



universität
wien

DIPLOMARBEIT / DIPLOMA THESIS

Titel der Diplomarbeit / Title of the Diploma Thesis

**„Investigating the structural determinants defining the
substrate specificities of the human Bile Acid transporters
NTCP (SLC10A1) and ASBT (SLC10A2)“**

verfasst von / submitted by

Viktoria Gamsjäger

angestrebter akademischer Grad /

in partial fulfilment of the requirements for the degree of

Magistra der Pharmazie (Mag.pharm.)

Wien, 2020 / Vienna, 2020

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

A 449

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

Diplomstudium Pharmazie

Betreut von / Supervisor:

Univ.-Prof. Mag. Dr. Gerhard Ecker

Mitbetreut von / Co-Supervisor:

Claire Colas, PhD

Acknowledgements

First and foremost I want to thank **Prof. Dr. Gerhard Ecker** for giving me this unique opportunity of working in his group. He always ensured that I was feeling comfortable when entering this new world of research. Moreover he was always supporting me with his profound knowledge and guided me through arising obstacles with his great treasure of experience.

Next I want to thank **Claire Colas** for her guidance, endless patience and heart-warming personality. You dedicate your time in the most admirable and rousing way to your research. I marvel your passion and engagement for this sometimes very demanding work and hope you will always keep this enthusiasm for transporters and the unraveling of their mysteries.

My appreciation as well goes to every member of the **Pharmacoinformatics Research Group** for creating such a motivating and yet relaxed working atmosphere. From the first day I felt accepted as a part of the team and was glad about the prevailing mutual helpfulness.

Finally my special gratitude is dedicated to my beloved ones: I want to thank my parents, **Brigitte** and **Rudolf**, and my boyfriend **Michael** for never letting me down. You always had good advice and knew how to motivate or put my mind at ease when hard times arose. Thank you for all of your words of encouragement and your endless affection.

Table of contents

1. Introduction.....	1
1.1. The Enterohepatic Circulation and Recycling of Bile Acids.....	1
1.2. NTCP and ASBT: Striking members of the SLC10 Family.....	1
1.3. Structure of native Bile Acids.....	3
1.4. Transporter abnormalities and correlated diseases.....	4
2. Aim of the thesis.....	5
3. Material and Methods.....	7
3.1. Homology Modeling.....	7
3.1.1. Validation by Enrichment.....	9
3.2. Docking Study.....	10
3.2.1. Ligand and Protein Preparation.....	10
3.2.2. Receptor Grid generation.....	10
3.2.3. Docking of ligands.....	11
3.3. Induced Fit Docking	12
3.4. Heat map and PLIF	13
3.5. Ligand-based Pharmacophore Modeling – Phase	13
4. Results and discussion.....	15
4.1. Homology Modeling of ASBT.....	15
4.1.1. Structure of ASBT and NTCP.....	16
4.1.2. Template selection and Alignment.....	20
4.1.3. Model building, assessment and refinement.....	20
4.1.4. Re-Docking of TCH for Binding site refinement.....	22
4.1.5. Model validation by Enrichment.....	27
4.1.6. Modeling and preparation of hNTCP.....	29
4.2. Docking of BA in ASBT inward open conformation.....	30
4.2.1. Preparation and Receptor Grids.....	30
4.2.2. Docking of Ligands – Affinity inward open	30
4.2.3. “Structural water Hypothesis”	34
4.2.4. Comparison for NTCP.....	34

4.3. Induced Fit Docking of BA in ASBT (outward open)	35
4.3.1. Clustering based on Volume Overlap	37
4.3.2. Trend in orientation of clustered poses	37
4.3.2.1. Selectivity: “Ser-Thr-Lock Theory”	38
4.3.2.2. Affinity calculations of clustered poses	39
4.3.3. Heat map and PLIF	42
4.4. Ligand-based Pharmacophores of known ASBT substrates	46
4.5. Phylogenetic Relationship within the SLC10 family	50
5. Conclusions and Outlook	57
6. References	60
7. Appendix	65
7.1. Supplemental material	65
7.2. Abstract	68
7.3. Zusammenfassung	69
7.4. List of Abbreviations	70

1. Introduction

1.1 The Enterohepatic Circulation and Recycling of Bile Acids

Primary **Bile Acids** (BA) are generated from cholesterol during a complex cascade of synthesis in the liver, then conjugated with taurine or glycine and stored in the gallbladder as a major component of the human bile. BA function as detergents to aid digestion and are facilitating the absorption of fats, fat-soluble vitamins and the solubilization of cholesterol, which are released in the intestine through contractions of the gallbladder when having a meal.¹ Their activity as emulsifying agent is possible because of their amphiphilic structure containing one hydrophilic (hydroxyl-groups) and one hydrophobic side (methyl-groups and steroid scaffold).²

The Enterohepatic Circulation can be summarized as a process of BA recycling between the liver and the intestine mainly facilitated by the two transporters NTCP and ASBT.

Due to the efficient work of those transporters, over 90% of BA can be reclaimed from the intestine and brought back to the liver through the systematic blood circulation, which results in less than 10% *de novo* hepatic synthesis.^{3,4} Those primary BA that circumvent the reabsorption via ASBT in the terminal ileum, are further on chemically modified by colonic enterobacteria and transformed into secondary BA via bacterial 7-dehydroxylation.

1.2 NTCP and ASBT: Striking members of the SLC10 Family

The solute carrier family 10 (SLC 10) consists of seven influx transporters of Bile Acids (BA), steroidal hormones or a diversity of substrates, which are involved in physiological processes of the human body.⁴

The first two discovered members, NTCP (SLC10A1) and ASBT (SLC10A2) are both sodium-dependent co-transporters of bile acid and therefore contribute a major part to the enterohepatic circulation (EHC).⁵

NTCP, also called sodium/taurocholate co-transporting polypeptide, is exclusively located in the sinusoidal membrane of hepatocytes in the liver and responsible for the uptake of BA from the portal blood circulation into the hepatocytes.

