore creation and screening.....	23

4.2.3.	Docking studies.....	25
4.2.3.1.	Determining most promising binding pocket .....	25
4.2.3.2.	Separation of inhibitors and non-inhibitors in docking studies .....	26
4.2.3.3.	Sulfonamide docking studies .....	27
4.2.3.4.	Taurocholate docking studies .....	30
4.2.3.5.	Interaction fingerprint clustering .....	31
4.2.3.6.	Literature in support of proposed binding mode .....	34
5.	Conclusion and Outlook .....	36
6.	References .....	38
7.	Appendix .....	42
7.1.	Supplemental material .....	42
7.1.1.	Data-based approach .....	42
7.1.1.1.	TGD-based pairs .....	42
7.1.1.2.	GpiDAPH3-based pairs .....	43
7.1.1.3.	MACCS-based pairs .....	43
7.1.2.	Structure-based approach.....	44
7.1.2.1.	Sequence alignment .....	44
7.1.2.2.	Residues of calculated binding pockets.....	46
7.1.2.3.	Visualization of binding pockets and e-pharmacophores.....	47
7.2.	Abstract .....	51
7.3.	Zusammenfassung.....	52
7.4.	Abbreviations.....	53

# 1. Introduction

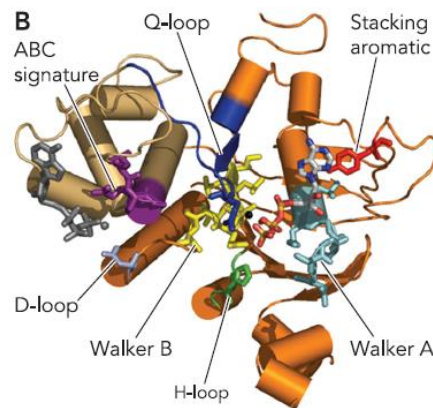
## 1.1. ABC transporter family

The ABC transporters constitute one of the largest protein families, that can be found in bacteria, archaea and eukaryotes. They facilitate active transport of substrates upon ATP binding in their conserved nucleotide binding domains, which led to the name of ABC (= ATP-binding-cassette) transporters. They can act as importers and exporters, however in mammals only export mechanisms are known.<sup>1</sup>

There are 48 human ABC transporter genes, that are subdivided into seven families, named A to G. Their endogenous role is to export and therefore influence homeostasis of hormones, lipids, ions, hormones and other compounds.<sup>2</sup> The ABCA family is involved in lipid trafficking. Mutations were associated with diseases such as Tangier disease T1 and familial high-density lipoprotein deficiency. The ABCB subgroup is unique to mammals and plays an important role in drug discovery due to the transporters' role in multi-drug resistance (MDR) in cancer. The family contains four full-transporters and seven half-transporters.<sup>3</sup> Mutations were connected to diseases such as diabetes type 2, coeliac disease and several cholestatic liver diseases. The subgroup C is best known for its member ABCC7, or rather CFTR, the cystic fibrosis gene. The different members of the group transport diverse substrates and have also been implicated in MDR. Diseases connected to this family include cystic fibrosis, Dubin-Johnson syndrome and diabetes type 2. ABCD members are also known as "adrenoleukodystrophy (ALD) or peroxisomal transporters. Mutations in these members cause ALD and Zellweger syndrome. Subfamilies E and F only contain the ATP-binding domain, but no transmembrane domains, suggesting that these members are not acting as transporters. The ABCG genes have been associated with sterol accumulation, disorders and atherosclerosis. ABCG2 is also known for its MDR activity.<sup>3</sup>

Structurally, the transporters consist of two so-called nucleotide binding domains (NBDs) and two transmembrane domains (TMDs). The NBDs are located in the cytoplasm and are responsible for ATP binding and hydrolysis. The motif of the NBDs is conserved among all ABC transporter types. It consists of a catalytic core site and an  $\alpha$ -helical domain. The catalytic core site includes the Walker A motif (P-loop) which interacts with the phosphate group of ATPs. Additionally, a glutamate residue of the Walker B motif acts as a base to activate a water molecule for the nucleophilic attack of the  $\gamma$ -phosphate of ATP. The  $\alpha$ -helical domain contains the ABC-family signature motif LSGGQ, that is

involved in nucleotide binding.<sup>1</sup> The D, H and Q loops, which are involved in pi-stacking interactions, are also unique to the family.<sup>4</sup>



*Figure 1: Conserved motifs of nucleotide-binding-domain*<sup>5</sup>

The TMDs are comprised of membrane-spanning  $\alpha$ -helices and are responsible for substrate binding and translocation. The sequence of the TMDs is not conserved among the subfamilies which explains the diverse substrates and different selectivity of ABC transporters.<sup>5</sup> Each NBD and TMD are coupled through an  $\alpha$ -helix, that is located in the cytoplasmic loop of the TMDs. The coupling helix is necessary to facilitate structural change in the TMDs upon ATP binding in the NBDs.<sup>1</sup>

Several transport mechanisms for the ABC family have been proposed in the past, including the switch model, the constant contact model and the reciprocating twin-channel model. Experimental findings suggest that different ABC transporters use different ways of transporting their substrates.<sup>6</sup> However, the basic principle of the mechanism has been established. The nucleotide binding domains must dimerize for efficient ATPase activity. As the NBDs come together, the TMDs are pulled apart by the coupling helices. It is believed that the substantial conformational change in the TMDs lowers the affinity for the substrate and therefore translocation out of the cell occurs. The exact order and different intermediate states of the transport mechanism are still unknown for most of the clinically relevant human transporters, especially the ABCB subfamily.<sup>6</sup>



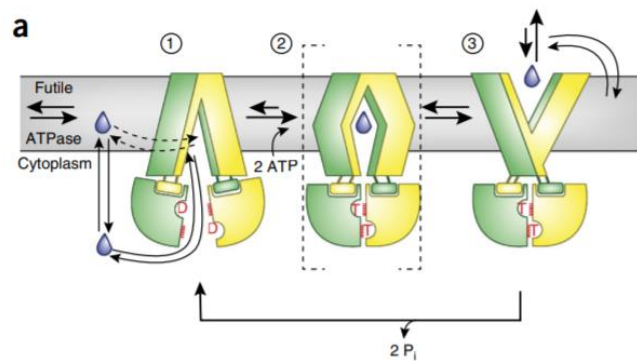


Figure 2: Alternating access model; substrate binding might not only be possible from the cytoplasm but also the inner leaflet of the lipid bilayer, the three states of inward-open, inward-occluded and outward-open are depicted<sup>6</sup>

Overall, ABC transporters play a major role in drug discovery due to their interactions with xenobiotics. They affect diverse pharmacological properties, including oral bioavailability and hepatobiliary, intestinal and urinal excretion.<sup>7</sup> Due to this fact a lot of research has been dedicated to this area to gain insight into structural mechanisms of protein-ligand interactions with these transporters.

## 1.2. Enterohepatic circulation of bile acids

### 1.2.1. Bile acids – structure, synthesis and physiological function

Bile flow is an important pathway for facilitating the elimination of endogenous compounds and metabolites, as well as xenobiotics. The major components of bile are bile acids (BAs).<sup>8</sup> BAs are synthesized from cholesterol via the cytochrome P450 proteins in the endoplasmic reticulum of hepatocytes. Subsequently, they are transported into peroxisomes and conjugated with taurine or glycine, increasing their hydrophilicity, reducing their toxicity and adding their typical amphiphilic character. Primary BAs include taurocholic acid (TC), glycocholic acid (GC), taurochenodeoxycholic acid (TCDC) and glycochenodeoxycholic acid (GCDC).<sup>9</sup>

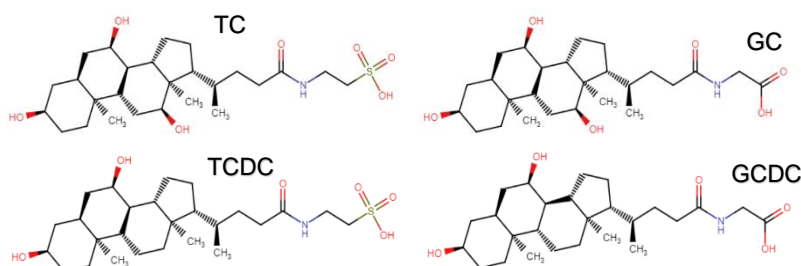


Figure 3 (from left to right and top to bottom): Main substrates of bile salt export pump

After synthesis, BAs are excreted from the liver into the bile and stored in the gallbladder. Upon food ingestion, bile acids are released to the small intestine to act as emulsifier for dietary lipids and fat-soluble vitamins.<sup>10</sup> In the intestinal lumen several biotransformations

mediated by bacteria of the microbiota influence the structural diversity of BAs. Conjugated BAs get hydrolyzed, resulting in their unconjugated and amino acid moieties. Unconjugated BAs can be further transformed to secondary BAs such as deoxycholic acid and lithocholic acid.<sup>9</sup> After the release to the small intestine, most of the bile acids get reabsorbed and return to the liver.<sup>10</sup> Approximately 5 % are not transported back, but are eliminated through the feces.<sup>11</sup> The exact pathway is described in 1.2.2.

Apart from their role as fat emulsifier, bile acids also act as signaling molecules via interaction with several receptors, including the farnesoid X receptor (FXR), the pregnane X receptor (PXR), the vitamin D receptor (VDR) and others. Through these pathways they are involved in energy, glucose, lipids and lipoprotein metabolism.<sup>10</sup>

### 1.2.2. The role of transporters in enterohepatic circulation

The enterohepatic circulation of bile acids heavily depends on transporter proteins. Figure 4 shows the transporters involved at each stage of the cycle. The secretion of bile acids from the hepatocytes into the gall bladder and bile duct is mediated by the bile salt export pump (BSEP) and the multidrug-resistance-associated proteins (MRP). If bile accumulation in the liver is excessive, basolateral export systems of the liver mediated by MRP or the organic anion transporting polypeptide 2 (OATP2) are used. With the help of the apical sodium-dependent bile acid transporter (ASBT), the bile acids get absorbed in the distal ileum. Binding to the ileal bile acid-binding protein (IBABP), they travel across the enterocyte and get exported into the portal circulation by the organic solute transporter alpha and beta (OST $\alpha/\beta$ ). The uptake into the hepatocyte is mediated by the sodium/taurocholate co-transporting polypeptide (NTCP) and the organic anion transporting polypeptide 1 (OATP1), where the cycle starts again.<sup>9</sup>

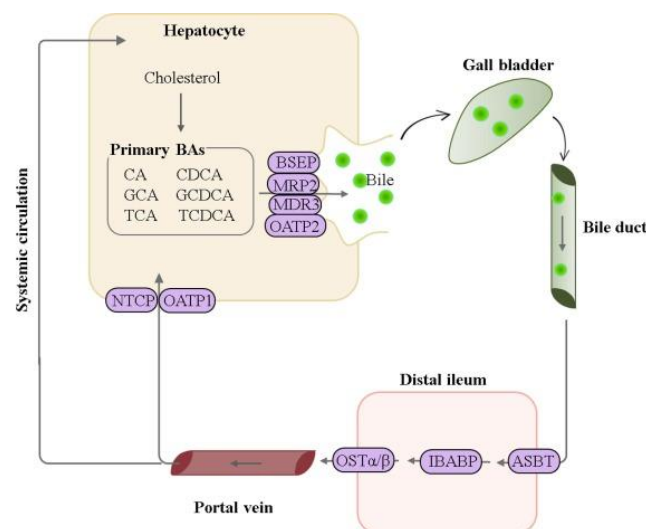


Figure 4: Depiction of the enterohepatic circulation as described in 1.2.2.<sup>9</sup>

Genetic mutations or inhibition of involved transporters can lead to impaired bile flow and the accumulation of bile acids in the liver. Due to their amphiphilic character, BAs can become toxic if they persist in tissues with high concentrations. This ultimately leads to liver diseases such as necrosis, steatosis and cholestasis, with the latter one accounting for most of the cases. As the bile salt export pump serves as primary route of canaliculi elimination of bile acids, this transporter has been implicated in severe liver injury upon impaired transport capacity.<sup>12</sup>

### 1.3. Bile salt export pump

The bile salt export pump is encoded by the gene ABCB11, which is located on the chromosome 2q24 in humans, and belongs to the ABC transporter superfamily. The protein is encoded by 27 exons, following the first untranslated exon, and consists of 1321 amino acids.<sup>13</sup> Expression of the transporter takes place in hepatocytes, with its main localization in the canalicular membrane. BSEP expression is regulated by the farnesoid X receptor, which forms a heterodimer with the retinoid X receptor. Upon ligand binding, the heterodimer binds to an FXR response element in the promoter region of BSEP.<sup>14</sup>

Structurally, BSEP is a full-length ABC transporter, consisting of two TMDs and two NBDs in one polypeptide chain. The resolved cryo-EM structure shows the transporter in its inward-facing apo structure. Each TMD consists of 6 transmembrane helices and is connected through coupling helices with the NBD to allow structural change upon ligand binding. Notably, the Walker B motif shows a degenerated active site, as the catalytic residue glutamate is replaced by methionine.<sup>15</sup>

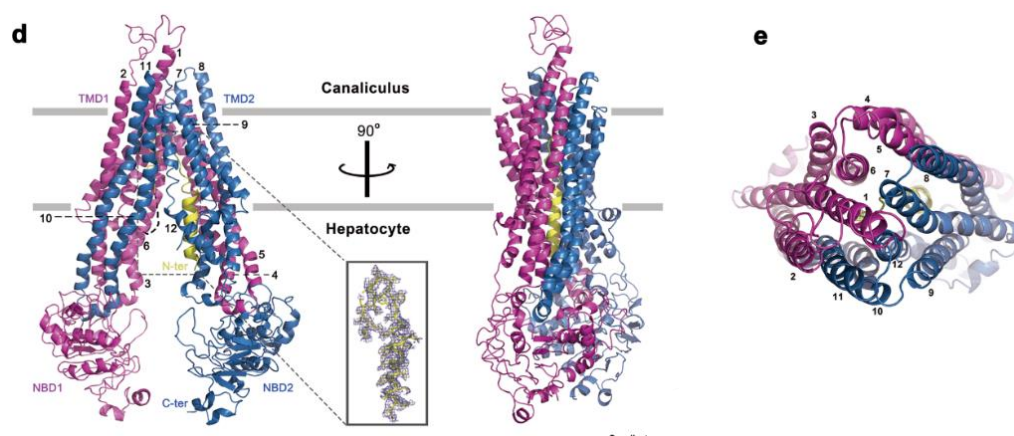


Figure 5: Cryo-EM structure of BSEP released in April 2020 by Wang et. al with two perpendicular and one top view<sup>15</sup>

As set out above, the endogenous role of BSEP is to export BAs from hepatocytes into the canaliculi. The transporter mainly exports monovalent bile acids, such as taurocholic

acid and glycocholic acid, as well as the secondary deoxycholic acid.<sup>16</sup> Substrate clearance was determined to be in the order of TCDC > GCDC > TC > GC, which is favorable due to the higher toxicity of chenodeoxycholate compared to cholate derivatives.<sup>17</sup> The transporter is highly specific for its substrates, it can however be inhibited by several classes of drugs, which has been implicated in the risk of cholestatic drug-induced liver injury (DILI). In vitro assays implied that intracellular accumulation of BAs is one of the lead drivers of DILI.<sup>18</sup>

Apart from BSEP inhibition, several mutations of the transporter are known, which lead to severe liver diseases including progressive familial intrahepatic cholestasis (PFIC), benign recurrent intrahepatic cholestasis (BRIC), low phospholipid associated cholelithiasis (LPAC), Wilson's disease, and others.<sup>19</sup>

#### 1.4. Drug-induced liver injury

Drug-induced liver injury is the most common cause of acute liver failure in the USA and Europe. It is also one of the top adverse-drug reactions (ADRs) responsible for attrition of compounds in the drug development process, as well as withdrawals from the market.<sup>20</sup> Due to the partly poor correlation of hepatic side effects between animals and humans, DILI is regularly undetected in preclinical studies, leading to an increased risk of serious consequences, including death of patients.<sup>21, 22</sup> DILI can be categorized in two groups: intrinsic and idiosyncratic DILI. Intrinsic DILI is dose-dependent and to high extents predictable by animal studies. Idiosyncratic DILI however shows a late onset of symptoms and no clear dose dependency.<sup>20</sup> Various mechanisms were described to be involved in idiosyncratic DILI, including inhibition of transporters, mitochondrial injury and oxidative stress.<sup>23</sup> In 2016 the Critical Path Institute's Predictive Safety Testing Consortium (C-Path PSTC) addressed the major role of BSEP inhibition and perturbation of bile acid homeostasis in DILI, which led to an industry-wide consensus on the importance of assessing BSEP inhibition in the early stages of drug discovery.<sup>24</sup>

## 2. Aim of the thesis

Due to the strong correlation between bile salt export pump inhibition and DILI-associated diseases such as cholestasis, it is of great importance to elucidate structural patterns that cause the inhibition of the transporter. Several machine-learning based classification models and a few ligand-based pharmacophore approaches have been published in the past. However, since the protein structure of human BSEP was only released in April 2020, so far, no structure-based investigations have been reported on this transporter, except for one homology model-based approach. Therefore, the thesis aims to combine data-based and structure-based methods to investigate molecular features causing BSEP inhibition, as well as potential binding sites and protein-ligand interactions of inhibitors. The goal was to discover new molecular aspects to contribute to a better understanding of BSEP inhibition and therefore addressing key issues of DILI.

In order to achieve this goal, inhibitory data on the transporter must be collected in the first step. Different data science approaches can be used to find trends in bioactivity data. In this case, descriptor calculation, matched molecular pair analysis and fingerprint analysis were implemented, to find structural differences, or physicochemical property shifts between inhibitors and non-inhibitors.

Subsequently, the results of the data-based approach were planned to be investigated with molecular docking studies. Since there is only an unbound structure of the transporter available, possible binding pockets of the protein must be investigated first. Different tools are available to achieve this task. In the next step, structure-based pharmacophores were created to determine the most likely binding pocket. Following, inhibitors and non-inhibitors were docked to investigate differences in binding and possible causes for differences in inhibitory activity. In the end, a binding mode hypothesis was established. The hypothesis was reevaluated by docking the substrate taurocholate in the proposed binding pocket.

## 3. Material and Methods

### 3.1. Data-based approach

The data-based approach was conducted doing a detailed data analysis of inhibitors and non-inhibitors of BSEP. After collecting available bioactivity data, physicochemical descriptor calculation, matched molecular pairs analysis and fingerprint clustering were performed. All depicted structures were drawn using Marvin JS by ChemAxon.

#### 3.1.1. KNIME Analytics Platform

All data-based approaches were conducted using the KNIME Analytics Platform by the University of Konstanz. KNIME is an open-source software based on the Eclipse platform that enables accessing and processing data using workflow systems. Workflow systems are built up of different types of connected data transformation points, which are referred to as nodes. A graphical user interface using drag-and-drop options for nodes enable the user to build individual workflows. Different plug-ins are available including CDK, Weka, Python programming and Schrödinger. The plug-ins offer applications specifically for cheminformatic and drug discovery purposes such as computing quantitative structure-activity relationship (QSAR) descriptors, implementing machine learning algorithms or visualizing molecular structures.<sup>25</sup>

##### 3.1.1.1. Data collection

Bioactivity data on BSEP was collected from the manually curated bioactivity database ChEMBL and the open chemistry database PubChem using the search for the encoding gene ABCB11. A KNIME workflow was built to access the desired data. Available data on ChEMBL was downloaded through the ChEMBL extractor node using the ID CHEMBL6020. Only data containing activity values with the operator “=” were filtered and all values were converted to the unit  $\mu\text{M}$ . The structural information was available as SMILES code. PubChem was accessed via PugRest and the GeneID 8647 was used to retrieve compound IDs of bioactivity records and corresponding SMILES codes of the molecules.

Next, duplicates were filtered out and the dataset was standardized using the standardizer metanode by Jennifer Hemmerich.<sup>26</sup> Mixtures of molecules and molecules including nonorganic atoms were filtered out. Only entries containing IC<sub>50</sub> values were kept. For molecules with several measured activity values, the higher IC<sub>50</sub> value was retained. For the data-based approach the stereochemistry was removed and the isomers with higher IC<sub>50</sub> values were kept. Binary classification was added to the data

set using a threshold of 10  $\mu\text{M}$ . Furthermore, the activity data set was combined with an in-house binary classification data set, where the threshold for activity was also set to 10  $\mu\text{M}$ . For the structure-based approach the original stereochemistry of the data set was included.

#### 3.1.1.2. Physicochemical property characterization

Descriptor calculation is a powerful method to characterize different properties of small molecules and is regularly used in drug discovery for purposes of classification and regression models. Descriptors are based on different mathematical calculations, starting from simple counts, e.g. number of oxygen atoms, to complex, rather uninterpretable descriptors employing e.g. quantum chemistry.<sup>27</sup>

The following descriptors were calculated in KNIME with the RDKit Descriptor Calculation node<sup>28</sup> for the analysis of the data set:

- SlogP: partition coefficient of solubility in octanol:water ; measure for lipophilicity
- AMW: molecular weight [g/mol]
- SMR: molar refractivity
- TPSA: topological polar surface area
- NumRotBonds: number of rotatable bonds
- NumHBA: number of hydrogen bond acceptors
- NumHBD: number of hydrogen bond donors

#### 3.1.1.3. Matched-Molecular-Pairs (MMPs) Workflow

Matched-Molecular-Pairs are based on the concept of structure-activity-relationship (SAR). SAR links changes in the ligand structure to changes in biological activity on the target protein. The systematic approach of quantitatively linking these changes to adding or removing a certain functional group is the idea of MMPs. The matched pairs can be described as “molecules that differ only by a particular, well-defined, structural transformation”.<sup>29</sup> Overall, MMP analysis has become a useful tool in drug discovery, especially for the field of medicinal chemistry.<sup>30</sup>



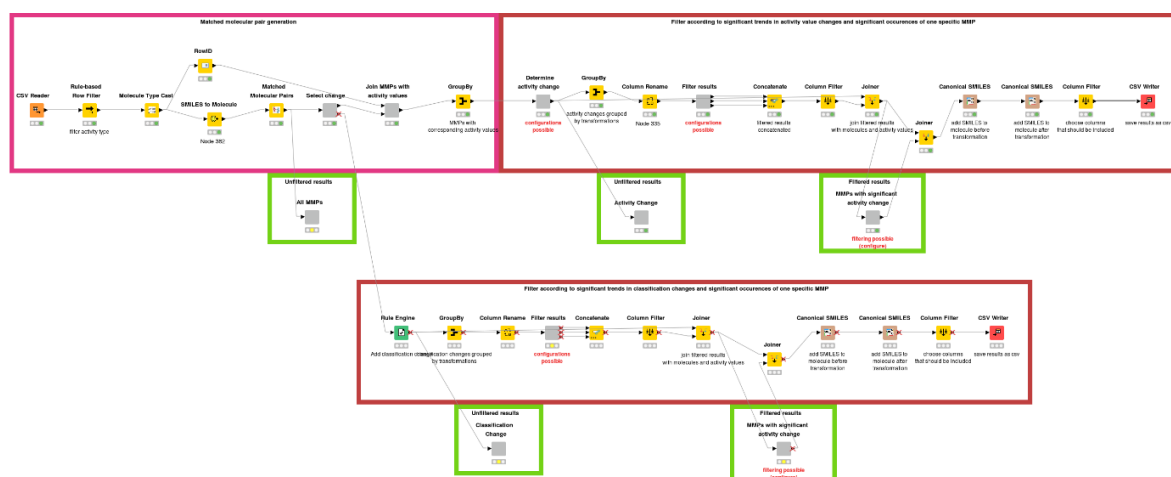


Figure 6: Matched-Molecular-Pairs workflow built in KNIME

The Matched-Molecular-Pairs workflow was built around the available Matched-Molecular-Pairs node by the Chemical Computing Group. The workflow allows to read in bioactivity data in csv format containing structural information in SMILES code, which is transformed to SDF and passed on to the Matched-Molecular-Pairs node. Different thresholds can be set to obtain pairs, Figure 7 shows the chosen settings for the analysis.

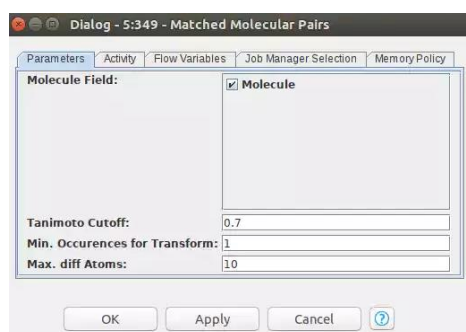


Figure 7: Possible threshold settings for the MMP node with entered values for the conducted analysis

Subsequently, the workflow provides a detailed analysis of the retrieved MMPs. It is possible to use a data set based on concrete activity values, but also data sets containing only binary classification. The workflow visualizes the quantity and distribution of changes in activity. Individual thresholds can be set to filter interesting pairs based on frequency and predefined minimum activity difference. End results can be saved as csv file containing SMILES code of the matched pairs, structural difference in SMILES code, activity difference and activity values of both molecules.

#### 3.1.1.4. TGD-, GpiDAPH3- and MACCS-Fingerprint Clustering Workflow

Another powerful method for identifying structurally related compounds is clustering based on 2D molecular fingerprints. Binary fingerprints describe the presence or absence of a certain property in the molecule. Properties can reach from simple atoms



to functional groups, as well as pharmacophoric patterns and other features. The calculated fingerprints are used to determine similarity between molecules and therefore, homogenous molecule clusters can be created.<sup>31</sup>

A fingerprint clustering workflow was built to further analyze the bioactivity data. Three different fingerprint methods were used:

- TGD (Typed-graph distances, 2-point 2D pharmacophore)
- GpiDAPH3 (Graph-P-Donor-Acceptor-Polar-Hydrophobe-Triangle, 3-point 2D pharmacophore)<sup>31</sup>
- MACCS (Molecular ACCEss System, structure based)<sup>32</sup>

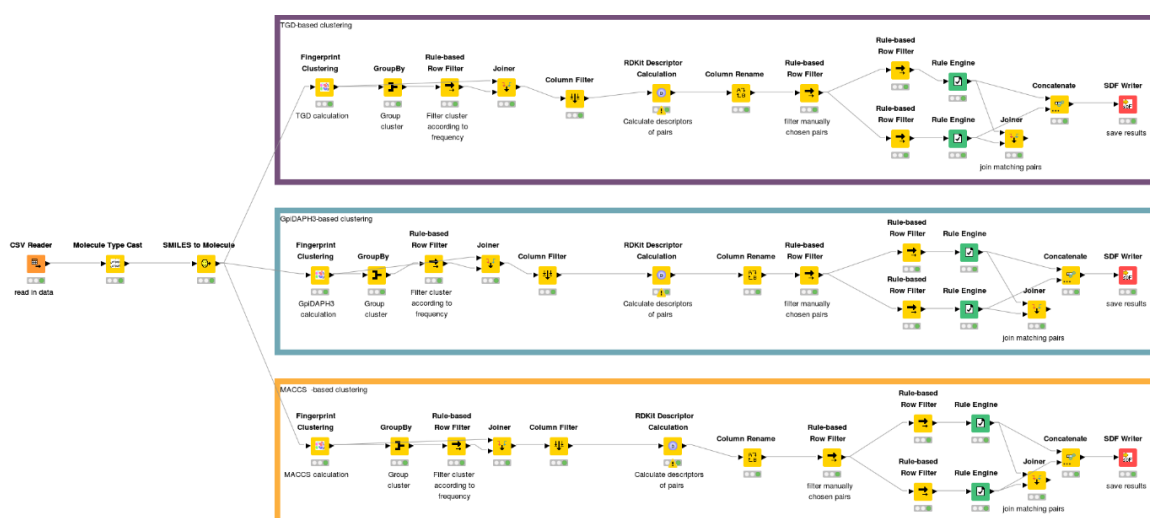


Figure 8: fingerprint clustering workflow in KNIME; purple box displaying TGD-clustering, blue box displaying GpiDAPH3-clustering, orange box displaying MACCS-clustering

The workflow creates clusters based on the three aforementioned methods, which are filtered by predefined minimum cluster size. A minimum cluster size of 5 containing molecules was chosen. Following, physicochemical descriptors were calculated. In the next step pairs of interest were manually chosen and saved.

## 3.2. Structure-based approach

Structure-based investigations included binding site detection of the available apo structure of BSEP and Glide docking, as well as induced-fit docking of inhibitors and the natural substrate taurocholate.

### 3.2.1. Cryo-EM structure BSEP

The structure of the human BSEP was extracted from the protein data bank using the PDB code 6LR0. The apo-form of the inward-open transporter was solved at a resolution of 3.5 Å using cryogenic electron microscopy. The structure was obtained by Wang et al.

by overexpressing ABCB11 in HEK293 cells to obtain the recombinant protein. ATPase assays were conducted ensuring that the protein samples were in a relevant physiological state. The extension of the N-terminus could not be resolved and therefore residues Ser32-Val43 are missing in the reported structure. However, the scientists constructed a truncated version which showed no significant difference in bioassays and therefore these residues do not seem to be essential in substrate binding or translocation.<sup>15</sup>

### 3.2.2. Protein and ligand preparation

Certain preparation steps for proteins and ligands need to be conducted prior to docking studies. Preparation of proteins include adding hydrogen bonds, optimizing hydrogen bonds and removing atomic clashes. For ligands 3D structures need to be computed and possible tautomers and ionization states need to be considered. These preparations are necessary to obtain comprehensive and significant docking results.<sup>33</sup>

In MOE, the protein was prepared using the protonate 3D and the structure preparation panel with default settings. Protein preparation in Maestro was conducted using the Protein Preparation Wizard by Schrödinger. Default settings were used to preprocess the protein, minimize the structure and optimize hydrogen bonds. The standard settings were used except for the pH range, which was set to  $7.0 \pm 0.5$ .

Ligand preparation was done using LigPrep by Schrödinger. Chiralities were defined from the 3D structure and the pH range was changed to  $7.0 \pm 0.5$ . All other settings were left as default.

### 3.2.3. Binding site detection

The investigation of possible binding pockets was done by analyzing a related co-crystallized transporter and by binding site calculations using different algorithms.

#### 3.2.3.1. Sequence alignment of P-glycoprotein and BSEP

The closest related transporter of BSEP is P-glycoprotein (P-gp) with a sequence identity of approximately 49 %. A co-crystallized structure of P-gp with the inhibitor tariquidar is available on the protein data bank (PDB code: 7A6E).

The sequences were aligned by the Clustal Omega program offered by Uniprot. The UniProt identifiers P08183 (P-gp) and O95342 (BSEP) were used. The binding pocket of P-gp in complex with MRK16 Fab and tariquidar was analyzed in LigandScout and relevant residues were translated to the corresponding residues in the BSEP structure.

### 3.2.3.2. Binding site calculation

Binding site calculation was done using different algorithms. The two main programs used were SiteFinder in MOE by the Chemical Computing Group and SiteMap in Maestro by Schrödinger.

SiteFinder is a geometry based binding site detection algorithm, where no energy calculations are considered. Alpha spheres are calculated by populating receptor cavities with 3D points, that are triangulated, which results in simplexes. The results are filtered by criteria of accessibility and solvent exposition. Filtered alpha spheres are clustered and ranked by the number of hydrophobic contacts with the receptor atoms. Every pocket must contain at least one hydrophobic ranked alpha sphere.<sup>34</sup>

SiteMap on the other hand is an energy-based cavity finding algorithm. The binding site detection occurs in three steps. The detection of cavities takes place first, where a 1 Å grid of site points is defined over the whole protein. Site points that collide with receptor atoms, do not show sufficient enclosure, or are too far away from the protein are filtered out. Site points are merged by a predefined threshold distance, which by default is 6.5 Å. During the second step the found cavities are characterized as hydrophobic or hydrophilic using Van der Waals and electric field grids. The last step considers different scores to evaluate the characterized sites, including the SiteScore, the DScore, number of site points, exposure/enclosure, contact and site volume. Additionally, hydrophobic/hydrophilic character and donor/acceptor properties can be visualized in the binding site.<sup>34</sup>

The default settings of the described algorithms were used to investigate potential binding pockets in BSEP.

In addition to SiteFinder and SiteMap, additional cavity detecting tools were used to detect further pockets and to support the previous found pockets. Other programs used include LigandScout, DoGSiteScorer and P2RANK.

### 3.2.4. Structure-based 3D-Pharmacophore creation and screening

IUPAC defines pharmacophores as “an ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target and to trigger (or block) its biological response”. Pharmacophore modelling can either be done by considering only structural features of known ligands or by using the structural information of the protein in structure-based pharmacophore designs.<sup>35</sup>

Structure-based pharmacophores can be obtained by analyzing ligand-protein complexes or by investigating residue features in possible binding sites.<sup>35</sup> Since the BSEP structure is only available in its apo form, structure-based pharmacophores were created using e-pharmacophore models by Phase. Phase calculates pharmacophore sites which are characterized by type, location, and if applicable, direction. Six different pharmacophoric features are described: hydrogen bond acceptor (A), hydrogen bond donor (D), hydrophobic (H), negative ionizable (N), positive ionizable (P) and aromatic ring (R).<sup>36</sup>

Structure-based e-pharmacophores were built using the properties of residues surrounding the receptor cavity in 3 Å distance. In case of a non co-crystallized protein, e-pharmacophores are built by docking fragments to the receptor using Glide XP. Common features that maximize the binding energy are chosen. Excluded volumes are added for the regions that are occupied by receptor atoms.<sup>37</sup> The standard settings were used, calculating 7 pharmacophoric features.

The e-pharmacophores were used to evaluate the calculated binding sites doing a pharmacophore screening of inhibitors and non-inhibitors. The screening was conducted using phase ligand screening.<sup>38</sup> Compounds were set to must match at least 5 out of 7 features to be identified as hits.

### 3.2.5. Docking studies

Since the 1980s molecular docking has gained great importance in the drug discovery area. Docking studies can predict ligand conformation and ligand-protein interactions within a target binding site. Energy calculations provide docking scores that rank different ligands in the order of the stability of the proposed ligand-receptor complexes.<sup>39</sup> Interactions between ligand and target receptor can be divided into five major molecular forces, including covalent bonding, Van der Waals interactions, hydrophobic interactions, hydrogen bonding and ionic interactions.<sup>40</sup>

Docking studies can either be performed as rigid docking, semi-flexible docking or flexible docking. During rigid docking the structures of both protein and ligand cannot change, semi-flexible docking allows changes in ligand conformation and flexible docking treats ligands as well as proteins as flexible structures. The computational effort and cost increases with the flexibility.<sup>40</sup>

#### 3.2.5.1. Glide Docking

Glide stands for **g**rid-based **l**igand **d**ocking with **e**nergetics and is a rigid docking algorithm. It was designed to perform as close to exhaustive search algorithms as

possible while retaining sufficient computational speed. In the preprocessing step a receptor grid is generated, representing shape and properties of the protein for more accurate scoring of ligand poses. Following, initial ligand conformations are produced by searching for local minima of the ligand torsion-angle space in an exhaustive manner. After this preprocessing step, the most promising poses of each ligand are minimized using an OPLS-AA force field. Finally, the three to six most promising poses of each ligand are subjected to a Monte Carlo procedure to examine nearby torsional minima. For scoring the ligands a combination of the GlideScore, which describes the ligand-receptor molecular mechanics interaction energy, and the ligand strain energy are used.<sup>41</sup>

The grid generation was done calculating the centroid of specified residues. The residues were manually entered, considering all residues within 3 Å of the calculated binding sites by SiteFinder and SiteMap. Prepared ligands were docked using default settings.