ASBT, further known as the apical sodium-dependent bile acid transporter, is mainly expressed in the brush boarder membrane of enterocytes (*ileocytes*) in the terminal Ileum. Its major task is the initial uptake of BA across the enterocyte brush border membrane and therefore clears BA from the ileum to the portal blood vein, where they are, as a part of the EHC, again being redelivered to the liver via NTCP (figure 1).⁴

ASBT is known to have a **narrow substrate specificity**, transporting **solely bile acids**. In contrast it's shown that **NTCP** additionally transports **sulfo-conjugated BA** and **steroid sulfates** (oestrone-3-sulfate, DHEAS e.g.).^{4,6}

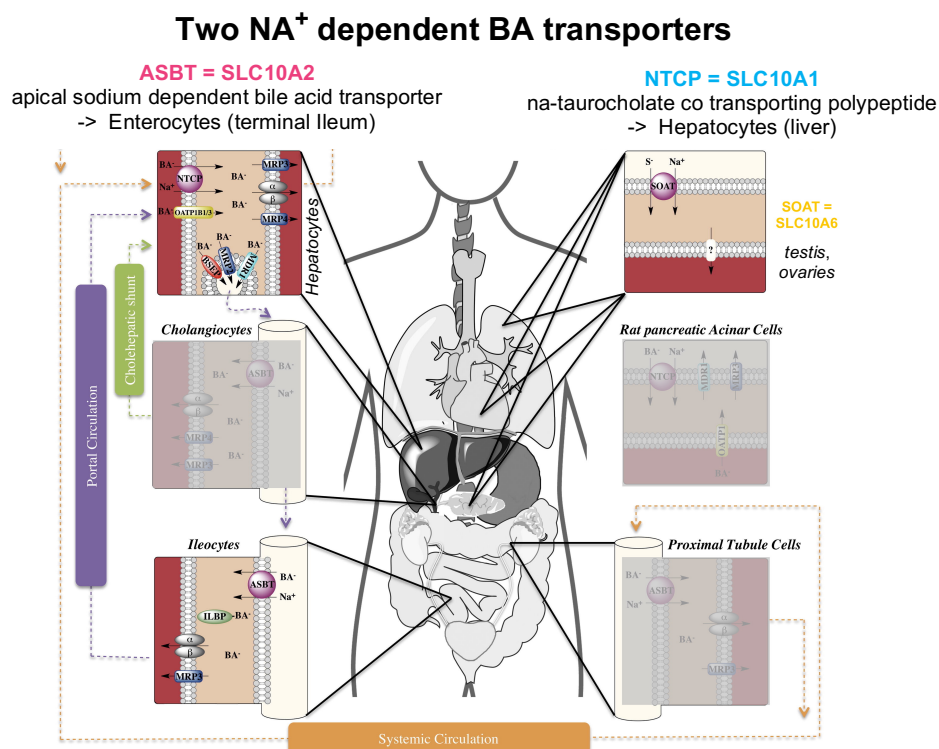


Figure 1: The Enterohepatic circulation and the contribution of ASBT and NTCP.⁴

ASBT being responsible for the initial BA uptake from the intestine and transport to the portal circulation. Whereas NTCP clears BAs from the portal blood vein and returns them into the hepatocytes as a part of the BA recycling process.

NB: in this picture another SLC10 family member, SOAT (SLC10A6) is marked, who is structurally closely related to ASBT but strictly transports sulfated steroids. Therefore SOAT is used as another source of information.

In this regard NTCP and ASBT can be seen as the leading and rate-limiting mediators of BA-uptake and homeostasis in the liver and intestine.⁴

1.3 Structure of native Bile Acids

Since BA are the major physiological substrates of ASBT and NTCP it is important to pay attention to the structural requirements of those natural compounds in order to understand the prevailing transport mechanisms and their different substrate preferences:

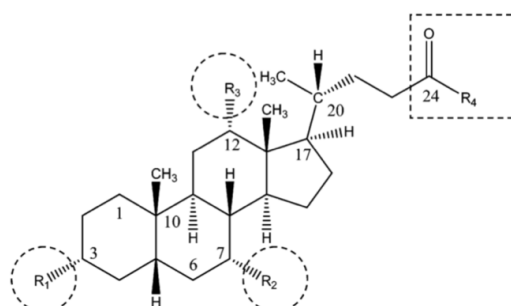
As mentioned above **primary BA** are synthesized from cholesterol in hepatocytes, hence they consist of a steroidal scaffold and possess different hydroxylation patterns. Generally, they are di- or tri-hydroxylated at position C₃, C₇ and/or C₁₂, with each hydroxyl-group in α -position, except Ursodeoxycholate and its conjugated derivatives (β -position).

Furthermore, BA can be divided into **unconjugated BA** (free carboxylic acid) or **conjugated BA** (glycine or taurine substituent) at position C₂₄ (table 1).

Secondary BA are deduced from primary ones, that escaped the reabsorption by ASBT, via 7-dehydroxylation or 7-epimerisation (α - to β -OH) by the bacterial metabolism in the small intestine. Again, there is an additional classification in unconjugated and conjugated secondary BA.

The Nomenclature of the different BA can be seen in table 1.³

Table 1 Bile acid nomenclature and structure ²				
bile acid	R ₂ (C-7)	R ₃ (C-12)	R ₄ (C-24)	
Primary Bile Acids				
cholate	OH	OH	OH	
glycocholate			NHCH ₂ COOH	
taurocholate			NH(CH ₂) ₂ SO ₃ H	
chenodeoxycholate	OH	H	OH	
glycochenodeoxycholate			NHCH ₂ COOH	
taurochenodeoxycholate			NH(CH ₂) ₂ SO ₃ H	
Secondary Bile Acids				
deoxycholate	H	OH	OH	
glycodeoxycholate			NHCH ₂ COOH	
taurodeoxycholate			NH(CH ₂) ₂ SO ₃ H	
lithocholate	H	H	OH	
glycolithocholate			NHCH ₂ COOH	
tauroolithocholate			NH(CH ₂) ₂ SO ₃ H	
ursodeoxycholate	OH (β)	H	OH	
glycoursoxycholate			NHCH ₂ COOH	
tauroursoxycholate			NH(CH ₂) ₂ SO ₃ H	



Generally both proteins are known to transport each physiological di- and tri-hydroxylated BA, preferring conjugated over unconjugated ones.⁶

1.4 Transporter abnormalities and correlated diseases

Due to the importance of the function and contribution that NTCP and ASBT perform in the EHC, transporter abnormalities such as mutations can be involved in serious gastrointestinal disorders.

For example, loss of functions caused by point mutations of ASBT (*Thr262Met*; *Leu243Pro*) are linked to **Primary Bile Acid Malabsorption (PBAM)**. The inherited PBAM is accompanied by symptoms such as severe diarrhea, steatorrhea along with an elevated excretion of BA and therefore lowered plasma cholesterol levels and malnutrition.^{1,4}

It is likely that some other diseases like Familial Hypertriglyceridemia or Idiopathic Chronic Diarrhea are linked to the downregulation of the uptake transporters NTCP and ASBT.