#### 3.2.5.2. Induced-Fit Docking (IFD)

Induced-Fit Docking is a flexible docking algorithm, where movement of the receptor upon ligand binding is made possible. This can have a significant impact on the calculated energy of a ligand-protein complex. Schrödinger developed a technology that “accounts for receptor flexibility in ligand-receptor docking by iteratively combining rigid receptor docking (using Glide) with protein structure prediction and refinement (using Prime).”<sup>42</sup> This methodology allows small backbone shifts as well as significant side chain conformation changes.<sup>42</sup> The procedure itself is comprised of four essential steps, starting with an initial Glide docking into a rigid receptor. Secondly, sampling of the protein for each ligand pose is conducted using the refinement module of prime. Only residues with at least one atom in 5 Å distance of the ligand poses are sampled. This refinement results in low energy induced-fit structures. In the third step, ligands are redocked. Lastly the poses are scored by means of docking energy and receptor strain and solvation terms.<sup>43</sup>

For the IFD run the standard protocol was used, yielding up to 20 poses for each ligand. Prepared inhibitors and the substrate taurocholate were docked using default settings.

All energy values mentioned in this thesis are reported in kcal/mol.

#### 3.2.5.3. Structural Interaction Fingerprint (SIFt) analysis

Structural interaction fingerprints translate 3D interactions of a receptor-ligand complex to binary digits. As it is almost impossible to investigate every ligand pose in the times of high-throughput virtual screening, SIFt provides an algorithm for filtering, clustering and

analyzing docking poses more sufficiently.<sup>44</sup> The following protein-ligand interactions can be computed in Maestro.

- Any Contact
- Backbone Interaction
- Sidechain Interaction
- Polar Residues (including Arg, Asp, Glu, His, Asn, Gln, Lys, Ser, Thr)
- Hydrophobic residues (including Phe, Leu, Ile, Tyr, Trp, Val, Met, Pro, Cys, Ala)
- Hydrogen Bond Acceptor
- Hydrogen Bond Donor
- Aromatic Residue (including Phe, Tyr, Trp)
- Charged Residue (including Arg, Asp, Glu, Lys)<sup>45</sup>

After calculating chosen fingerprints, it is possible to cluster ligands based on their SIFts.

## 4. Results and Discussion

### 4.1. Data-based approach

#### 4.1.1. Data collection and physiochemical characterization

After filtering and merging the collected data as described in 3.1.1.1, a data set containing 1513 molecules was obtained, consisting of 298 inhibitors and 1215 non-inhibitors. 367 molecules contain concrete activity information (IC<sub>50</sub>), the rest of the compounds contain only binary classification information. This can be explained by data of publications, where IC<sub>50</sub> values are published as “higher” or “lower” instead of a concrete value. An example would be a paper by Morgan et. al, where the highest concentration measured was 133 μM. If there was no measurable BSEP inhibition upon this concentration, the compound was classified as inactive.<sup>46</sup> Analysis of different properties of inhibitors (red) and non-inhibitors (green) with known IC<sub>50</sub> value were conducted, plotting the pIC<sub>50</sub> values against SlogP, SMR, AMW and TPSA. Descriptors were calculated using the RDKit Descriptor Calculation node, pIC<sub>50</sub> values were calculated using the following equation.

$$pIC_{50} = -\log(IC_{50} [\mu M] * 10^{-6})$$

Equation 1

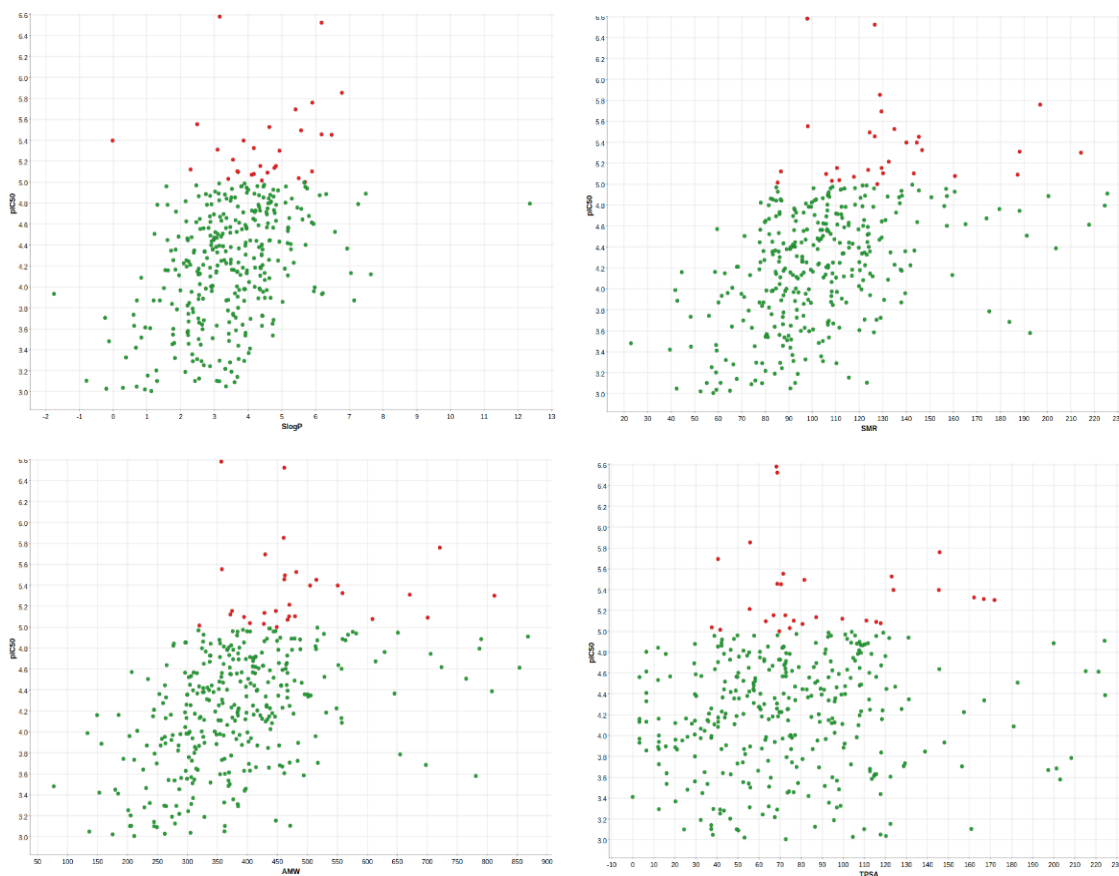


Figure 9 (from left to right and top to bottom): pIC<sub>50</sub> values as function of SlogP, SMR, AMW and TPSA

The figures show the most significant trend for SlogP. A steep ascent of pIC50 values can be observed when going to more lipophilic molecules. A similar trend can be observed for the molecular weight (AMW) and the molar refractivity (SMR). However, in these plots the trend is not as clear as for SlogP. For TPSA, no real activity trend can be observed depending on this descriptor. Nevertheless, these plots need to be interpreted with caution, since only a small number of compounds show concrete IC50 values and could therefore be included.

*Table 1: Mean values and standard deviations of calculated descriptor values of non-inhibitors and inhibitors*

<b>Descriptor</b>	<b>Non-inhibitors</b>	<b>Inhibitors</b>
SlogP	2.3 ± 2.1	4.1 ± 1.6
AMW [g/mol]	344.3 ± 129.8	470.1 ± 111.1
SMR	91.3 ± 33.5	126.0 ± 28.1
TPSA	82.9 ± 50.2	90.9 ± 35.0
NumRotatableBonds (NRotB)	4.8 ± 3.5	6.8 ± 3.3
NumHBA	4.9 ± 2.8	6.1 ± 2.4
NumHBD	2.0 ± 1.8	1.6 ± 1.2

The mean descriptor values support the information obtained from the activity plots. The SlogP is shifted from 2 to 4 between non-inhibitors and inhibitors. Additionally, inhibitors have a higher molecular weight, by approximately 120 g/mol on average. Since SlogP and AMW are correlated with each other these supporting trends are not surprising. Additionally, considering that BSEP is a membrane-spanning transporter, it seems logical that more lipophilic molecules reach the transporter's binding pocket more easily and therefore have a higher potential of inhibiting it. Another observation already made in the activity plots is the substantially higher SMR of inhibitors. Since molar refractivity is dependent on the polarizability of a molecule, this could mean that electronegative moieties, such as halogens, on the molecules might be beneficial for BSEP inhibition. However, global descriptors such as SMR are quite difficult to interpret structurally. TPSA does not show a significant difference between inhibitors and non-inhibitors. Hydrogen-bond acceptors and donors also seem to be in the same range for both classes. However, it seems unexpected that inhibitors show quite a high mean value for HBAs since higher lipophilicity was strongly correlated with higher inhibitory activity. SlogP and hydrogen bonding are not mutually exclusive, but they are interesting partners. Structure-based approaches might give some further insight into this matter. Inhibitors contain 2 more rotatable bonds on average, which could mean that higher flexibility is important for BSEP inhibition.



#### 4.1.2. MMP analysis and fingerprint clustering

For the activity data (367 compounds) the MMP workflow detected 42 MMPs, but no significant activity change trends within the pairs. Using binary classification as activity change parameter (1513 compounds), 406 MMPs were detected, however again no relevant activity changes with a valid number of examples were found.

Since no significant trends could be detected using the approach of MMPs, the data set was further investigated using different approaches of fingerprint clustering. The pharmacophore clustering yielded 17 (TGD) and 5 (GpiDAPH3) clusters, whereas 15 clusters were obtained by using MACCS fingerprints. After manually inspecting the clusters for significant activity changes 5 (TGD), 2 (GpiDAPH3) and 2 (MACCS) clusters were filtered and for each cluster the most and least active molecule were plotted, and descriptor changes were investigated to find further trends. An example can be seen in Figure 10, all investigated pairs are attached in the Appendix under section 7.1.1.

Table 2: Descriptor values of 2 found pairs by TGD-based clustering

Mol.	IC50 [ $\mu$ M]	SlogP	AMW [g/mol]	SMR	TPSA	NRotB	HBA	HBD
1	49.8	4.4	336.2	86.1	62.5	3	3	2
2	18.5	4.8	369.7	86.1	62.5	3	3	2

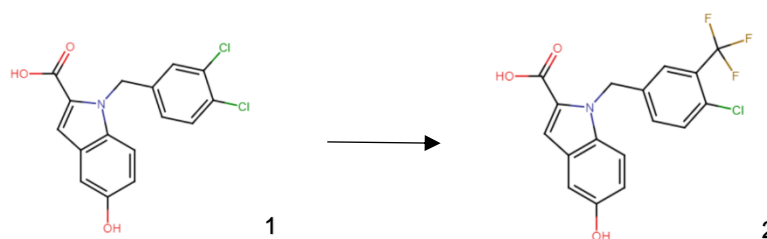


Figure 10: Example of matched pair found by TGD-based clustering

For all compared pairs found by the three algorithms the most significant trends were observed for SlogP and AMW.

Clustering the data revealed prominent structural groups in the data set, including sulfonamides, piperazines, carboxylic acids and steroid based scaffolds. The groups were extracted using SMILES codes. For steroids there was an issue using SMARTS, due to the difference in double bonds in the core structures. Therefore, a similarity search using the SMILES of chenodeoxycholic acid was used to get an estimated number of related molecules in the data set. One structure can belong to several classes.

Table 3: Structural moieties with high abundance in the data set with active/inactive ratio

Class	SMARTS	Hits	Active/inactive [%]
Sulfonamides	<chem>[*]S([*])(=O)=O</chem>	144	32/68
Carboxylic acids	<chem>[*](=O)[O*H,OX2H1]</chem>	346	15/85
Piperazine	<chem>[*]-1-[*]-[*]-[*]-[*]-[*]-1</chem>	131	40/60
Steroid-based	-	~ 249	12/88

The high abundance of these classes makes sense, since they are structurally related to the natural substrates, except for the piperazine moiety. Interestingly, steroid-based scaffolds were found to be mostly inactive, although they contain the structural motif of the natural substrates. However, since this class could only be filtered using similarity search, which is not completely accurate, this trend needs to be interpreted with caution. Sulfonamides and piperazines show the highest actives ratios, which could mean that a positive ionizable group can be beneficial for binding. Since the structural classes were extracted manually, it is not known whether there are other features with high abundance, that were not found using the fingerprint clustering.

Considering all the results, it can be said that lipophilicity and molecular weight show a distinct trend in the data set. This observation is also supported by several literature sources, where these descriptors were already found to be important for BSEP inhibition.<sup>47,48,49</sup> Pedersen et. al also stated that strong inhibitors were correlated with higher flexibility in their models, which supports the observed higher number of rotatable bonds. Additionally, a positive correlation with the abundance of halogen atoms was observed. This could possibly be interpreted as a backup for the higher molar refractivity; however, halogen atoms also lead to an increase of lipophilicity and molecular weight. Interestingly, the publication states that hydrogen bond acceptors were negatively correlated with BSEP inhibition.<sup>48</sup> These findings somewhat contradict the observed trend in Table 1. However other papers suggested hydrogen acceptor properties to be important based on created ligand-based pharmacophores.<sup>47</sup> The observed importance and trends of molecular descriptors, certainly depend on the size and heterogeneity of the data set, as well as the workflow for identifying important descriptors. Nevertheless, the observed molecular trends for BSEP inhibition presented in this thesis support previous findings from literature.

The approach of using matched molecular pair analysis did not work out in this case. The question is, if there is simply not enough concrete bioactivity data for this transporter available at this point, or if the inhibitory activity of compounds cannot be led back to one

concrete structural change. Most likely, it is a combination of both. Especially for heterogeneous data sets and/or for large flexible proteins, the MMP approach might not be feasible. However, the presented MMP workflow can be used for a prescreening of a dataset, giving first insights into possible trends or even yielding concrete QSAR information. The fingerprint clustering method aided in elucidating prominent structural features in the data set, which will be implemented in the structure-based approach.

## 4.2. Structure-based approach

For structure-based purposes the stereoisomers were reincluded, resulting in a dataset of 1832 compounds, with 357 inhibitors and 1475 non-inhibitors.

### 4.2.1. Binding site detection and sequence alignment

The sequence alignment of P-glycoprotein and BSEP showed a sequence identity of 49.2% and 433 similar positions (whole sequence alignment Appendix Figure 28).

```

P08183 MDR1_HUMAN      1  -----MDLEGDRNGGAKKNFFKLNNK--SEKDKKEKKPTVSVFMSFRYSNWLDKL      49
O95342 ABCB1_HUMAN      1  MSDSVILRSIKKFGGEENDGFESDKSYNNDKKSLQDEKKDGVVRVGFQLEFRSSSTDIW      60
      :.  *:.*. *  :.:. :. :. *  :.:. *  :. *:.*. *  :.
P08183 MDR1_HUMAN      50  YMVVGTLAAIIHGAGLPLMLLVFGEMTDIFANAGNLEDLMS-----NIT      93
O95342 ABCB1_HUMAN      61  LMFVGSGLCAFLHGIAQPGVLLIFGTMTDFIDYDVELQELQIPGKACVNNITVWNTSSLN    120
      *:.*. *:.*. *  :. *  :. *:. *  *:. *  :. :.

```

Figure 11: Part of sequence alignment of P-gp and BSEP; \* indicating same residue, : indicating related residue, . indicating non-related residue

Following, the binding pockets of the two bound tariquidar molecules in P-gp were visualized in LigandScout. Important residues for binding were extracted (Appendix Figure 29, Figure 30) and translated to BSEP using the results of the alignment.

Table 4: Translated residues of P-gp to BSEP relevant for inhibitor binding

P-gp	BSEP
Met69	Leu80
Phe72	Phe83
Trp232	Ile259
Leu236	Val263
Ile306	Ile333
Tyr307	Phe334
Tyr310	Tyr338
Phe336	Leu364
Phe728	Tyr772
Phe732	Phe776
Glu875	Gln918
Met876	Thr919
Phe978	Ile1021
Phe983	Leu1026
Phe994	Tyr1038

Subsequently, the binding site detection results calculated by SiteFinder and SiteMap were compared to the determined amino acids. SiteFinder calculated 70 pockets, which are ranked by hydrophobic residue interactions. The first three pockets obtained contained similar residues (Appendix Table 20) to the observed interacting residues in MDR1 and were therefore chosen for closer investigation. Using default settings, SiteMap detected 5 binding pockets, of which two pockets were located in the NBD and therefore not considered as ligand binding sites. The following sites were investigated in more detail.

Table 5: Three top ranked pockets calculated by SiteFinder

Pocket	Size	PLB	Hyd	Side
MOE1	212	5.57	71	142
MOE2	126	4.20	39	84
MOE3	78	1.70	18	48

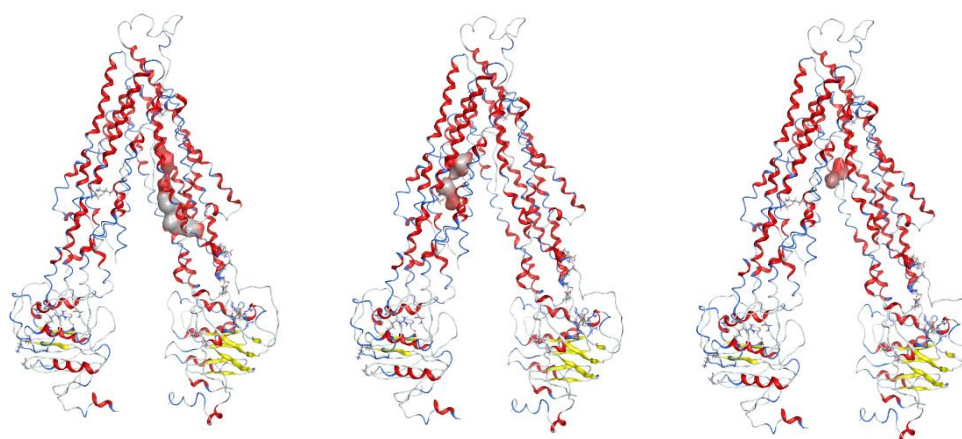


Figure 12: Surfaces of binding pockets detected by SiteFinder; f.l.t.r.: MOE1, MOE2 and MOE3

Table 6: Calculated pockets by SiteMap located in the TMD

Pocket	Size	SiteScore	Dscore	Expos./Enclos.	Phobic/Philic	Don/Acc
Maestro1	400	1.132	0.962	0.281/0.896	0.188/1.582	0.750
Maestro2	147	1.022	0.962	0.533/0.731	0.209/1.279	0.847
Maestro4	135	1.118	1.180	0.604/0.773	1.171/0.688	1.225

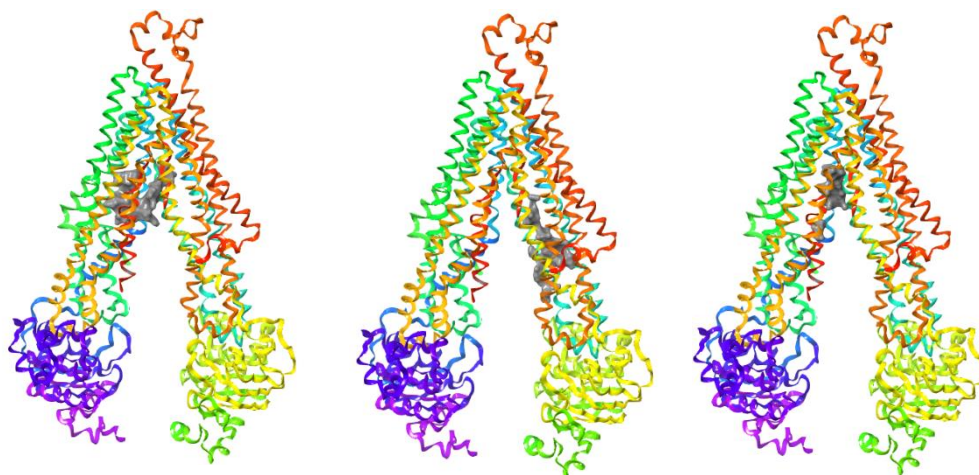


Figure 13: Binding pockets calculated by SiteMap in Maestro, f.l.t.r.: Maestro1, Maestro2 and Maestro4

Looking at the location of MOE1 and Maestro2 and of MOE2 and Maestro1, it can be observed that these pockets are located in the same area and show similar residues (Appendix Table 20). However, they show differences in size. MOE3 and Maestro4 are both located in the center of the transporter, but they also differ in size and interacting residues.