Furthermore it is believed that malfunctions or downregulation of the BA transporter ASBT are further affecting the intestinal function and could be involved in Chronic Ileitis, Cholesterol and Black Pigment Gallstone disease, Crohn's disease, and even the contribution to Colon cancer is debated.¹

Since NTCP is the leading transporter for hepatic BA uptake, transport impairing single nucleotide polymorphisms (SNPs) could affect liver function or drug disposition. The therapeutic effect of many drugs is known to be dependent on the intact enterohepatic circulation. The deviation from normal re-absorption processes could have unpredictable consequences e.g. on the half-life or plasma level of the drug.⁷ Yet little is known about the impact and occurrence of SNPs causing loss of functions regarding hepatocyte damage, consequences on cholesterol excretion or assumed rising serum BA levels, which strengthens the need for further investigation.⁸

2. Aim of thesis

Since ASBT and NTCP are known to have a key role in the EHC, they have been more and more in the focus of research aiming to treat widespread diseases like **Hypercholesterolemia**.

ASBT inhibition would lead to a high elimination rate of not-absorbed BA via fecal excretion and as a consequence reduce plasma cholesterol levels.

In order to compensate the lack of recycled BA the hepatic *de novo* synthesis would be propagated, leading to a greater cholesterol consumption. Additionally, an up-regulation of hepatic LDL receptors and therefore increased plasma LDL-cholesterol uptake into the liver can be seen.^{1,4}

Secondly ASBT is an interesting target for **prodrug approaches**, since it is likely to improve oral bioavailability by utilizing its uptake mechanism.

Two different ways of tackling the problem could be imaginable: On one hand it's possible to link a drug to a natural substrate of the transporter, so called "Trojan Horses" for delivery of therapeutics mentioned by *Polli et al.*³ On the other hand, "substrate mimicry" can be performed, where the 3D structure of the drug mirrors natural substrates, in order to enhance the active transport. This procedure could lead to tremendous success of targeted delivery for currently poor bioavailable drugs due to the localization of NTCP in the liver and of ASBT in the ileum.

Because of all these diverse mentioned application possibilities the **aim of my thesis is to understand the determinants of the substrate's specificities and gaining insight in the transport mechanism of ASBT and NTCP**. Moreover, it would be valuable to understand the crucial structural differences causing the boarder substrate specificity of NTCP and discovering the reason for ASBT's strict limitation to BAs. Based on this knowledge it would be **possible to either synthesize new compounds** or conduct **virtual screening of already existing drugs**. This would allow a **classification and**

repurposing as **substrates or inhibitors**, thus achieving the above-mentioned benefits.

Due to the complexity of this topic we essentially concentrated on the elucidation of ASBT as a key target and then applied the gained information to analyze NTCP.

3. Material and Methods

3.1 Homology Modeling

The methodology of Homology Modeling is based on the assumption that **functionally related proteins share common structural properties and therefore own similar fold-motifs**. As there is only a limited number of possible folds, two proteins with a sequence identity of about 30% to 40% are likely to share similar shapes, which builds the foundation to postulate comparable transport mechanisms.⁹ SLC transporters seem to constitute an exception and allow to draw structural conclusions even from low sequence identity about only 10% due to a common evolutionary conserved fold motif.¹⁰

In absence of an experimentally determined crystal structure of the target-protein, homology modeling is a reliable computational prediction method which allows an accurate structure prediction. The sought three dimensional-structure is calculated, in respect to a known phylogenetic related template, based on the knowledge that a protein's fold can be deduced from its primary amino acid sequence.¹¹

Since the applied structure prediction is based on probed structures of close related proteins (homologs) it's also known as "Comparative modeling".

The process consists of **four major steps** (figure 2):

At first the **template selection** is done, aiming to identify a known structure closely related to the target protein. This can be done with a tool called **BLAST** (Basic Local Alignment Search Tool) or **HHPred** (Homology detection and structure prediction by HMM-HMM comparison). By comparing protein sequences BLAST evaluates and ranks considerable matching proteins according to their similarity, making them suitable to be chosen as a template structure.^{12,13} HHPred applies a similar procedure for identifying the most homologous protein sequence and moreover allows to discover targets with (semi)conserved motifs, through sequence search.¹⁴

Homology modeling

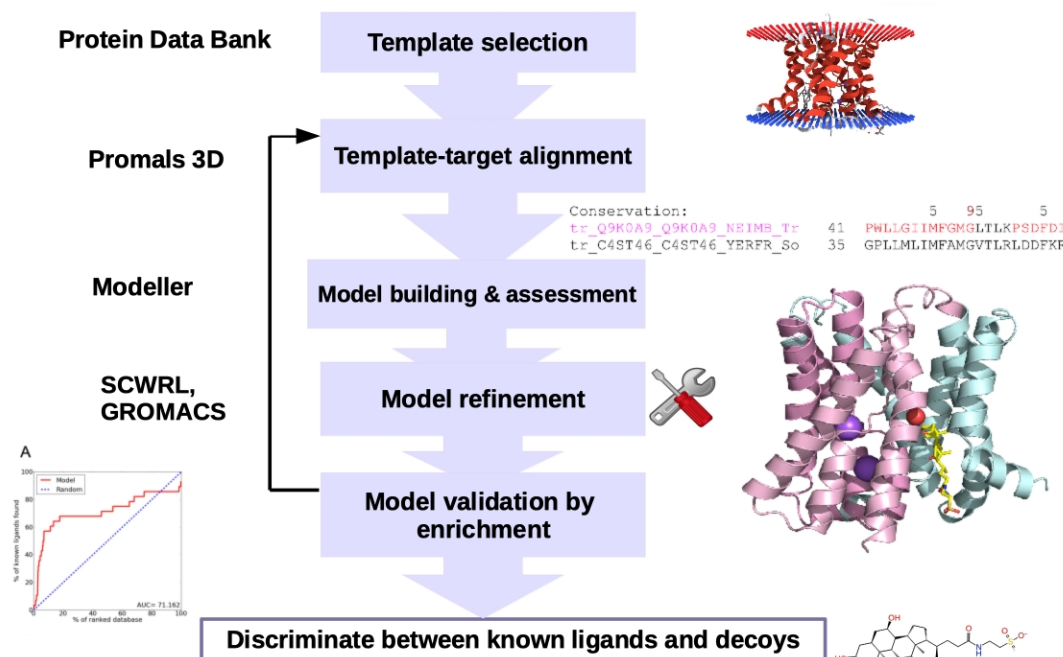


Figure 2: Major steps of Homology Modeling

An iterative process which is highly dependent on the accuracy of the underlying chosen homolog-template.

Another important source of information about the chosen protein is the Protein Data Bank (PDB), which contains crucial data of the template crystal structure such as ligands or mutations.

The sequence identity between template and target should ideally be above 40% or higher, since then according to *Sali et al* the modeled protein atoms are expected to differ only with an RMSD* of 1 Å (Angström) because of great correspondence between the x-ray structure and the selected protein.¹⁵ Interestingly, a lot of SLC transporters share the same fold despite lower identities (< 30% till 10%), which is an additional challenge for the modeling process.¹⁰ The second step, the **template-target alignment**, is conducted with an alignment tool such as PROMALS3D.¹⁶ Here the structural and sequence information of the known homolog and the targeted protein are

* Root-mean-square deviation: a quantitative measure of similarity between proteins

combined in order to obtain an accurate predicted alignment of the sequences.¹⁷

The alignment is then used **to generate the actual homology model**. This is done with the help of tools like MODELLER.¹⁸ As an output a 3D structure is obtained, that satisfies the spatial constraints set as accurate as possible. The focus is set to e.g. conformational limitations of the main-chain due to the type of occurring amino acids, dihedral angle restraints or main-chain N-O distances as *Sali et al* specify in their work.¹⁹

The fourth step is the **model refinement**, where the alignment can be manually adjusted by minimizing gaps or aligning functionally important residues, the process itself is iterative until a suitable model is generated.¹⁰

3.1.1 Validation by Enrichment

The **Validation** is done **by Enrichment**, to assess the performance and predict the capacity of the attained model to discriminate between known ligands and decoys using docking. Decoys are compiled by the webtool DUD E (Database of Useful Decoys: Enhanced).^{10,18}

Decoys are molecules generated with similar physicochemical properties e.g. molecular weight, quantity of hydrogen bond donors and acceptors but a diverging 2D topology to serve as a negative control for the enrichment procedure. They are originally designed to not actually bind the target protein. The calculated Enrichment Curve displays the ratio of ranked known ligands (true actives) in respect to the screened database (in %), additionally containing the generated decoys. The model evaluation is then done by calculating the AUC (area under the curve) of the enrichment plots, representing how the docked poses of the known ligands are better ranked as compared to the poses of the non-binding decoys against the generated homology model.

The obtained data is used to improve the homology model by refining side chains, inspecting the binding site for regions of shifted amino acids or uncover misalignments on the protein to reach AUC scores far better than 50%.^{20,21,22,23}

3.2 Docking Study

The process of molecular docking is performed to predict the orientation of a ligand placed in the binding site of a protein (= ligand pose) and therefore gain information about the stated interactions. As a result a depiction of the ligand-receptor complex is obtained.²⁴ For unknown structures docking against a homology-modeled protein is a widely spread opportunity to elucidate the interactions of a protein and its ligand.²⁵

This enables us to create a possible binding hypothesis for less investigated proteins. The docking was conducted using the molecular docking program Glide provided by Schrödinger.²⁶

3.2.1 Ligand and Protein Preparation

The protein and ligand preparation steps are done to ensure chemical correctness and optimization of the structures for further usage in the docking process with Glide.

The **Ligand Preparation** can be done in the “**LigPrep** panel” and secures an optimally processed ligand e.g. with added hydrogens or various ionization states and correct chiral properties to start the docking procedure with.

Using the “**Protein Preparation Wizard** panel”, the refinement and hydrogenation of input files can be conducted, since usually PDB input files only consist out of heavy atoms, sometimes contain incorrect bond orders or atomic clashes.^{26, 27}

3.2.2 Receptor Grid generation

Since the aim of ligand docking is to identify, predict and depict relevant interactions between different ligands and a receptor, the correct shape and important protein properties need to be defined. The so called “grid” determines the conformational area of the protein’s ligand binding site which is considered to be investigated during the docking process itself.²⁸ This grid set-up can be done by using Glide’s “**Receptor Grid Generation** panel”.

Most of the time this is done manually according to structure and known properties of the binding site. The more precise space limitation are set during this step, the more precise docking poses and scoring values can be expected afterwards.²⁹

3.2.3 Docking of ligands

As shortly mentioned above, docking of ligands is a widely used and acknowledged computational tool to identify, predict and depict relevant interactions between different ligands and a receptor (protein). It enables the illustration and explanation of ideal protein-ligand interactions, allows to estimate binding affinities of various ligands, and paves the way for screening new unknown ligands. It yields for an exhaustive elucidation of a protein's transport mechanism and preferred ligands based on rational docking calculations.^{28,30}

For the calculations Schrödinger's docking program **Glide** (*Grid-based Ligand Docking with Energetics*) was used.²⁶ Glide has been designed for docking calculations with a rigid receptor structure and if chosen flexible ligands in order to dock a large number of ligands with a suitable output in moderate time. Whenever more precise output is required, time-consuming calculations as induced fit docking needs to be operated (see 3.3 Induced Fit Docking). The standard docking protocol for Glide includes the former mentioned protein preparation, ligand preparation, grid generation and finally the docking process itself.

As an output a so called "ligand pose" is acquired containing the ligand's specification and spatial orientation in relation to the protein's binding site. Glide is able to sieve and narrow down the calculated ligand poses within the ongoing process through a sequence of hierarchical filters which rates ligand poses according to various values as GlideScore, an adapted version of the ChemScore and used for binding affinity prediction and rank-ordering, or calculated energies between the ligand and receptor. This hierarchical process of selection aims to terminate irrelevant ligand conformations and therefore provides accurate ligand poses for further investigation.^{30,26}

3.3 Induced Fit Docking (IFD)

Since the standard docking process in Schrödinger's Glide is performed with a rigid receptor this may lead to a wrong/excessive energy penalty for ligands preferring interactions with a slightly alternated receptor conformation as the one used from the input file. This could lead to a wrong classification as a "non-binder" of an actual active compound or poor scoring of "true binders".³¹

In fact a **receptor is able to alter its binding site** to complement the binding mode of a ligand, this circumstance will be considered by exhaustive induced fit docking calculations.