#### 4.2.2. Pharmacophore creation and screening

E-pharmacophores were created for all six binding sites. The following patterns were obtained. Visualizations of all pharmacophore hypotheses are included in the Appendix (Figure 31 – Figure 35).

Table 7: e-Pharmacophores of calculated binding pockets

Pocket	A	D	H	N	R	P
MOE1	1	1	1	1	3	-
MOE2	2	4	-	-	1	-
MOE3	1	2	-	1	3	-
Maestro1	1	4	-	-	1	-
Maestro2	1	4	-	1	1	-
Maestro4	2	2	-	-	3	-

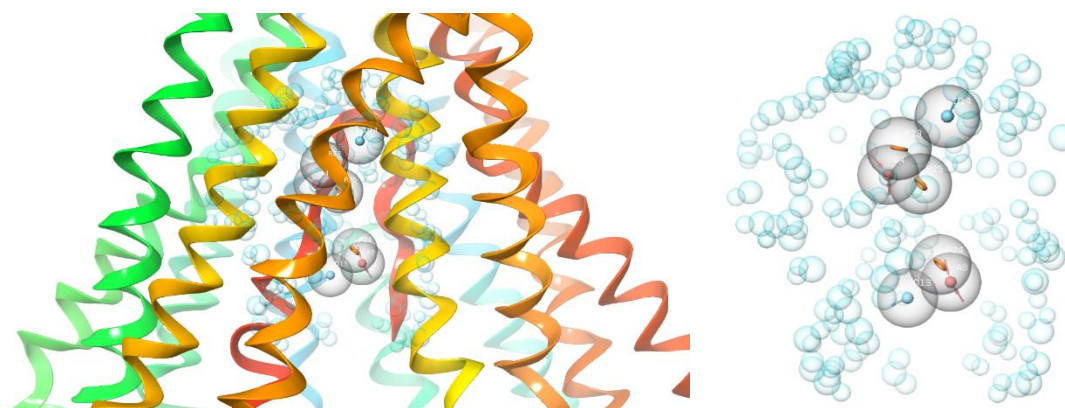


Figure 14: Pharmacophore of Maestro4 with surrounding transporter helices (left) and only with excluded volumes (right); blue vector = HBD; red vector = HBA; orange ring = aromatic; blue spheres = exclusion volumes

The pharmacophores of MOE2, Maestro1 and Maestro2 reflect their hydrophilic character, which could already be observed in Table 5 and Table 6. Every pocket shows at least one hydrogen bond acceptor and one hydrogen bond donor property. On the one side, this is not surprising, since backbones of proteins are made up of amide bonds, that can easily interact in hydrogen bonding. However, the e-pharmacophores do not include all possible pharmacophores, but rather important features for ligand binding, which makes this trend quite unexpected, considering the fact that higher lipophilicity results in higher inhibitory activity. Interestingly, no pocket contains positive ionizable features and only one pocket contains a hydrophobic feature.

Considering the structure of bile acids, one would expect more hydrophobic/aromatic features on one side of the pocket and hydrophilic features on the other side, provided that one of the found cavities is the orthosteric binding site. This spatial arrangement is mostly given in Maestro4 and MOE1.

In the next step, the pharmacophore screening was conducted, and the results were analyzed.

Table 8: Results from pharmacophore screening of six proposed binding pockets

Pocket	Hits [%]	Active/inactive [%]	Found actives [%]	Prominent groups
MOE1	50	25/75	64	Piperazines
MOE2	25	19/81	24	Steroids
MOE3	31	28/72	44	Sulfonamides
Maestro1	29	13/87	19	Steroids
Maestro2	13	11/89	7	Steroids
Maestro4	53	28/72	76	Sulfonamides, carboxylic acids



Table 8 shows that separation of actives and inactives was not possible with the created e-pharmacophores. Interestingly, the more hydrophilic pockets (MOE2, Maestro1, Maestro2) found more molecules with a steroid core structure, whereas the other pockets found more sulfonamides, carboxylic acids and piperazines and no or only a few steroids. This can be explained by the molecular structure of the steroid based molecules in the data set, which contain several hydroxyl or carbonyl moieties. However, most of these structures are not active, explaining the low number of active hits for these three pharmacophores.

MOE1, MOE3 and Maestro4 show the highest number of found actives and the best active/inactive ratios. Nevertheless, they still show a high rate of inactives, making the pharmacophores not specific enough for filtering inhibitors.

For further input on the most likely binding site, the six pockets were further investigated in docking studies.

#### 4.2.3. Docking studies

##### 4.2.3.1. Determining most promising binding pocket

To reevaluate the hit rate of the structure-based pharmacophores, all active molecules were docked to all six binding pockets using rigid docking (Glide). Following, the number of molecules with a docking score lower than or equal to -5.0 were determined to compare the results.

*Table 9: Docking results of active inhibitors to all six potential binding pockets*

Pocket	Docked actives (all) [%]	Docked actives (score <= -5.0) [%]
MOE1	94	4
MOE2	79	19
MOE3	52	18
Maestro1	96	20
Maestro2	83	3
Maestro4	96	<b>86</b>

Interestingly, the percentage of docked molecules is significantly higher than the number of hits in the pharmacophore screening. This can be explained by the more restrictive filtering in pharmacophore screening, where compounds must match at least 5 out of 7 features to be identified as hits. In docking studies on the other hand, the ligand must only fit into the binding pocket and result in a negative binding free energy. Although all pockets show somewhat similar results in terms of the number of docked actives, looking

at molecules with a score lower than or equal to -5.0, Maestro4 greatly outperformed all other pockets. Consequently, Maestro4 was chosen for further docking studies.

#### 4.2.3.2. Separation of inhibitors and non-inhibitors in docking studies

To support the binding site hypothesis of Maestro4 and to identify important structural aspects for BSEP inhibition, all molecules were docked to Maestro4.

Table 10: Docking results of all inhibitors and non-inhibitors in Maestro4

	<b>Docked molecules (all) (%)</b>	<b>Docked molecules (score &lt;= 5.0) (%)</b>
Inhibitors	96	86
Non-inhibitors	92	78

Although inhibitors show better results, there is rarely a distinction between the two classes. These results raise the question whether Maestro4 truly is a binding pocket of BSEP inhibitors or whether it is just a promiscuous binding pocket where diverse compounds can be docked obtaining high docking scores.

One reason for the insufficient separation might be that compounds of the so-called middle class blur the line between inhibitors and non-inhibitors. It can be expected that an inhibitor with an IC<sub>50</sub> value of 9.0  $\mu$ M and a non-inhibitor with an IC<sub>50</sub> value of 20.0  $\mu$ M do not show significant differences in docking studies. To address this issue, one can exclude activity values that show weak activity to improve separation between true inhibitors and non-inhibitors. This approach was tried using only inhibitors with activity values smaller or equal to 7.0  $\mu$ M and non-inhibitors with values equal or higher than 700.0  $\mu$ M. 69 molecules were included, out of which 29 were inhibitors and 40 non-inhibitors.

Table 11: Docking results of data set without middle class in Maestro4

	<b>Docked molecules (all) (%)</b>	<b>Docked molecules (score &lt;= 5.0) (%)</b>
Inhibitors	91	35
Non-inhibitors	95	46

The results show that excluding the middle class shifts higher docking scores even more towards non-inhibitors than inhibitors. However, in docking studies not only scores must be analyzed in detail, also docking poses, together with literature searches, give significant information on the likeliness of a certain binding mode. The structurally diverse sets made it difficult to evaluate and compare different binding modes, which is why the approach was switched to analyzing one specific subgroup of the data set. Due to the



structural relation of the sulfonyl group and the sulfonamide moiety and the high abundance of sulfonamides in the pharmacophore screening, this subset was chosen for further investigations of the binding mode.

#### 4.2.3.3. Sulfonamide docking studies

44 sulfonamide containing inhibitors were extracted from the data set using SMARTS code as stated in 3.1.1.4.

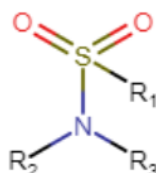


Figure 15: Structural depiction of SMARTS pattern used to filter data set;  $R_1 = R_2 = R_3 = H$ ,  $CR_3$

Since hydrogen bond acceptors have been implicated in increased inhibitory activity, and the natural substrates contain a hydrophilic head group, it was postulated that the sulfonamide group most likely interacts with the protein through hydrogen bonding. Consequently, the binding site was rechecked for potential hydrogen binding spots.

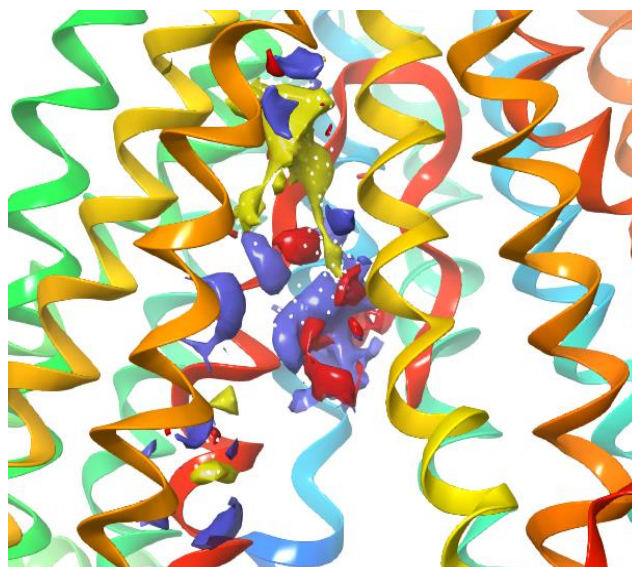


Figure 16: Beneficial ligand features in binding pocket Maestro4 ; HBD in blue, HBA in red, hydrophobic/aromatic in yellow

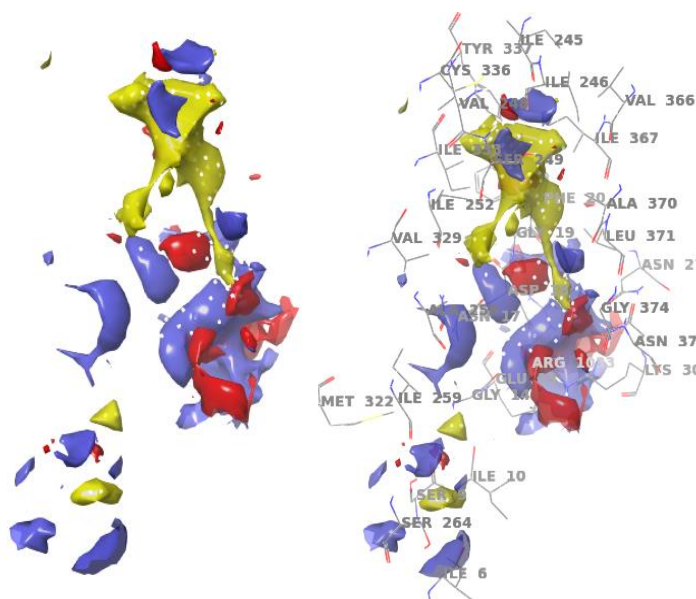


Figure 17: Beneficial ligand features in binding pocket Maestro4 ; HBD in blue, HBA in red, hydrophobic/aromatic in yellow; (right) with residues

Looking at the surface, it becomes apparent, that there is one major spot for hydrogen bonding, including residues Asn17 (HBA), Asp18 (HBA), Lys30 (HBD) and Asn375 (HBD). For a more detailed analysis of possible binding modes, induced-fit docking for all sulfonamide inhibitors was conducted. 844 poses were obtained using the standard protocol. SIFts were calculated based on hydrogen bond acceptors and donors.

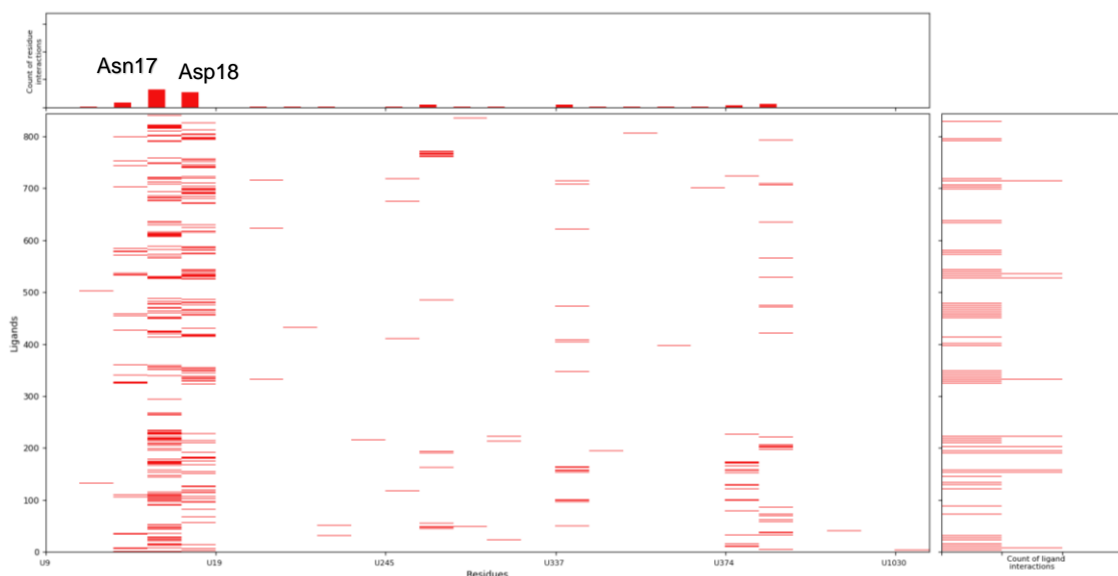


Figure 18: Structural interaction fingerprints calculated based on hydrogen bond acceptor properties of receptor with sulfonamide moiety containing inhibitors

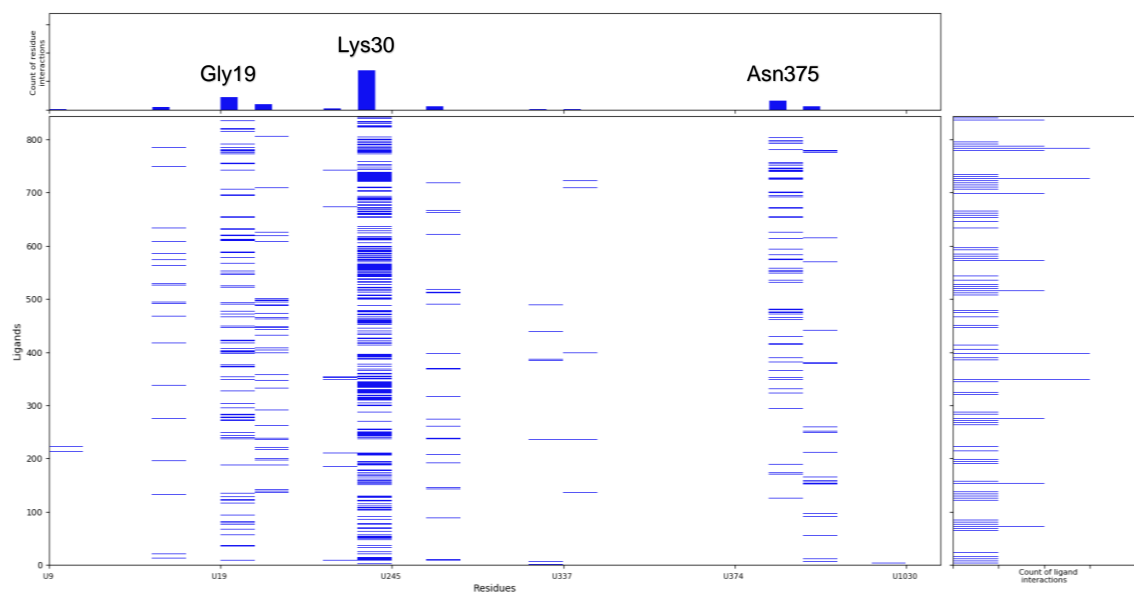


Figure 19: Structural interaction fingerprints calculated based on hydrogen bond donor properties of receptor with sulfonamide moiety containing inhibitors

The hydrogen bond acceptor plot shows a high abundance of ligand poses interacting with residues Asn17 and Asp18. For hydrogen bond donor properties, Lys30 shows the highest interaction frequency with inhibitors. Since sulfonamides contain hydrogen bond acceptor features (sulfonyl oxygens) next to hydrogen bond donor features (H-substituted nitrogen) in case of primary and secondary sulfonamides, it would make sense that the molecules bind in a hydrogen bond acceptor and donor rich area. The observed important residues in Figure 18 and Figure 19 span a sub pocket with a favorable environment for hydrophilic groups such as sulfonamides.

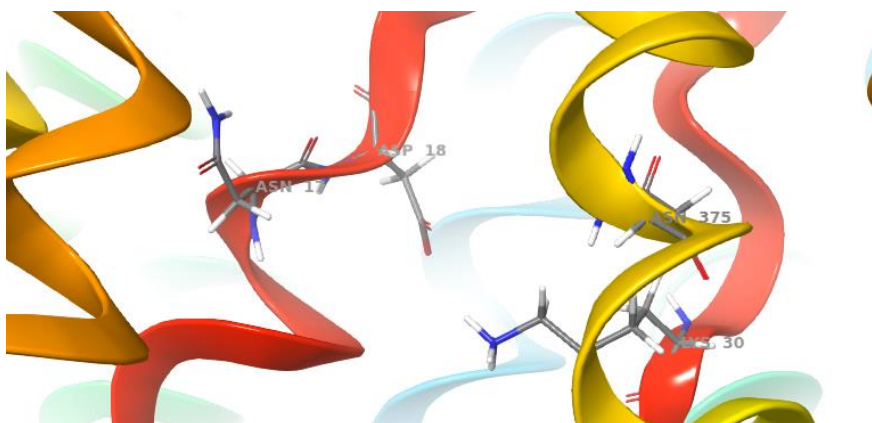


Figure 20: Hydrophilic subpocket spanned by Asn17, Asp18, Lys30 and Asn375

These findings suggest that the sulfonamide group interacts in this hydrophilic sub pocket, with the rest of the molecule extending into the hydrophobic/aromatic cavity of the calculated pocket. This binding mode could indeed be observed for several poses of inhibitors.

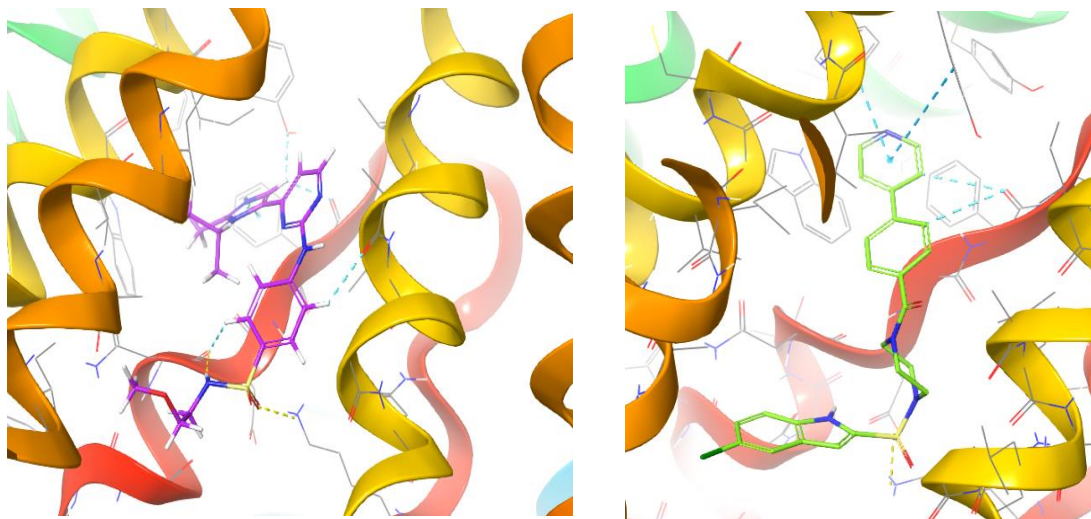


Figure 21: 2 inhibitors bound to Maestro4 showing interactions between the sulfonamide moiety and Lys30

To support the proposed binding mode, the natural substrate taurocholate was docked to Maestro4 using IFD.

#### 4.2.3.4. Taurocholate docking studies

The highest ranked pose of TC shows the same hydrogen bonding pattern as observed in the sulfonamide docking study. The deprotonated oxygen of the sulfonic acid interacts with Lys30 via HB and forms a salt bridge to the nitrogen of Lys30. The amido nitrogen of TC acts as hydrogen bond donor of Asn17. The steroid core of the molecule is enclosed by the hydrophobic residues of the pocket, including Phe20, Leu332, Ile333, Cys336, Tyr337, Val366, Ile367, Ala370 and Leu371.

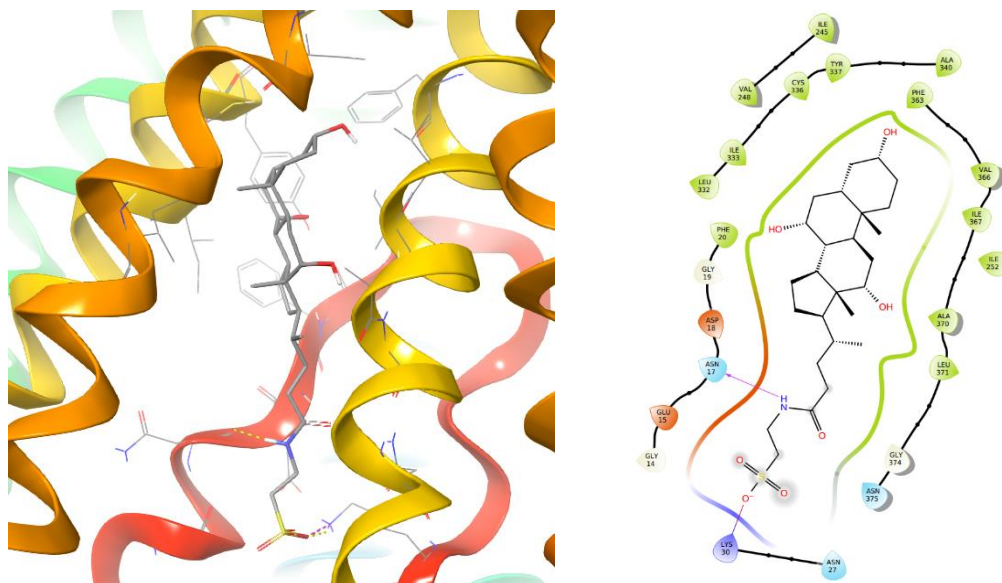


Figure 22: Docking pose of natural substrate taurocholate in binding pocket Maestro4; HBD residues shown in dark blue, HBA residues shown in red, polar residues shown in light blue, neutral residues shown in white, hydrophobic residues shown in green

#### 4.2.3.5. Interaction fingerprint clustering

Interaction fingerprints were used to filter and cluster poses, for a better overview on different binding modes. Since Lys30 appears to be the most important residue for hydrogen bonding interactions, the poses were filtered for hydrogen bonding with Lys30 using SIFt.

Table 12: Mean values of scores of poses dependent on interaction with Lys30

SIFt Lys30	Poses	IFD Score	Prime Energy	Glide Emodel
yes	360	-1828.0 $\pm$ 5.2	-36407.4 $\pm$ 96.4	-76.3 $\pm$ 9.5
no	337	-1829.2 $\pm$ 5.5	-36396.8 $\pm$ 105.1	-75.3 $\pm$ 9.5

As apparent in Table 12, 360 poses show hydrogen bonding with Lys30, whereas 337 do not show interactions with this residue. For both clusters, the scoring values are similar. However, 16 out of 18 proposed poses of TC show hydrogen bonding with Lys30. Therefore, it was decided to move forward with this cluster.

The poses were further clustered using SIFTs based on backbone and sidechain interactions. 23 clusters were obtained. Only clusters with at least 20 poses were analyzed in more detail.

Table 13: Mean scores of poses from different clusters grouped by backbone and sidechain interactions

Cluster	Poses	IFD Score	Prime Energy	Glide Emodel
2	53	-1825.2 ± 2.0	-36337.7 ± 236.8	-75.0 ± 6.7
5	29	-1825.8 ± 3.2	-36366.6 ± 60.3	-73.1 ± 5.0
6	43	-1827.7 ± 6.5	-36394.7 ± 125.2	-78.7 ± 7.8
18	22	-1828.6 ± 5.1	-36424.1 ± 95.9	-71.8 ± 10.6
19	145	-1828.9 ± 5.0	-36423.5 ± 91.7	-76.9 ± 10.2

Overall, clusters 6, 18 and 19 show the highest scores. Group 6 contains 14 inhibitors and no pose of TC. Cluster 18 contains only 10 inhibitors and no pose of the natural substrate as well. Number 19 includes 36 inhibitors and 14 poses of TC. Therefore, this cluster summarizes the most shared features among all inhibitors and the substrate TC and will be used for further structural investigations. It was tried to reduce the cluster size, however this was not possible. Nevertheless, the poses were analyzed looking into the interaction profile of the sulfonamide group. The poses were categorized into the following groups:

- Sulfonyl oxygen forms HB with Lys30 (3 stars)
- Sulfonyl oxygen forms HB with other residue (1 star)
- Sulfonyl oxygen does not interact through HB (no star)

Ranking of the poses with first priority IFD score and second priority Glide Emodel score, resulted in the following order (complete figure Appendix, Figure 36):

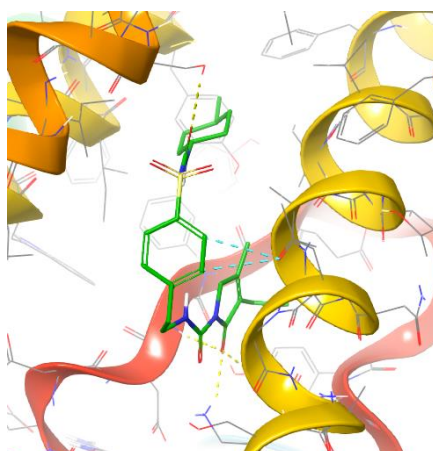
Title	Stars	IFDScore	Prime Energy	glide emodel
Cluster 19 (145)				
ligprep_3.sdf:35	☆☆☆	-1845.2	-36729.8	-89.1
ligprep_3.sdf:36	☆☆☆☆	-1844.0	-36695.5	-95.7
ligprep_3.sdf:36	☆☆☆☆	-1841.8	-36683.5	-76.4
ligprep_3.sdf:24	☆☆☆	-1841.4	-36647.0	-91.7
ligprep_3.sdf:24	☆☆☆	-1840.3	-36647.5	-76.5
ligprep_3.sdf:24	☆☆☆☆	-1839.7	-36628.8	-82.6
ligprep_3.sdf:24	☆☆☆	-1838.7	-36646.1	-65.6
ligprep_3.sdf:34	☆☆☆	-1837.2	-36592.5	-80.4
ligprep_3.sdf:20	☆☆☆	-1837.1	-36554.1	-87.7
ligprep_3.sdf:20	☆☆☆	-1836.6	-36551.9	-82.5
ligprep_3.sdf:28	☆☆☆	-1836.4	-36560.7	-84.4
ligprep_3.sdf:28	☆☆☆	-1836.3	-36571.3	-74.5
ligprep_3.sdf:20	☆☆☆	-1836.2	-36550.0	-80.2
ligprep_3.sdf:20	☆☆☆	-1836.2	-36543.8	-84.4
ligprep_3.sdf:28	☆☆☆	-1836.0	-36564.5	-83.9
ligprep_3.sdf:28	☆☆☆	-1835.8	-36572.5	-74.9
ligprep_3.sdf:46	☆☆☆	-1835.7	-36542.7	-83.3
ligprep_3.sdf:46	☆☆☆	-1835.7	-36539.3	-79.0
ligprep_3.sdf:24	☆☆☆	-1835.3	-36532.1	-93.3
ligprep_3.sdf:28	☆☆☆	-1835.2	-36553.9	-75.3
ligprep_3.sdf:22	☆☆☆	-1835.2	-36548.7	-77.7
ligprep_3.sdf:20	☆☆☆	-1835.0	-36548.2	-68.9
ligprep_3.sdf:23	☆☆☆	-1835.0	-36552.1	-74.2
ligprep_3.sdf:24	☆☆☆	-1835.0	-36540.6	-79.4
ligprep_3.sdf:23	☆☆☆☆	-1834.9	-36514.0	-90.4
ligprep_3.sdf:24	☆☆☆	-1834.7	-36524.9	-90.1
ligprep_3.sdf:24	☆☆☆	-1834.7	-36536.5	-79.1
ligprep_3.sdf:28	☆☆☆	-1834.4	-36556.2	-69.6
ligprep_3.sdf:23	☆☆☆☆	-1834.1	-36502.6	-87.7
ligprep_3.sdf:24	☆☆☆	-1834.0	-36532.6	-71.1
ligprep_3.sdf:42	☆☆☆☆	-1834.0	-36458.3	-116.0

Figure 23: Ranking of poses in cluster 19 according to 1) IFD score and 2) glide emodel score

It can be observed that interaction of the sulfonyl oxygen with Lys30 is abundant among the top ranked poses. However, not for every inhibitor this pose is top ranked or actually



exists. This is true specifically in cases, where another hydrogen bond acceptor is available terminally of the molecule or the sulfonamide moiety is sterically hindered in interacting with Lys30.



*Figure 24: Example of different binding mode for inhibitor containing several hydrogen bond acceptor features*

This could also explain the inactivity of some promising looking non-inhibitors.

Apart from the hydrophilic part of the pocket, there seem to be several modes for occupying the hydrophobic/aromatic cavity, which explains the numerous numbers of poses. In the poses of cluster 19, hydrophobic and aromatic parts of the molecule are primary interacting with residues such as Ile245, Val248, Ile367 and Ala370. However, looking at the poses in cluster 6, a different mode is observed for inhibitors containing aromatic features. In this case, the aromatic rings interact with residues such as Phe20, Phe334, Trp330 and Tyr337. It seems logical, that aromatic substructures interact with the aromatic part of the cavity, whereas hydrophobic parts, e.g. the steroid core of TC, rather interact with apolar residues. Figure 25 and Figure 26 show the two different modes for one inhibitor.

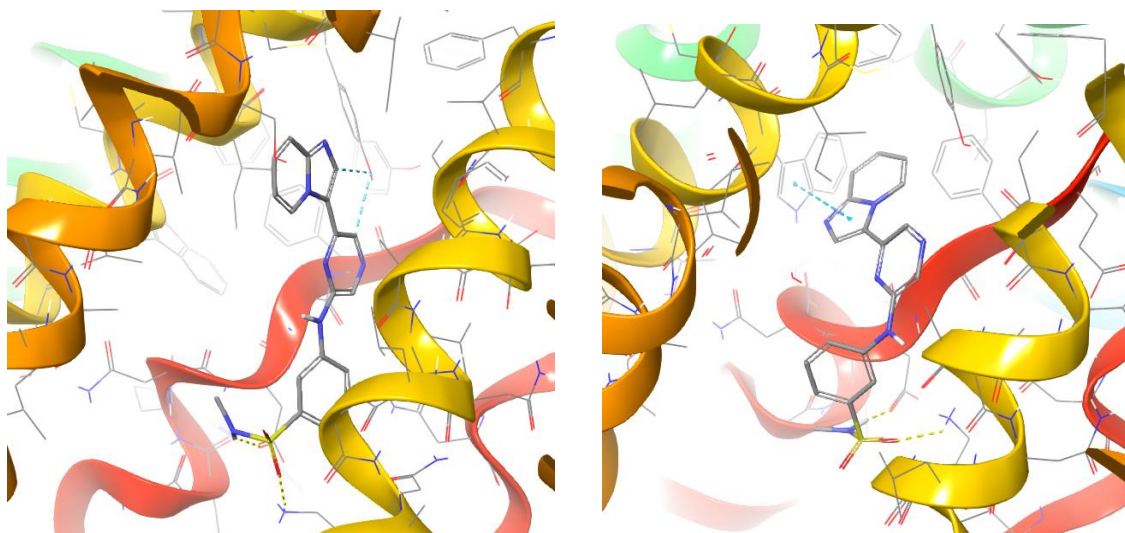


Figure 25: left: top ranked pose of structure 2 in cluster 19 showing hydrogen bonding to Asp18 and Lys30 and aromatic hydrogen bonding with Tyr337; right: top ranked pose of structure 2 in cluster 6 showing hydrogen bonding to Lys30 and Asp18 and pi-stacking with Trp330

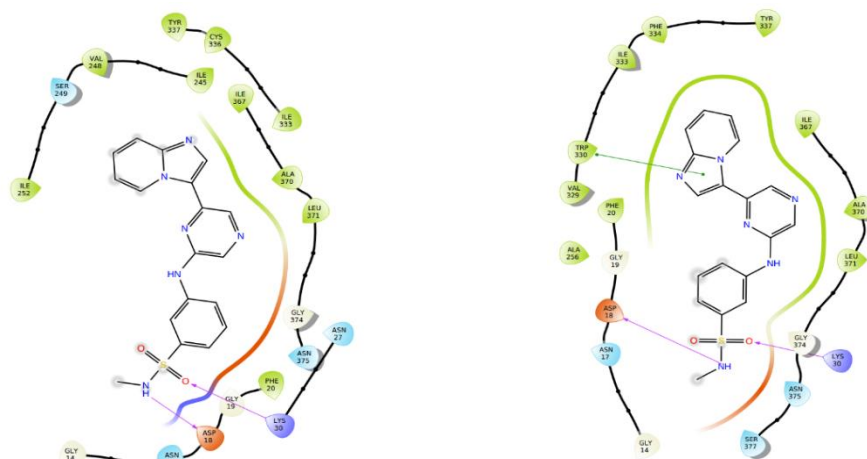


Figure 26: Ligand interaction diagrams of structure 2 in cluster 19 (left) and cluster 2 (right); HBD residues shown in dark blue, HBA residues shown in red, polar residues shown in light blue, neutral residues shown in white, hydrophobic residues show in green

However, it was found that these two modes are observed for quite similar molecules, or identical molecules with different ionized states, which makes the clustering quite biased. In order to support the proposed binding mode, docking a more structurally diverse set with concrete bioactivity values would be important.