The idea behind these calculations is based on the so called "**induced fit theory**" originated by *Koshland*, assuming the protein-ligand interactions to be a permanent process of binding site adaption depending on the current ruling ligand properties, resulting in multiple possible binding site conformations.³² In order to evade this falsely ranking of ligands it is often suggested that both, the protein (receptor) and the ligand, should be assumed as flexible during the docking procedure. The probability of misleading docking results can be reduced and additional protein conformations can be achieved.²⁴ This concept is implemented in Schrödinger's induced fit docking protocol, which uses Glide and Prime to execute the calculations of new receptor side chain conformations and various ligand binding modes.^{33,31}

These time- and processing power-consuming calculations can be carried out when e.g. x-ray structures of the ligand-protein complex are missing and can be seen as a tool to predict the actual binding mode and its stated interactions for this chosen conformation.³¹

The induced fit docking protocol procedure can be divided into three major steps: first, various ligand poses are calculated with Glide by regularly docking the ligand into the initial protein structure, then, as a second step, Prime performs the "actual induced fit" part by adapting each binding-site to its specific output-ligand. In the end the ligand is then redocked into these newly obtained receptor conformations, assessed and graded according to the GlideScore for redocking and the Prime energy.³¹

3.4 Heatmap and PLIF

The **Heatmap** calculated is based on the **Protein Ligand Interaction Fingerprints** data (PLIF). The PLIF is a tool that merges all interactions between the ligand and the protein depicting it as a Fingerprint scheme, in order to reveal frequent interaction patterns. Features such as “hydrogen bonding, hydrogen bond acceptor or donor and ionic interactions” are taken into account for the Fingerprint creation.³⁴ The PLIF input data was calculated with Maestro using the “Interaction Fingerprints” panel, then uploaded to MOE (Molecular Operating Environment), a molecular modeling software, where an inhouse script was used for creating a **Heatmap**. This Heatmap is a **graphic representation** of the occurring protein-ligand **interaction types** (hydrophobic, polar,...) plotted against the **involved amino acid (aa) residues** and illustrated with a color code.³⁵

3.5 Ligand-based Pharmacophore Modeling – Phase

In 1909 the novel concept of “Pharmacophores” was introduced by Paul Ehrlich, which are described as a “*molecular framework that carries (phoros) the essential features responsible for a drug's (pharmacon) biological activity*”. A Pharmacophore model resembles a 3D description of critical chemical ligand properties (=features) needed to interact with a receptor (=protein). On one hand **structure-based Pharmacophore models** can be generated, directly deriving necessary features for binding from crystal structures of the protein-ligand complex reflecting the essential properties for interaction.

On the other hand the **ligand-based** approach is made, in absence of x-ray structure information, by using superimposed structures of known active binders to identify and accumulate common features.^{36,37}

Due to the fact that a homology model is only an approximation and some interactions could be left unaccounted, we opted for the ligand-based approach. Using known BA substrates of ASBT as a training set served to identify so called “common Pharmacophores”. Considering only shared ligand features enables to establish a common *pharmacophore hypothesis*, a

detailed description of ligand binding illustrated on the basis of displayed steric features.³⁸

All calculations were conducted in Schrödinger's application Phase.^{39,40}

4. Results and Discussion

4.1 Homology Modeling of ASBT and NTCP

So far neither the crystal structure of human ASBT (**hASBT**), nor of human NTCP (**hNTCP**) could be solved. However, as x-ray structures of two prokaryotic homologs ASBT_{NM} and ASBT_{Yf} were available, we were able to build our homology models on the basis of information obtained by these templates. The crystal structures of the homolog **ASBT_{NM}** are acquired from *Neisseria meningitidis* with 26% sequence identity and 54% similarity to hASBT. Furthermore, strictly conserved binding site residues and a shared common motif can be seen.

The two conformational structures of **ASBT_{Yf}** were resolved from *Yersinia frederiksenii* with 22% sequence identity and 59% similarity towards hASBT. Additionally, a molecule of **TCH** (Taurocholate), a physiological ligand, bound to the binding site of the **inward-open** state is included, providing essential information. In addition, a similar topology and conserved binding site residues can be found, which makes it to a valuable template.^{5,41}

An overview of the homolog properties can be seen in table 2.

Both of the homolog organisms were used to build **three different homology models** of **hASBT** with the advantage of altered conformations (inward-open, outward-open) of the transport cycle and differing ligands bound, in order to gain as much structural information as possible.

Likewise we built **two homology models** of **NTCP**, based on the same prokaryotic homologs, in the inward-open (3ZUY) and outward-open (4N7X) conformation, to facilitate the comparison of those two BA transporters within altered conformations.

Table 2 Summary of Homolog properties

ASBT _{Yf}	ASBT _{NM}
22% sequence identity	26% sequence identity
59% similarity to hASBT	54% similarity to hASBT
PDB ID: 4N7X <i>outward-open</i>	PDB ID: 3ZUX <i>artificial mutation</i>
PDB ID: 4N7W <i>inward-open</i>	PDB ID: 3ZUY <i>inward-open</i>
Citrate in binding site (<i>inward open</i>) → from crystallization buffer	TCH in binding site (<i>inward open</i>) → physiological ligand

4.1.1 Structure of ASBT and NTCP

According to several hydrophobicity analyses and experimental data *Hagenbuch et al.* came to the conclusion that human ASBT and NTCP consist out of an **odd number of transmembrane domains** (TMDs), most likely seven or nine.^{42,6}

Sequence analysis exhibited 10 TMDs of the ortholog ASBT_{NM}, arranged into two inverted 5 TMD repeats which deviates from the assumption of seven or nine TMDs for hASBT.

It is still speculated and not entirely clarified whether hASBT and NTCP are assembled by 9-TM or 7-TM sequences. Döring *et al* argue in 2012 that, because of an inexistent and therefore unpredictable first bacterial transmembrane segment, hASBT consists out of 9 segments, while Swaan and his group's topological prediction experiments from 2004 favored the 7-TM assumption.^{42,43} It is further speculated if so called "substrate sensible re-entrant loops" are a potential reason for these ambiguous outcomes of several hydrophobicity analyses through the years. These special loops are winding towards the membrane, partly entering and then reversing to its origin. Their

reason for forming and influence is often unrevealed and requires further inspection.^{43,44} Supplementary analyses are also needed according to *Swaan* and colleagues to evaluate the suitability of ASBT_{NM} for depiction of mammalian ASBT as it is not completely clarified if this bacterial homolog transporter exclusively transports bile acids.⁴

Since a 9-TM model was built by *Geyer et al.* and *Zhou et al.* (appendix 1), we decided to build our model according to their more recent findings and therefore excluded the first bacterial transmembrane region for our calculations as suggested in their work.^{41,45}

A comparison of the transmembrane topology between the prokaryotic and mammalian ASBT (predicted with seven TMs) can be seen in figure 3 and a depiction of hASBT's secondary structure in figure 4.^{42, 4}

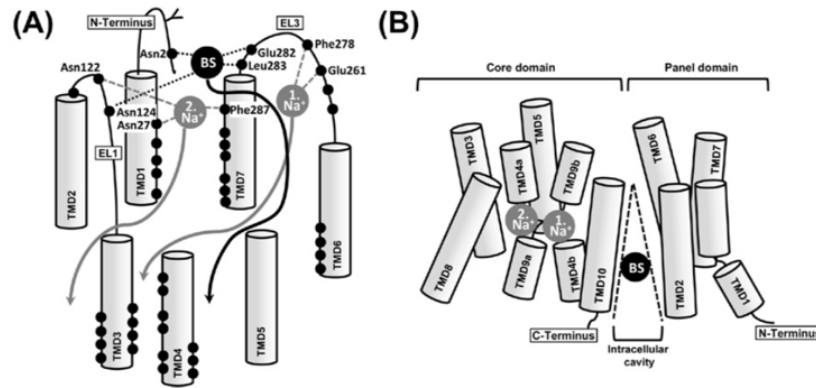


Figure 3: Transmembrane domain topology of hASBT(A) and ASBT_{NM} (B)

(A) Depicting the possible seven TM assembling and assumed interactions of hASBT's amino acid residues with the bile acid substrates
(B) Comparing the 10 TM structure and steric representation from template homolog ASBT_{NM} (derived from its crystal structure) .

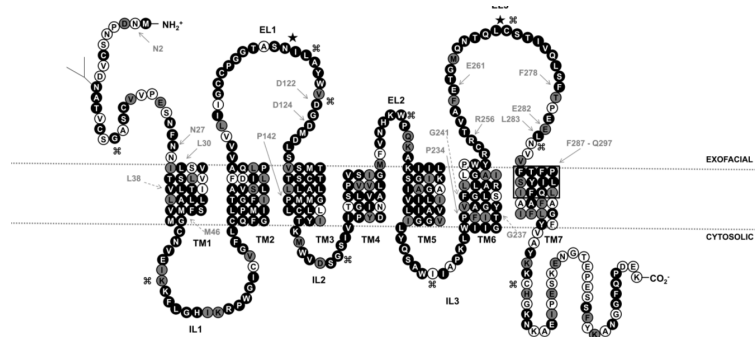


Figure 4: predicted secondary structure of hASBT

The ten bacterial transmembrane (TM) segments are assembled as following described: The first two TM Segments (TM1 and TM2; TM6 and TM7) are forming the typically “**V-motif shape**” and create the **scaffold domain**. The **core motif** consists out of three helices (TM 3 to TM5 and TM8 to TM10) and is forming the transport domain (also see figure 3B).

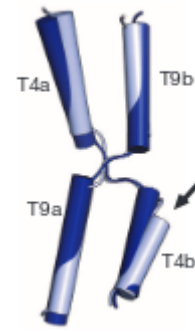


Figure 5: Alignment of crossover region in ASBT_{Yf} (light blue) and ASBT_{NM} (dark blue)

Furthermore, a particular characteristic feature entitled as the “**crossover region**” (figure 5)⁴¹ can be seen, including the discontinuous TMs 4a,b and 9a,b which are intersecting at their breakpoints.

Another known example for discontinuous transmembrane helices is the sodium/proton antiporter NhaA. Interestingly, NhaA and ASBT_{NM} share an RMSD[†] of 2.9 Å, which shows an unexpected similarity of these two independent transporters.^{5,46} Indeed ASBT is not related to NhaA, but they are defined by mutual motifs and a similar fold, which could help understanding the transport mechanism.

As ASBT and NTCP need to bind two sodium ions (Na⁺) in order to translocate one BA molecule, two **sodium-binding sites** (Na₁,Na₂) could be discovered. They are located in the core domain close to the crossover region. These ion-coordinating residues are highly conserved within the bacterial homologs and as well as the mammalian ASBT and NTCP (including human see PROMALS alignment figure 27).^{5,41}

The inspection of the distinct X-ray structures available of the bacterial homologous protein in various conformations, suggested an “**elevator mechanism**”(figure 6) of transport.¹⁰ In this

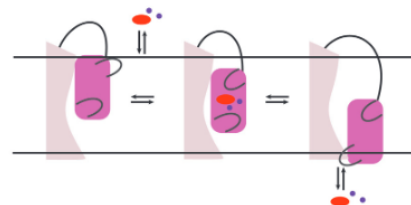


Figure 6: elevator mechanism

The transport domain (pink) is translocating the substrate with a perpendicular movement across the membrane, whereas the scaffold domain (light pink) remains static.

[†] Root-mean-square deviation: a quantitative measure of similarity between protein structures

RMSD = 1-3 Å for similar proteins

model, the protein can be divided in a **transport domain**, moving across to the membrane, containing major parts of the binding site, and a static **scaffold domain**.

For ASBT the conformational change from outward- to inward-facing states is characterized by only a small slide movement of its transport domain as compared to other transporters like NhaA (figure 7 A and B).⁴⁶

As the substrate binds to the transport domain, the conducted perpendicular movement allows the electrogenic driven translocation from the extracellular- into the intracellular-side of the membrane, accomplished with the substrate release.^{5,10,46}