#### 4.2.3.6. Literature in support of proposed binding mode

Uniprot collects different mutations of the transporter reported in literature. Several mutations are described, that are known to decrease transport capacity of taurocholate. However, in most cases these mutations cause a decrease in protein expression, hinders apical membrane localization or leads to a misfolding of the transporter, thus giving no information on the binding area of the substrates. Two relevant mutations were observed



for residues 336 and 337 in progressive familial intrahepatic cholestasis 2, however it is not clear whether they are involved in taurocholate binding.<sup>50,51</sup>

No structure-based analysis of the bile salt export pump was done since the structure was released. However, one structural analysis was performed on a homology model of BSEP by Jain et. al. The approach focused on structure-based classification of inhibitors and non-inhibitors. The docking results presented in the work show the same area of the binding site, however, interacting residues differ. The authors state that Phe334, Leu364, Tyr772 and Phe776 might be important for inhibitory activity. However, since no substrate was docked, it is difficult to compare the results with each other.<sup>52</sup> No other docking studies on the BSEP transporter are known of.

Comparing the binding pockets of the closest related transporter MDR1, it can be observed that in the case of tariquidar (PDB: 7A6E), zosuquidar (PDB: 7A6F) and elacridar (7A6C), two inhibitors are bound at the same time. Since these structures can be compared to the sulfonamides and taurocholate concerning size, it might be possible that not one but two inhibitor molecules at once inhibit the transporter. However, addressing this question in basic docking studies might be a difficult task. Additionally, it must be kept in mind, that P-glycoprotein is known for its diverse substrates and inhibitors<sup>53</sup> indicating a large, promiscuous binding site, whereas BSEP is known to be more selective towards interacting ligands.<sup>54</sup>

## 5. Conclusion and Outlook

In summary, the work presented in this master thesis shows insights into properties correlated with BSEP inhibition and gives a suggestion for an orthosteric binding site and ligand-protein interactions with the transporter.

The data-based approach demonstrated the importance of lipophilicity and molecular weight, which is supported by several literature sources.<sup>47,48,49</sup> Additional features, that were found to be important, were molar refractivity, rotatable bonds and hydrogen bond acceptors. No concrete matched molecular pairs were found in the data set. Fingerprint clustering resulted in the finding of abundant functional groups in the data set, including sulfonamides, carboxylic acids, piperazines and steroid-based molecules.

The structure-based approach resulted in a possible binding site of the substrate taurocholate and a subset of inhibitors. Overall, six binding pockets were analyzed in more detail, using structure-based pharmacophores and docking studies. The proposed binding site is located in the center of the transporter and has a size of 135 amino acids. The pocket can be categorized into a hydrophilic cavity and a hydrophobic/aromatic part. When thinking of the natural substrates, this character reflects the amphiphilic structures of bile acids such as taurocholate. Upon induced-fit docking studies of taurocholate and a subset of inhibitors, containing a sulfonamide moiety, it was found that one specific interaction is conserved among nearly all inhibitors and the substrate, being a hydrogen bond with residue Lys30. It was further observed that, depending on the nature of the rest of the molecule, different binding modes are possible. Hydrophobic core structures, such as steroids, tended to interact with hydrophobic residues such as Ile245, Val248, Ile367, Ala370 and Leu371. Aromatic rings rather showed interactions with residues Phe20, Phe334, Trp330 and Tyr337. However, this trend was only observed for a small portion of inhibitors and needs to be further investigated.

The results presented in this thesis could be used as starting point for further structural investigations. Since it is known that several carboxylic acid derived structures inhibit the transporter, it would be interesting to see whether this moiety also interacts with Lys30 in docking studies. Together with that, the natural substrate glycocholate could be docked, which also contains a carboxylic acid moiety.

Furthermore, structure-based pharmacophores could be created of the proposed binding mode to gather further insight on the plausibility of the binding pocket. For a higher reliability on the proposed results, the missing residues of the transporter should be modelled, to avoid overlooking any steric clashes that might occur.

In general, more concrete bioactivity data on the transporter would be beneficial for further QSAR or structural analyses. Especially a congeneric series of compounds could help elucidate important structural features for inhibitory activity. For the purposes of validating the proposed binding mode, mutation studies specifically altering the involved residues would be of great benefit. Lastly, it must be said that only high resolution co-crystallized structures can provide the desired information on the location and interactions of ligands with a high certainty. Nevertheless, since elucidating the structures of membrane-bound proteins has been quite difficult to this day, homology modeling and working on apo structures are valuable tools for addressing structure-based issues.

Overall, the thesis provides a systematic workflow on how to combine data science and structure-based approaches for mapping structural differences with changes in activity, not only from a ligand's point of view but also considering the structure of the protein. In this case, the approach did not work out as planned, however the described workflow can be used for any target protein and data set.

## 6. References

1. Beis K. Structural basis for the mechanism of ABC transporters. *Biochem Soc Trans.* 2015;43. doi:10.1042/BST20150047
2. Robey RW, Pluchino KM, Hall MD, Fojo AT, Bates SE, Gottesman MM. Revisiting the role of ABC transporters in multidrug-resistant cancer. *Nat Rev Cancer.* 2018;18(7). doi:10.1038/s41568-018-0005-8
3. Dean M, Hamon Y, Chimini G. The human ATP-binding cassette (ABC) transporter superfamily. *J Lipid Res.* 2001;42(7). doi:10.1016/s0022-2275(20)31588-1
4. Linton KJ, Higgins CF. The Escherichia coli ATP-binding cassette (ABC) proteins. *Mol Microbiol.* 1998;28(1). doi:10.1046/j.1365-2958.1998.00764.x
5. Linton KJ. Structure and function of ABC transporters. *Physiology.* 2007;22(2). doi:10.1152/physiol.00046.2006
6. Locher KP. Mechanistic diversity in ATP-binding cassette (ABC) transporters. *Nat Struct Mol Biol.* 2016;23(6). doi:10.1038/nsmb.3216
7. Schinkel AH, Jonker JW. Mammalian drug efflux transporters of the ATP binding cassette (ABC) family: An overview. *Adv Drug Deliv Rev.* 2003;55(1):3-29. doi:10.1016/S0169-409X(02)00169-2
8. Li M, Cai SY, Boyer JL. Mechanisms of bile acid mediated inflammation in the liver. *Mol Aspects Med.* 2017;56. doi:10.1016/j.mam.2017.06.001
9. Chen M jun, Liu C, Wan Y, et al. Enterohepatic circulation of bile acids and their emerging roles on glucolipid metabolism. *Steroids.* 2021;165. doi:10.1016/j.steroids.2020.108757
10. Di Ciaula A, Garruti G, Baccetto RL, et al. Bile acid physiology. *Ann Hepatol.* 2017;16. doi:10.5604/01.3001.0010.5493
11. Lefebvre P, Cariou B, Lien F, Kuipers F, Staels B. Role of bile acids and bile acid receptors in metabolic regulation. *Physiol Rev.* 2009;89(1). doi:10.1152/physrev.00010.2008
12. Garzel B, Zhang L, Huang S-M, Wang H. A Change in Bile Flow: Looking Beyond Transporter Inhibition in the Development of Drug-induced Cholestasis. *Curr Drug Metab.* 2019;20(8). doi:10.2174/1389200220666190709170256
13. Strautnieks SS, Kagalwalla AF, Tanner MS, et al. Identification of a locus for progressive familial intrahepatic cholestasis PFIC2 on chromosome 2q24. *Am J Hum Genet.* 1997;61(3). doi:10.1086/515501
14. Sohail MI, Dönmez-Cakil Y, Szöllősi D, Stockner T, Chiba P. The bile salt export pump: Molecular structure, study models and small-molecule drugs for the treatment of inherited bsep deficiencies. *Int J Mol Sci.* 2021;22(2). doi:10.3390/ijms22020784
15. Wang L, Hou WT, Chen L, et al. Cryo-EM structure of human bile salts exporter ABCB11. *Cell Res.* 2020;30(7). doi:10.1038/s41422-020-0302-0
16. Stieger B. Recent insights into the function and regulation of the bile salt export pump (ABCB11). *Curr Opin Lipidol.* 2009;20(3). doi:10.1097/MOL.0b013e32832b677c
17. Hayashi H, Takada T, Suzuki H, Onuki R, Hofmann AF, Sugiyama Y. Transport

- by vesicles of glycine- and taurine-conjugated bile salts and tauroolithocholate 3-sulfate: A comparison of human BSEP with rat Bsep. *Biochim Biophys Acta - Mol Cell Biol Lipids*. 2005;1738(1-3). doi:10.1016/j.bbalip.2005.10.006
18. Ogimura E, Sekine S, Horie T. Bile salt export pump inhibitors are associated with bile acid-dependent drug-induced toxicity in sandwich-cultured hepatocytes. *Biochem Biophys Res Commun*. 2011;416(3-4). doi:10.1016/j.bbrc.2011.11.032
  19. Kubitz R, Dröge C, Stindt J, Weissenberger K, Häussinger D. The bile salt export pump (BSEP) in health and disease. *Clin Res Hepatol Gastroenterol*. 2012;36(6). doi:10.1016/j.clinre.2012.06.006
  20. Kullak-Ublick GA, Andrade RJ, Merz M, et al. Drug-induced liver injury: Recent advances in diagnosis and risk assessment. *Gut*. 2017;66(6). doi:10.1136/gutjnl-2016-313369
  21. Watkins PB. Drug safety sciences and the bottleneck in drug development. *Clin Pharmacol Ther*. 2011;89(6). doi:10.1038/clpt.2011.63
  22. Mosedale M, Watkins PB. Drug-induced liver injury: Advances in mechanistic understanding that will inform risk management. *Clin Pharmacol Ther*. 2017;101(4). doi:10.1002/cpt.564
  23. Russmann S, Jetter A, Kullak-Ublick GA. Pharmacogenetics of drug-induced liver injury. *Hepatology*. 2010;52(2). doi:10.1002/hep.23720
  24. Critical Path Institute's Predictive Safety Testing Consortium. Current trends in BSEP inhibition and perturbation to bile acid homeostasis as mechanisms of drug-induced liver injury. Published 2016. <https://c-path.org/current-trends-in-bsep-inhibition-and-perturbation-to-bile-acid-homeostasis-as-mechanisms-of-drug-induced-liver-injury/>
  25. Tiwari A, Sekhar AKT. Workflow based framework for life science informatics. *Comput Biol Chem*. 2007;31(5-6). doi:10.1016/j.compbiolchem.2007.08.009
  26. Hemmerich J. {KNIME} Structure Standardisation Workflow. Published January 31, 2020. Accessed March 23, 2021. <https://github.com/PharminfoVienna/Chemical-Structure-Standardisation>
  27. Todeschini R, Consonni V. *Molecular Descriptors for Chemoinformatics*. Vol 2.; 2010. doi:10.1002/9783527628766
  28. RDKit Descriptor Calculation. <https://kni.me/n/6Fdt1ozl1QcUu1BX>
  29. Leach AG, Jones HD, Cosgrove DA, et al. Matched molecular pairs as a guide in the optimization of pharmaceutical properties; a study of aqueous solubility, plasma protein binding and oral exposure. *J Med Chem*. 2006;49(23). doi:10.1021/jm0605233
  30. Griffen E, Leach AG, Robb GR, Warner DJ. Matched molecular pairs as a medicinal chemistry tool. *J Med Chem*. 2011;54(22). doi:10.1021/jm200452d
  31. Nisius B, Bajorath J. Reduction and recombination of fingerprints of different design increase compound recall and the structural diversity of hits. *Chem Biol Drug Des*. 2010;75(2). doi:10.1111/j.1747-0285.2009.00930.x
  32. Fernández-De Gortari E, García-Jacas CR, Martínez-Mayorga K, Medina-Franco JL. Database fingerprint (DFP): an approach to represent molecular databases. *J Cheminform*. 2017;9(1). doi:10.1186/s13321-017-0195-1
  33. Madhavi Sastry G, Adzhigirey M, Day T, Annabhimoju R, Sherman W. Protein

and ligand preparation: Parameters, protocols, and influence on virtual screening enrichments. *J Comput Aided Mol Des*. 2013;27(3). doi:10.1007/s10822-013-9644-8

34. Schmidtke P, Souaille C, Estienne F, Baurin N, Kroemer RT. Large-scale comparison of four binding site detection algorithms. *J Chem Inf Model*. 2010;50(12). doi:10.1021/ci1000289
35. Yang SY. Pharmacophore modeling and applications in drug discovery: Challenges and recent advances. *Drug Discov Today*. 2010;15(11-12). doi:10.1016/j.drudis.2010.03.013
36. Dixon SL, Smondyrev AM, Knoll EH, Rao SN, Shaw DE, Friesner RA. PHASE: A new engine for pharmacophore perception, 3D QSAR model development, and 3D database screening: 1. Methodology and preliminary results. *J Comput Aided Mol Des*. 2006;20(10-11). doi:10.1007/s10822-006-9087-6
37. Loving K, Salam NK, Sherman W. Energetic analysis of fragment docking and application to structure-based pharmacophore hypothesis generation. *J Comput Aided Mol Des*. 2009;23(8). doi:10.1007/s10822-009-9268-1
38. Dixon SL, Smondyrev AM, Rao SN. PHASE: A novel approach to pharmacophore modeling and 3D database searching. *Chem Biol Drug Des*. 2006;67(5). doi:10.1111/j.1747-0285.2006.00384.x
39. Ferreira LG, Dos Santos RN, Oliva G, Andricopulo AD. Molecular docking and structure-based drug design strategies. *Molecules*. 2015;20(7). doi:10.3390/molecules200713384
40. Hung CL, Chen CC. Computational approaches for drug discovery. *Drug Dev Res*. 2014;75(6). doi:10.1002/ddr.21222
41. Friesner RA, Banks JL, Murphy RB, et al. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J Med Chem*. 2004;47(7). doi:10.1021/jm0306430
42. Sherman W, Beard HS, Farid R. Use of an induced fit receptor structure in virtual screening. *Chem Biol Drug Des*. 2006;67(1). doi:10.1111/j.1747-0285.2005.00327.x
43. Sherman W, Day T, Jacobson MP, Friesner RA, Farid R. Novel procedure for modeling ligand/receptor induced fit effects. *J Med Chem*. 2006;49(2). doi:10.1021/jm050540c
44. Deng Z, Chuaqui C, Singh J. Structural Interaction Fingerprint (SIFt): A Novel Method for Analyzing Three-Dimensional Protein-Ligand Binding Interactions. *J Med Chem*. 2004;47(2). doi:10.1021/jm030331x
45. Singh J, Deng Z, Narale G, Chuaqui C. Structural interaction fingerprints: A new approach to organizing, mining, analyzing, and designing protein-small molecule complexes. *Chem Biol Drug Des*. 2006;67(1). doi:10.1111/j.1747-0285.2005.00323.x
46. Morgan RE, van Staden CJ, Chen Y, et al. A multifactorial approach to hepatobiliary transporter assessment enables improved therapeutic compound development. *Toxicol Sci*. 2013;136(1). doi:10.1093/toxsci/kft176
47. Ritschel T, Hermans SMA, Schreurs M, et al. In silico identification and in vitro validation of potential cholestatic compounds through 3D ligand-based pharmacophore modeling of BSEP inhibitors. *Chem Res Toxicol*. 2014;27(5).

doi:10.1021/tx5000393

48. Pedersen JM, Matsson P, Bergström CAS, et al. Early identification of clinically relevant drug interactions with the human bile salt export pump (BSEP/ABCB11). *Toxicol Sci.* 2013;136(2). doi:10.1093/toxsci/kft197
49. Welch MA, Köck K, Urban TJ, Brouwer KLR, Swaan PW. Toward predicting drug-induced liver injury: Parallel computational approaches to identify multidrug resistance protein 4 and bile salt export pump inhibitors. *Drug Metab Dispos.* 2015;43(5). doi:10.1124/dmd.114.062539
50. Jansen PLM, Strautnieks SS, Jacquemin E, et al. Hepatocanalicular bile salt export pump deficiency in patients with progressive familial intrahepatic cholestasis. *Gastroenterology.* 1999;117(6). doi:10.1016/S0016-5085(99)70287-8
51. Hu G, He P, Liu Z, Chen Q, Zheng B, Zhang Q. Diagnosis of ABCB11 gene mutations in children with intrahepatic cholestasis using high resolution melting analysis and direct sequencing. *Mol Med Rep.* 2014;10(3). doi:10.3892/mmr.2014.2349
52. Jain S, Grandits M, Richter L, Ecker GF. Structure based classification for bile salt export pump (BSEP) inhibitors using comparative structural modeling of human BSEP. *J Comput Aided Mol Des.* 2017;31(6). doi:10.1007/s10822-017-0021-x
53. Wang RB, Kuo CL, Lien LL, Lien EJ. Structure-activity relationship: Analyses of p-glycoprotein substrates and inhibitors. *J Clin Pharm Ther.* 2003;28(3). doi:10.1046/j.1365-2710.2003.00487.x
54. Wang EJ, Casciano CN, Clement RP, Johnson WW. Fluorescent substrates of sister-P-glycoprotein (BSEP) evaluated as markers of active transport and inhibition: Evidence for contingent unequal binding sites. *Pharm Res.* 2003;20(4). doi:10.1023/A:1023278211849

## 7. Appendix

### 7.1. Supplemental material

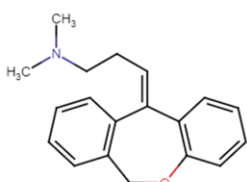
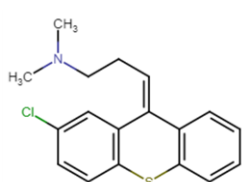
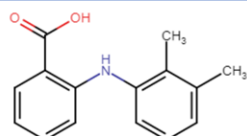
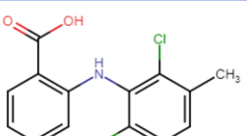
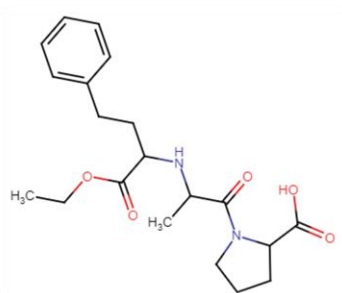
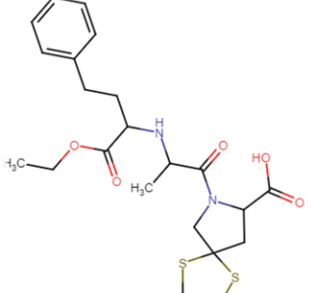
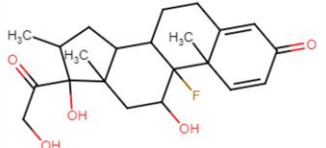
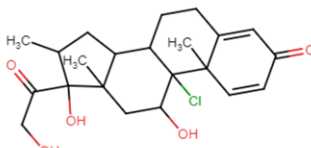
#### 7.1.1. Data-based approach