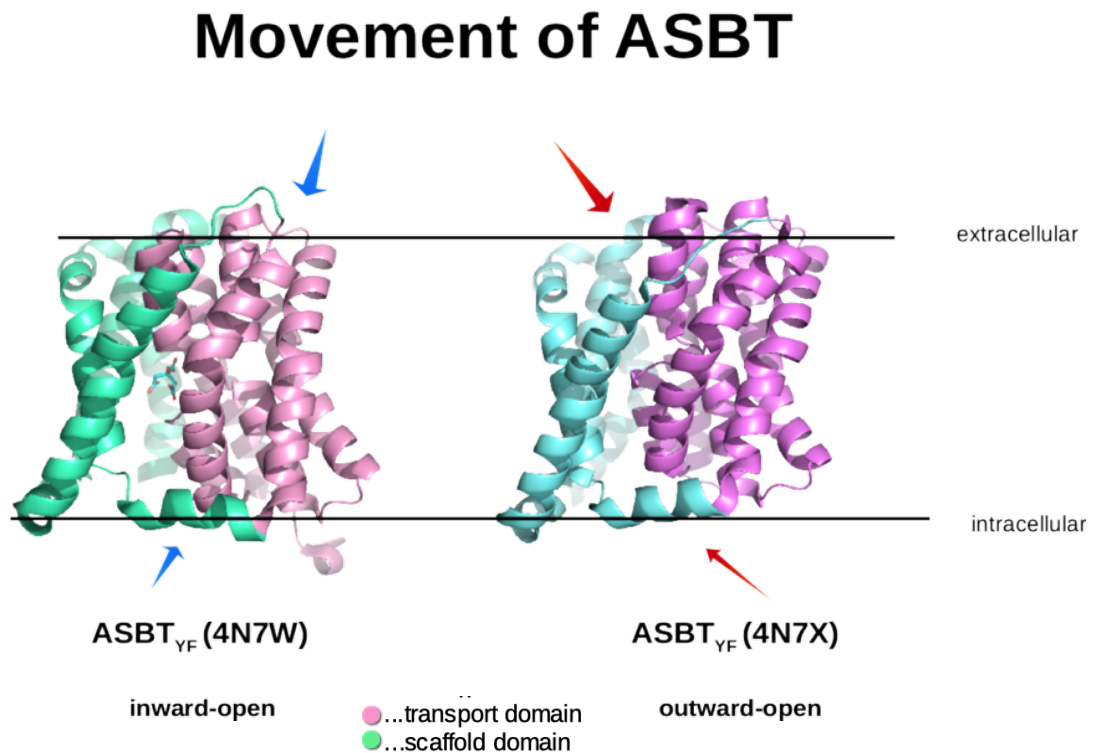


Figure 7A: The movement process of the elevator transport mechanism for the two bacterial homologs. The arrows are indicating the movement of the protein alternating its two structural conformations. The pink transport domain moves across the membrane to aid substrate release, while the scaffold domain (cyan blue) remains rather rigid.

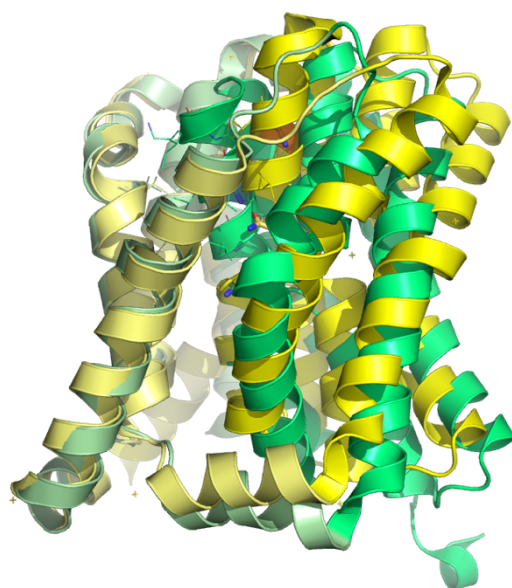


Figure 7B: Postulated movement process of human ASBT with superimposed structures of the inward- and outward-open conformation

Yellow: outward open (modeled from 4n7x)

Green: inward open (modeled from 4n7w)

The proteins are aligned with their statically fixed scaffold domain (pale color) to allow the visualization of the movement carried out by the transport domain (bold colors) between the two states.

4.1.2 Template selection and Alignment

As above stated the bacterial homolog sequences of ASBT_{NM} and ASBT_{Yf} are suitable templates for the homology modeling process. The primary sequence of each protein was downloaded in the *.fasta* format from **UniProt**, a database with a broad collection of protein sequences and further functional information.⁴⁷ The sequences of hASBT (UniProtKP: Q12908), hNTCP (Q14973), ASBT_{NM} (Q9K0A9) and ASBT_{Yf} (tr_C4ST46) were uploaded to **PROMALS3D**, an **alignment web tool**, which combines the structural and sequence information of the known homolog and the targeted protein in order to obtain an accurate predicted alignment of the sequences. As an output a colored alignment with predicted secondary structure and information about the amino acid conservation is obtained (see figure 27: PROMALS3D alignment of SLC10 family members; chapter 4.5 Phylogenetic Relationships).

4.1.3 Model building, assessment and refinement

This obtained PROMALS3D alignment was copied to in a text file in a *.pir* format (figure 8A), which is run with a python script (**build.py**) developed for **MODELLER** (figure 8B), a software for comparative modeling of proteins (homology modeling).¹⁹ First only five preliminary models were built according to each corresponding homolog-pdb file for an initial evaluation. On the bases

of the PROMALS alignment we were able to conduct further manual optimizations such as inserting small shifts, cutting loops or parts of the C termini by editing the *pir* file with a focus on the binding site residues. After each optimization step the DOPE scores were checked.

Figure 8A: Align.pir - Sequences are inserted in .pir format and then further adjusted for an optimal resulting model

```
>P1;asbt
sequence:asbt:1::LAST::::
MNDPNSCDNATVCSGASCVVPE SNFNILSVVLSTVLTILLALVMFS-MG
CNVEIKKFLGHIKRPWGICVGF LCQFGIMPLTGFILSVAFDILPLQAVVLIIGCCPGGTASNILAYWVD
GDMDLSVSMTTCTLLALGMMPLCLLIYTKMWVDS-GSIVIPYDNIGTSLVSLVVPVSIGMFVNHKWPQK
AKIILKIGSIAGAILIVLIAVVG GILYQS--AWIIPKLIIGTIFPVAGYSLGFL LARIAGLPWYRC
RTVAFETGMQNTQLCSTIVQLSF/LNVVFTFPLIYSIFQLAFAAIFLGFYVAYKKCHGKN.*

>P1;4n7w
structureX:4n7w: 1 :A:LAST :A::-1.00:-1.00
----MLVKITRLFPVWALLLSVAAYFRPTTFTGIGPYVGPLLMLIMFA-MG
VTLRLDDFKRVL SRPAPVAAATFLHYLIMPLTAWILAMLFMRPPDLSAGMVLVGSVASGTASNVM IYLAK
GDVALSVTISAVSTLVGVFATPLLRLYVDATISVD-----VVGMLKSILQIVVIPITAGLVIHHTFTKT
VKRIEYPYLPAMSMVCILAIISAVVAGSQ-SHIASVGFVVI IAVILHNGIGLLSGYWGKLF GFDESTC
RTLAIEVGMQNSGLAATLGKIYF/SPLAALPGALFSVWHNLSGSLLAGYWSGKPKVKDQE.*
```

Figure 8B: python script (build.py) which runs MODELLER

```
# Homology modeling with multiple templates
from modeller import *      # Load standard Modeller classes
from modeller.automodel import *  # Load the automodel class

log.verbose() # request verbose output
env = environ() # create a new MODELLER environment to build this model in

# directories for input atom files
env.io.atom_files_directory = ['.', './atom_files']
# Read in HETATM records from template PDBs
env.io.hetatm = True
env.io.water = False

a = automodel(env,
               alnfile = 'aln.pir', # alignment filename
               knowns = ('3zuy'),   # codes of the templates
               sequence = 'asbt')   # code of the target
a.starting_model= 1                # index of the first model
a.ending_model = 100               # index of the last model
                                   # (determines how many models to calculate)

a.assess_methods = (assess.GA341,
                    assess.DOPE,
                    assess.DOPEHR,
                    assess.normalized_dope)

a.make()                           # do the actual homology modeling
```

After the refinement was suitably completed, one hundred models of each conformation and each transporter were built to retrieve a sufficient **variety of binding site conformers**.