##### 7.1.1.1. TGD-based pairs

Table 14: Descriptor values of found pairs by TGD-based fingerprint clustering

Molecule	IC50 [μM]	SlogP	AMW [g/mol]	SMR	TPSA	NRotB	HBA	HBD
3	> 133.0	4.0	279.4	87.5	12.5	3	2	0
4	27.5	5.2	315.9	92.3	3.2	3	2	0
5	> 133.0	3.7	241.3	72.6	49.3	3	2	2
6	27.2	4.7	296.2	77.9	49.3	3	2	2
7	> 133.0	1.6	376.5	100.2	95.9	9	5	2
8	47.1	2.4	466.6	123.2	95.9	9	7	2
9	>133.0	1.9	392.5	99.9	94.8	2	5	3
10	23.7	2.2	408.9	104.7	94.8	2	5	3

Table 15: Structures of found pairs by TGD-based fingerprint clustering

Pair		
3 → 4		
5 → 6		
7 → 8		
9 → 10		

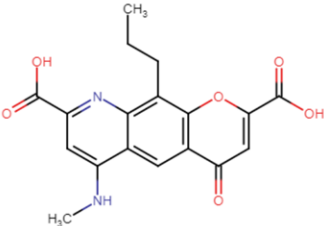
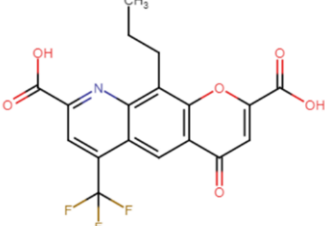
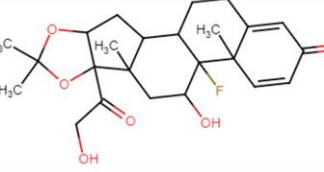
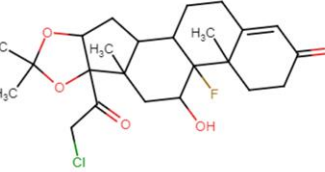


### 7.1.1.2. GpiDAPH3-based pairs

Table 16: Descriptor values of found pairs by GpiDAPH3-based fingerprint clustering

Mol.	IC50 [ $\mu$ M]	SlogP	AMW [g/mol]	SMR	TPSA	NRotB	HBA	HBD
11	> 1000.0	2.7	356.3	95.2	129.7	5	6	3
12	129.7	3.7	364.7	395.3	117.7	4	5	2
13	> 133.0	2.4	434.5	108.6	93.1	2	6	2
14	< 10.0	3.9	455.0	112.3	72.8	2	5	1

Table 17: Structures of found pairs by GpiDAPH3-based fingerprint clustering

Pair		
11 $\rightarrow$ 12		
13 $\rightarrow$ 14		

### 7.1.1.3. MACCS-based pairs

Table 18: Descriptor values of found pairs by MACCS-based fingerprint clustering

Molecule	IC50 [ $\mu$ M]	SlogP	AMW [g/mol]	SMR	TPSA	NRotB	HBA	HBD
15	> 133.0	1.6	337.4	81.7	99.6	2	6	2
16	7.6	2.3	371.8	86.7	99.6	2	6	2
17	> 133.0	1.9	392.5	99.9	94.8	2	5	3
18	16.4	2.8	376.5	98.5	74.6	2	4	2







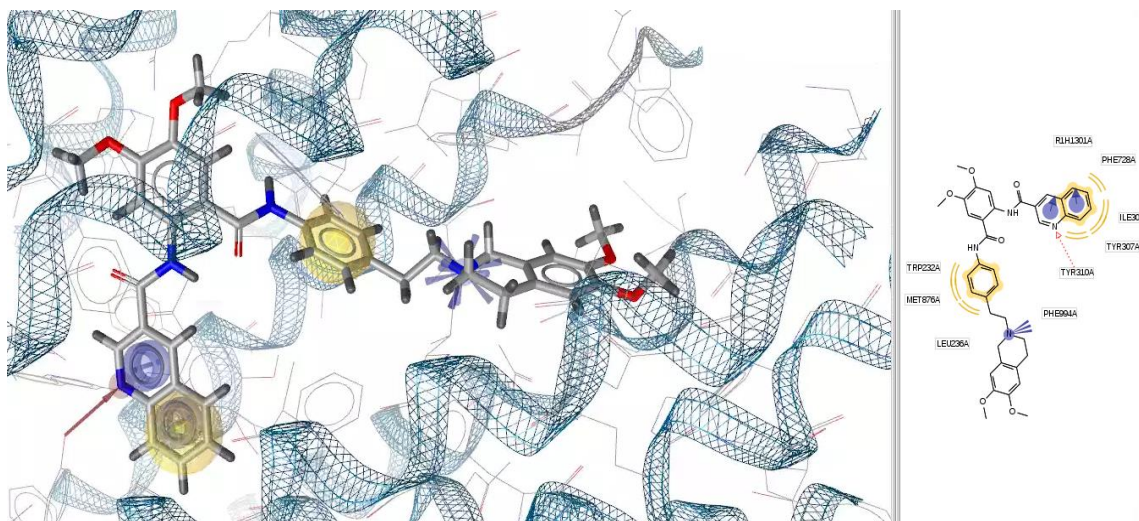


Figure 30: Binding site 2 of tariquidar in P-gp with interacting residues, pi-stacking as purple circle, hydrophobic interactions in yellow, positive ionizable in blue

#### 7.1.2.2. Residues of calculated binding pockets

Table 20: Residues surrounding binding pockets calculated by SiteFinder (MOE) and SiteMap (Maestro)

Pocket	Pocket surrounding residues (3Å)
MOE1	Glu21, Lys24, Ser25, Tyr26, Asn27, Asn28, Asp29, Lys30, Lys31, His72, Gln76, Ala167, Ala168, Ile171, Gln172, Arg175, Asn207, Asn210, Asp211, Ala214, Asp215, Gln216, Leu219, Gln222, Arg223, Ser226, Val368, Leu371, Asn372, Asn375, Pro378, Gln918, Leu922, Phe925, Ala926, Asp929, Lys930, Leu933, Leu967, Pro970, Phe971, Thr973, Ala974, Lys977, Tyr981, Phe985, Gln989
MOE2	Leu7, Arg8, Lys11, Lys12, Phe13, Glu15, Glu16, Asp18, Gly19, Phe20, Tyr26, Trp330, Asn765, Gly766, Val768, Thr769, Gln813, Gln816, Phe820, Gly870, Gly873, Ser874, Gln875, Gly877, Met878, Asn881, Ser882, Asn885, Leu1026, Thr1029, Ala1030, Arg1033, Tyr1037, Ser1040, Tyr1041
MOE3	Glu15, Asp18, Ser25, Tyr26, Asn27, Asn28, Asp29, Lys30, Phe910, Leu911, Ala912, Ser914, Gly915, Ala916, Gln918, Ala988, Ile991, Met992, Val1025, Thr1029, Gly1032, Arg1033
Maestro1	Arg8, Lys11, Lys12, Phe13, Glu15, Glu16, Asp18, Phe20, Tyr26 - Lys30, Trp330, Leu371, Asn375, Asn765, Thr769, Gln813, Gln816, Phe820, Gly870, Ser874, Gly877, Met878, Asn881, Ser882, Asn885, Leu911, Ala912, Ser914, Gly915, Thr919, Ala988, Ile991, Met992, Thr1029, Ala1030, Gly1032, Arg1033, Ser1036-Tyr1041
Maestro2	Asn27-Lys31, Val164, Ala167, Ala168, Ile171, Gln172, Arg175, Asn199, Phe202, Ser203, Asp204, Asn207, Lys208, Asn210, Asp211, Ala214, Asp215, Gln216, Leu219, Gln222, Arg223, Asn375, Pro378, Cys379, Ala382, Gln918, Leu922, Phe925, Ala926, Asp929, Lys930, Leu933, Glu934, Val936, Gly937, Gln938, Pro970, Thr973, Ala974, Lys977, Tyr981, Phe985, Gln989
Maestro4	Ile6, Ser9, Ile10, Gly14, Glu15, Asn17-Phe20, Asn27, Ile245, Ile246, Val248, Ser249, Ile252, Ala256, Ile259, Ser264, Met322, Val329, Ile333, Cys336, Tyr337, Val366, Ile367, Ala370, Leu371, Gly374, Asn375, Arg1033

### 7.1.2.3. Visualization of binding pockets and e-pharmacophores

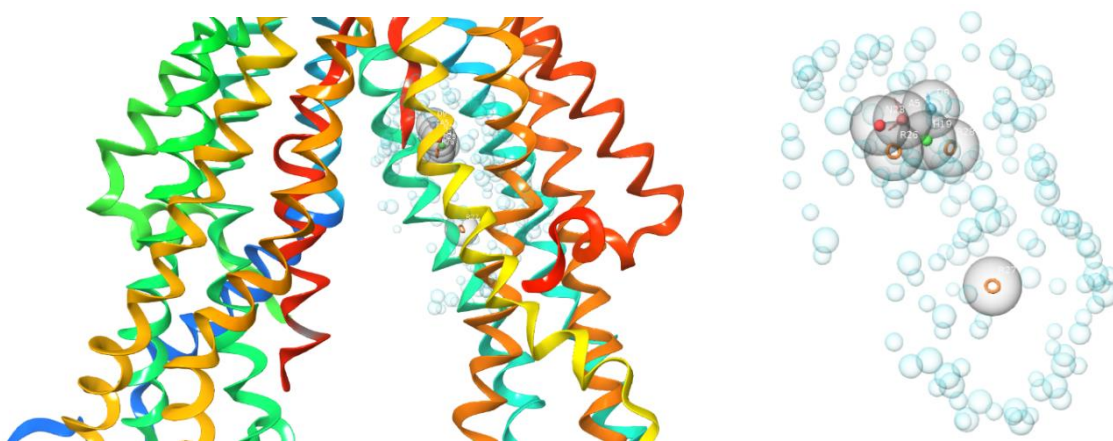


Figure 31: Pharmacophore of MOE1 with surrounding transporter helices (left) and only with excluded volumes (right); blue vector = HBD; red vector = HBA; orange ring = aromatic; blue spheres = exclusion volumes

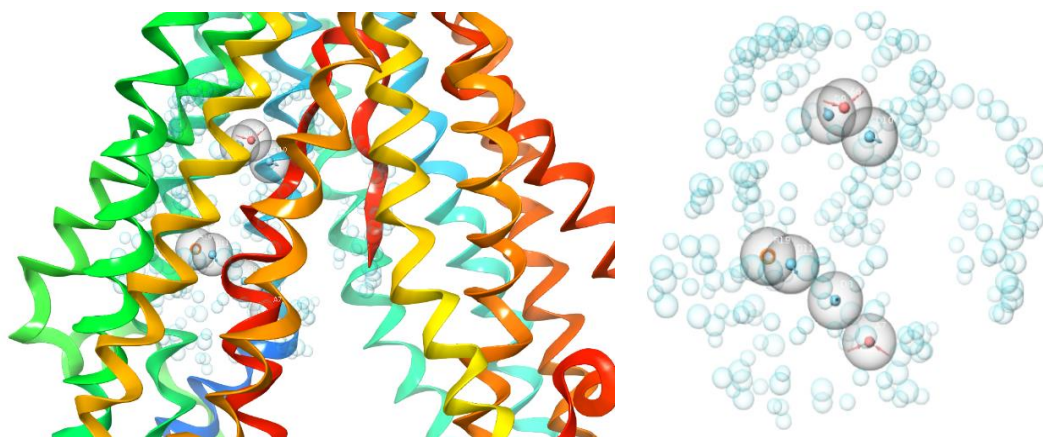


Figure 32: Pharmacophore of MOE2 with surrounding transporter helices (left) and only with excluded volumes (right); blue vector = HBD; red vector = HBA; orange ring = aromatic; blue spheres = exclusion volumes



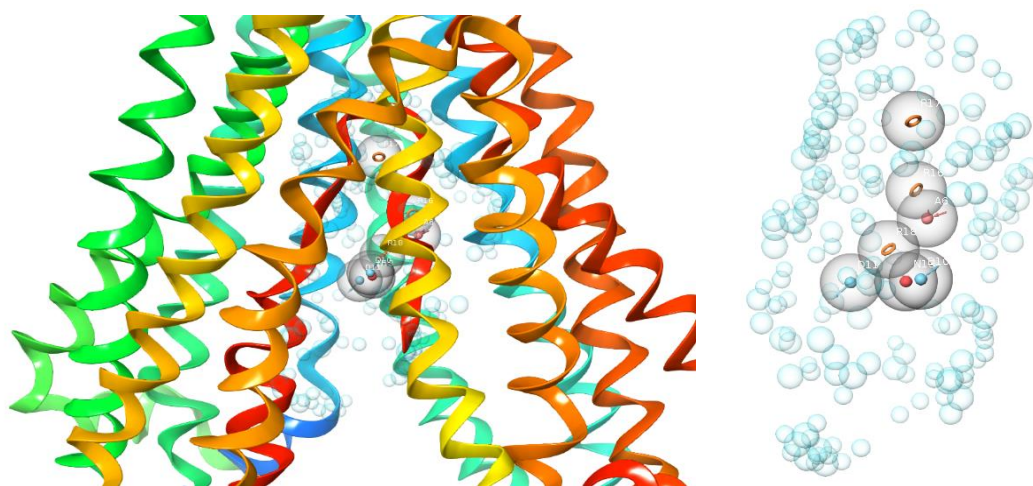


Figure 33: Pharmacophore of MOE3 with surrounding transporter helices (left) and only with excluded volumes (right); blue vector = HBD; red vector = HBA; orange ring = aromatic; blue spheres = exclusion volumes

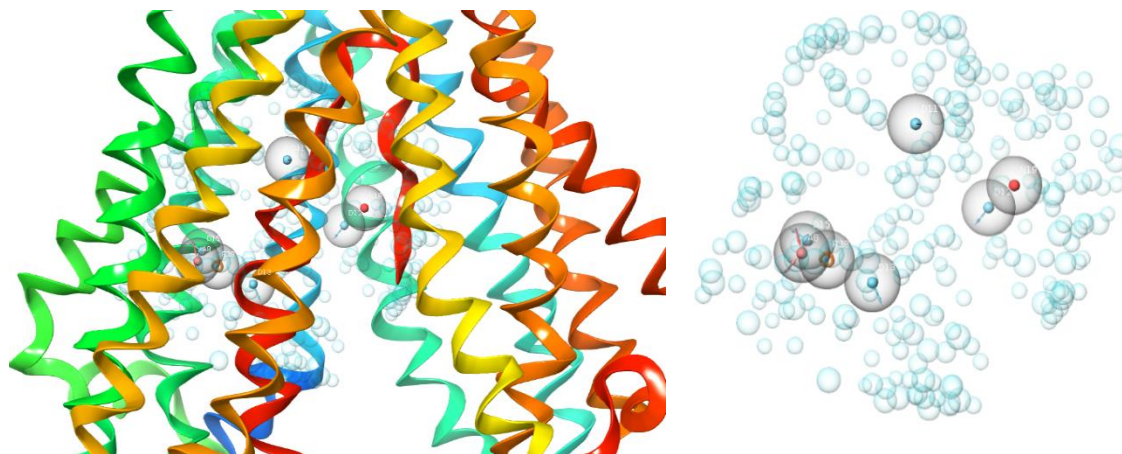


Figure 34: Pharmacophore of Maestro1 with surrounding transporter helices (left) and only with excluded volumes (right); blue vector = HBD; red vector = HBA; orange ring = aromatic; blue spheres = exclusion volumes

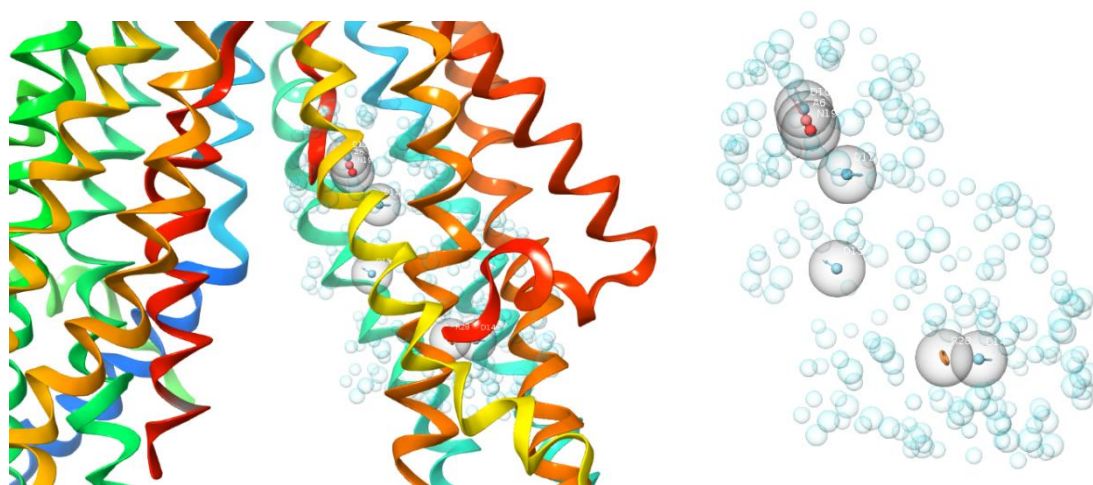


Figure 35: Pharmacophore of Maestro2 with surrounding transporter helices (left) and only with excluded volumes (right); blue vector = HBD; red vector = HBA; orange ring = aromatic; blue spheres = exclusion volumes

#### 7.1.2.4. Cluster19 ranking according to IFD Score and Emodel Score

Title	Stars	IFDScore	Prime Energy	glide emodel
Cluster 19 (145)				
ligprep_3.sdf:35	☆☆☆	-1845.2	-36729.8	-89.1
ligprep_3.sdf:36	☆☆☆	-1844.0	-36695.5	-95.7
ligprep_3.sdf:36	☆☆☆	-1841.8	-36683.5	-76.4
ligprep_3.sdf:24	☆☆☆	-1841.4	-36647.0	-91.7
ligprep_3.sdf:24	☆☆☆	-1840.3	-36647.5	-76.5
ligprep_3.sdf:24	☆☆☆	-1839.7	-36628.8	-82.6
ligprep_3.sdf:24	☆☆☆	-1838.7	-36646.1	-65.6
ligprep_3.sdf:34	☆☆☆	-1837.2	-36592.5	-80.4
ligprep_3.sdf:20	☆☆☆	-1837.1	-36554.1	-87.7
ligprep_3.sdf:20	☆☆☆	-1836.6	-36551.9	-82.5
ligprep_3.sdf:28	☆☆☆	-1836.4	-36560.7	-84.4
ligprep_3.sdf:28	☆☆☆	-1836.3	-36571.3	-74.5
ligprep_3.sdf:20	☆☆☆	-1836.2	-36550.0	-80.2
ligprep_3.sdf:20	☆☆☆	-1836.2	-36543.8	-84.4
ligprep_3.sdf:28	☆☆☆	-1836.0	-36564.5	-83.9
ligprep_3.sdf:28	☆☆☆	-1835.8	-36572.5	-74.9
ligprep_3.sdf:46	☆☆☆	-1835.7	-36542.7	-83.3
ligprep_3.sdf:46	☆☆☆	-1835.7	-36539.3	-79.0
ligprep_3.sdf:24	☆☆☆	-1835.3	-36532.1	-93.3
ligprep_3.sdf:28	☆☆☆	-1835.2	-36553.9	-75.3
ligprep_3.sdf:22	☆☆☆	-1835.2	-36548.7	-77.7
ligprep_3.sdf:20	☆☆☆	-1835.0	-36548.2	-68.9
ligprep_3.sdf:23	☆☆☆	-1835.0	-36552.1	-74.2
ligprep_3.sdf:24	☆☆☆	-1835.0	-36540.6	-79.4
ligprep_3.sdf:23	☆☆☆	-1834.9	-36514.0	-90.4
ligprep_3.sdf:24	☆☆☆	-1834.7	-36524.9	-90.1
ligprep_3.sdf:24	☆☆☆	-1834.7	-36536.5	-79.1
ligprep_3.sdf:28	☆☆☆	-1834.4	-36556.2	-69.6
ligprep_3.sdf:23	☆☆☆	-1834.1	-36502.6	-87.7
ligprep_3.sdf:24	☆☆☆	-1834.0	-36532.6	-71.1
ligprep_3.sdf:42	☆☆☆	-1834.0	-36458.3	-116.0
ligprep_3.sdf:32	☆☆☆	-1830.8	-36492.9	-74.6
ligprep_3.sdf:25	☆☆☆	-1830.6	-36444.9	-81.2
ligprep_3.sdf:11	☆☆☆	-1830.5	-36436.7	-80.9
ligprep_3.sdf:42	☆☆☆	-1830.5	-36453.5	-79.6
ligprep_3.sdf:15	☆☆☆	-1830.5	-36416.6	-84.8
ligprep_3.sdf:25	☆☆☆	-1830.2	-36452.8	-78.5
ligprep_3.sdf:25	☆☆☆	-1830.1	-36439.2	-80.5
ligprep_3.sdf:15	☆☆☆	-1830.0	-36432.0	-82.1
ligprep_3.sdf:15	☆☆☆	-1830.0	-36430.9	-82.0
ligprep_3.sdf:25	☆☆☆	-1829.9	-36452.8	-74.0
ligprep_3.sdf:15	☆☆☆	-1829.7	-36412.7	-77.3
ligprep_3.sdf:7	☆☆☆	-1829.7	-36442.0	-71.5
ligprep_3.sdf:25	☆☆☆	-1829.6	-36447.5	-77.5
ligprep_3.sdf:25	☆☆☆	-1829.5	-36441.8	-74.1
ligprep_3.sdf:37	☆☆☆	-1829.3	-36398.9	-86.0
ligprep_3.sdf:15	☆☆☆	-1829.0	-36412.8	-79.2
ligprep_3.sdf:15	☆☆☆	-1829.0	-36412.5	-75.7
ligprep_3.sdf:25	☆☆☆	-1829.0	-36451.4	-69.4
ligprep_3.sdf:25	☆☆☆	-1829.0	-36449.9	-66.4
ligprep_3.sdf:25	☆☆☆	-1828.9	-36444.3	-75.7
ligprep_3.sdf:26	☆☆☆	-1828.6	-36409.2	-77.5
ligprep_3.sdf:41	☆☆☆	-1828.6	-36418.3	-74.7
ligprep_3.sdf:8	☆☆☆	-1828.3	-36410.5	-82.3
ligprep_3.sdf:25	☆☆☆	-1828.2	-36392.4	-87.7
ligprep_3.sdf:8	☆☆☆	-1828.2	-36403.6	-74.2
ligprep_3.sdf:1	☆☆☆	-1828.1	-36423.9	-76.4
ligprep_3.sdf:26	☆☆☆	-1828.0	-36416.4	-71.2
ligprep_3.sdf:9	☆☆☆	-1827.8	-36395.6	-85.1
ligprep_3.sdf:15	☆☆☆	-1827.5	-36402.3	-65.5
ligprep_3.sdf:44	☆☆☆	-1827.4	-36375.5	-93.0
ligprep_3.sdf:44	☆☆☆	-1827.3	-36372.6	-94.1
ligprep_3.sdf:6	☆☆☆	-1827.2	-36387.9	-79.3
ligprep_3.sdf:44	☆☆☆	-1827.2	-36372.6	-89.3
ligprep_3.sdf:17	☆☆☆	-1827.1	-36386.3	-77.5
ligprep_3.sdf:29	☆☆☆	-1827.1	-36393.6	-83.0
ligprep_3.sdf:29	☆☆☆	-1827.0	-36378.2	-78.1
ligprep_3.sdf:29	☆☆☆	-1826.8	-36389.5	-82.4
ligprep_3.sdf:13	☆☆☆	-1826.8	-36392.8	-70.2
ligprep_3.sdf:26	☆☆☆	-1826.8	-36394.6	-68.4

Figure 36: Ranking of molecules in cluster 19 according to 1) IFD score and 2) Glide emodel score

Title	Stars	IFDScore	Prime Energy	glide emodel
ligprep_3.sdf:13	☆☆☆	-1826.7	-36396.1	-69.8
ligprep_3.sdf:29	☆☆☆	-1826.7	-36375.0	-81.3
ligprep_3.sdf:29	☆☆☆	-1826.5	-36382.5	-70.0
ligprep_taurocholic_acid.sdf:1	☆☆☆	-1826.5	-36346.7	-88.4
ligprep_3.sdf:25	☆☆☆	-1826.5	-36381.8	-71.7
ligprep_3.sdf:44	☆☆☆	-1826.4	-36359.1	-107.1
ligprep_3.sdf:6	☆☆☆	-1826.4	-36379.5	-70.7
ligprep_3.sdf:29	☆☆☆	-1826.1	-36385.5	-74.3
ligprep_3.sdf:25	☆☆☆	-1826.0	-36389.1	-73.4
ligprep_3.sdf:6	☆☆☆	-1826.0	-36380.2	-66.8
ligprep_3.sdf:26	☆☆☆	-1826.0	-36393.5	-61.7
ligprep_taurocholic_acid.sdf:1	☆☆☆	-1825.9	-36355.9	-72.8
ligprep_3.sdf:1	☆☆☆	-1825.9	-36419.2	-60.1
ligprep_3.sdf:6	☆☆☆	-1825.9	-36374.2	-73.8
ligprep_3.sdf:6	☆☆☆	-1825.9	-36370.6	-71.5
ligprep_taurocholic_acid.sdf:1	☆☆☆	-1825.8	-36350.3	-77.2
ligprep_3.sdf:6	☆☆☆	-1825.8	-36369.2	-72.6
ligprep_3.sdf:44	☆☆☆	-1825.8	-36355.0	-81.8
ligprep_3.sdf:25	☆☆☆	-1825.8	-36376.3	-74.8
ligprep_3.sdf:6	☆☆☆	-1825.8	-36374.1	-75.2
ligprep_3.sdf:6	☆☆☆	-1825.7	-36374.4	-75.5
ligprep_taurocholic_acid.sdf:1	☆☆☆	-1825.6	-36332.0	-82.5
ligprep_3.sdf:13	☆☆☆	-1825.6	-36379.8	-67.7
ligprep_taurocholic_acid.sdf:1	☆☆☆	-1825.6	-36355.6	-67.7
ligprep_3.sdf:16	☆☆☆	-1825.5	-36378.6	-54.8
ligprep_taurocholic_acid.sdf:1	☆☆☆	-1825.5	-36345.5	-77.9
ligprep_3.sdf:44	☆☆☆	-1825.4	-36355.3	-92.1
ligprep_3.sdf:18	☆☆☆	-1825.4	-36347.0	-79.8
ligprep_3.sdf:44	☆☆☆	-1825.3	-36355.3	-80.6
ligprep_3.sdf:27	☆☆☆	-1825.3	-36365.4	-65.8
ligprep_3.sdf:19	☆☆☆	-1825.3	-36338.1	-80.1
ligprep_3.sdf:25	☆☆☆	-1825.3	-36374.2	-66.9
ligprep_3.sdf:3	☆☆☆	-1825.2	-36344.8	-72.7
ligprep_3.sdf:6	☆☆☆	-1825.2	-36361.1	-69.7
ligprep_3.sdf:6	☆☆☆	-1825.2	-36372.7	-67.6
ligprep_3.sdf:6	☆☆☆	-1825.2	-36374.5	-62.3
ligprep_3.sdf:3	☆☆☆	-1825.2	-36355.7	-67.7
ligprep_3.sdf:16	☆☆☆	-1825.2	-36375.3	-55.0
ligprep_3.sdf:44	☆☆☆	-1825.1	-36358.1	-82.0
ligprep_3.sdf:6	☆☆☆	-1825.2	-36374.5	-62.3
ligprep_3.sdf:3	☆☆☆	-1825.2	-36355.7	-67.7
ligprep_3.sdf:16	☆☆☆	-1825.2	-36375.3	-55.0
ligprep_3.sdf:44	☆☆☆	-1825.1	-36358.1	-82.0
ligprep_taurocholic_acid.sdf:1	☆☆☆	-1825.1	-36355.3	-67.8
ligprep_3.sdf:16	☆☆☆	-1825.0	-36365.0	-60.3
ligprep_taurocholic_acid.sdf:1	☆☆☆	-1825.0	-36352.4	-68.5
ligprep_3.sdf:27	☆☆☆	-1825.0	-36339.2	-75.9
ligprep_3.sdf:25	☆☆☆	-1825.0	-36367.7	-71.8
ligprep_3.sdf:6	☆☆☆	-1825.0	-36370.6	-66.2
ligprep_taurocholic_acid.sdf:1	☆☆☆	-1824.9	-36355.4	-66.5
ligprep_taurocholic_acid.sdf:1	☆☆☆	-1824.9	-36353.9	-66.0
ligprep_3.sdf:21	☆☆☆	-1824.9	-36327.8	-91.7
ligprep_3.sdf:21	☆☆☆	-1824.5	-36321.3	-96.9
ligprep_3.sdf:6	☆☆☆	-1824.5	-36366.6	-63.7
ligprep_3.sdf:21	☆☆☆	-1824.4	-36335.7	-80.0
ligprep_taurocholic_acid.sdf:1	☆☆☆	-1824.4	-36356.8	-63.6
ligprep_taurocholic_acid.sdf:1	☆☆☆	-1824.3	-36342.4	-71.6
ligprep_3.sdf:25	☆☆☆	-1824.1	-36365.2	-64.1
ligprep_3.sdf:2	☆☆☆	-1824.1	-36389.6	-84.2
ligprep_taurocholic_acid.sdf:1	☆☆☆	-1824.0	-36347.5	-62.8
ligprep_3.sdf:2	☆☆☆	-1823.9	-36388.4	-81.6
ligprep_taurocholic_acid.sdf:1	☆☆☆	-1823.8	-36347.6	-55.9
ligprep_3.sdf:21	☆☆☆	-1823.8	-36322.7	-77.8
ligprep_3.sdf:21	☆☆☆	-1823.2	-36325.7	-76.5
ligprep_3.sdf:21	☆☆☆	-1823.2	-36331.8	-73.0
ligprep_3.sdf:21	☆☆☆	-1823.0	-36313.0	-76.9
ligprep_3.sdf:43	☆☆☆	-1822.9	-36383.7	-88.8
ligprep_3.sdf:21	☆☆☆	-1822.7	-36329.9	-68.8
ligprep_3.sdf:21	☆☆☆	-1822.4	-36311.9	-71.0
ligprep_3.sdf:5	☆☆☆	-1822.2	-36287.8	-73.3
ligprep_3.sdf:21	☆☆☆	-1822.1	-36315.9	-68.9
ligprep_3.sdf:16	☆☆☆	-1821.6	-36282.2	-62.6

Figure 37: Ranking of molecules in cluster 19 according to 1) IFD score and 2) Glide emodel score



## 7.2. Abstract

This master thesis focusses on a data-based and structure-based approach for the better understanding of the inhibition of the human bile salt export pump. The DILI-associated ABC transporter plays an important role in drug discovery and development since numerous drugs are discontinued in late clinical trials or withdrawn from market due to liver toxicity. Therefore, insights into molecular patterns causing inhibition and consequently impaired transport activity of BSEP would be of great benefit.

Data-based experiments such as molecular descriptor calculations and clustering methods were implemented in the attempt of finding structural features differentiating inhibitors and non-inhibitors. The data analysis showed a correlation between higher lipophilicity and higher molecular weights and increased inhibitory activity. These trends support previous literature findings. Additional descriptors, that were implicated in elevated inhibitory potential include higher molar refractivity values, higher flexibility and hydrogen-bond accepting properties.

Due to the release of the apo structure of the transporter in April 2020, structure-based investigations such as binding pocket detection and identification of important ligand features and ligand-protein interactions could be conducted. This thesis presents a possible orthosteric binding pocket with promising poses of the natural substrate taurocholate and a subgroup of inhibitors containing a sulfonamide moiety. Residues involved in protein-ligand interactions include the hydrogen bond donor Lys30, the hydrophobic residues Ile245, Val248, Ile367, Ala370 and Leu371, as well as aromatic residues such as Phe20, Phe334, Trp330 and Tyr337.

Further *in silico* investigations and *in vitro* assays are necessary to support the proposed binding hypothesis.

### 7.3. Zusammenfassung

Diese Masterarbeit beschäftigt sich mit der Analyse der Inhibierung des humanen Gallensäure-Exporters, kurz bezeichnet BSEP (bile salt export pump). Dieser ABC-Transporter wurde in der Vergangenheit mit schweren arzneimittelinduzierten Leberschäden assoziiert, die oftmals erst in fortgeschrittenen klinischen Studien oder nach Markteinführung erkannt werden. Erkenntnisse über die molekularen Zusammenhänge, die zu einer Inhibierung des Transporters und damit zu einer beeinträchtigten Transportaktivität führen, wären äußerst wertvoll für die Arzneistoffforschung.

Datenbasierte Experimente, wie molekulare Deskriptorberechnungen und Clusteringmethoden, wurden implementiert, um strukturelle Unterschiede zwischen Inhibitoren und Nicht-Inhibitoren auszumachen. Die Datenanalyse zeigte eine signifikante Korrelation zwischen höherer Lipophilie und höherem Molekulargewicht und erhöhter Inhibierungsaktivität. Diese Trends unterstreichen bereits gefundene Resultate in Literaturquellen. Zusätzliche Deskriptoren, die mit Inhibierung des Transporters in Zusammenhang gebracht wurden, inkludieren höheres molares Brechungsvermögen, höhere Flexibilität und Wasserstoffbrückenakzeptoren.

Die im April 2020 veröffentlichte Apostruktur wurde zur Detektion möglicher Bindungstaschen und in weiterer Folge zur Analyse wichtiger Ligand-Protein-Interaktionen herangezogen. In dieser Arbeit wird eine mögliche orthosterische Bindungstasche vorgestellt, die vielversprechende Posen mit dem natürlichen Substrat Taurocholsäure und einer Gruppe von Inhibitoren, die eine Sulfonamid-Einheit aufweisen. Aminosäuren, die Wechselwirkungen mit dem Liganden zeigen, inkludieren den Wasserstoffbrückenakzeptor Lys30, hydrophobe Reste wie Ile245, Val248, Ile367, Ala370 und Leu371, sowie die aromatischen Reste Phe20, Phe334, Trp330 und Tyr337.

Um die vorliegende Hypothese zu validieren, müssen weitere *in silico* sowie *in vitro* Untersuchungen durchgeführt werden.

## 7.4. Abbreviations

ABC ...	ATP-binding-cassette
AMW ...	molecular weight
ATP ...	adenosine triphosphate
BA ...	bile acid
BSEP ...	bile salt export pump
Cryo-EM ...	Cryogenic electron microscopy
DILI ...	drug-induced liver injury
GC ...	glycocholic acid
GCDC ...	glycochenodeoxycholic acid
Glide ...	grid-based ligand docking with energetics
GpiDAPH3 ...	graph-p-donor-acceptor-polar-hydrophobe-triangle
HB ...	hydrogen bonding, hydrogen bonds
HBA ...	hydrogen bond acceptor
HBD ...	hydrogen bond donor
IC50 ...	half maximal inhibitory concentration
IFD ...	induced-fit docking
KNIME ...	Konstanz information miner
MACCS ...	molecular access system
MDR ...	multi-drug resistance
MMP ...	matched molecular pair
Mol. ...	Molecule
NBD ...	nucleotide-binding domain
P-gp ...	P-glycoprotein
SAR ...	structure-activity-relationship
SIFt ...	structural interaction fingerprint
SMR ...	molar refractivity
TC ...	taurocholic acid
TCDC ...	taurochenodeoxycholic acid
TGD ...	typed-graph distances
TMD ...	transmembrane domain
TPSA ...	topological polar surface area