In between some of the top scored models were uploaded to **PyMOL**⁴⁸, a molecular visualization program for preliminary inspection (figure 9).

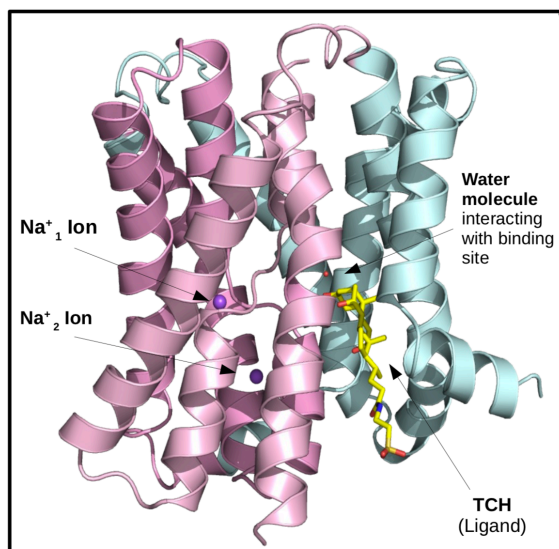


Figure 9: First actual homology model of hASBT modeled from 3ZUY (inward open). Showing the two required sodium ions (purple spheres) for transport and a physiological BA ligand (TCH) ready for release, as well as an included water molecule.

The final decision which of the exploratory models were chosen for further docking studies was on one hand depending on the re-docking results of TCH and finally made by means of the **enrichment process**. For our further investigation steps we initially proceeded with the **inward open** conformation of hASBT modeled from 3ZUY, since the bound substrate TCH was already included in its binding site. This served as a good starting point to examine possible interactions stated in the human protein

4.1.4 Re-Docking of TCH for Binding site refinement

Before calculating the enrichment curve, it was necessary to inspect and prepare the models' binding site in order to ensure mutual matching interactions between the ligand and protein for each complex.

During the literature research we found a **mutation study** that was conducted by *Zhou et al.* transforming putative relevant polar residues of **ASBT_{Yf}**'s binding site to Alanine or Valine. This was done to demonstrate how the loss of polarity from these residues would influence substrate binding of TCH as compared to the wild type. Those residues were expected to be involved in hydrogen bonding with hydroxyl-groups of the substrate Taurocholate (TCH) in a hypothetical horizontal binding pose. Three of the mutations (**T106V**,

N259V, H286A) happen to decrease TCH binding by more than 20%, which suggests that these residues are involved in ligand binding (figure 10).⁴¹

Conservation of important residues

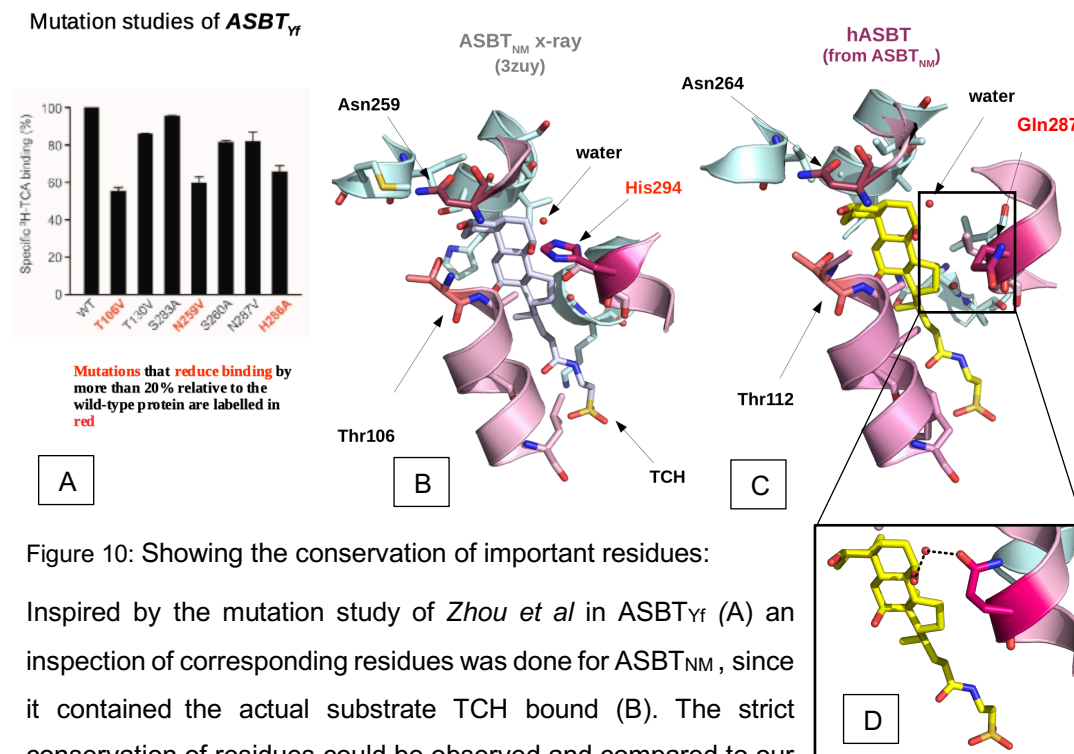


Figure 10: Showing the conservation of important residues:

Inspired by the mutation study of *Zhou et al* in ASBT_{Yf} (A) an inspection of corresponding residues was done for ASBT_{NM}, since it contained the actual substrate TCH bound (B). The strict conservation of residues could be observed and compared to our human ASBT model (C). Binding site of best ranked model with correct oriented Gln287 able to interact via hydrogen bond network with ligand TCH (D).

When inspecting the binding site of ASBT_{NM} (inward open) in PyMOL and according to their alignment, we found out that the residues mentioned were **highly conserved** within the two **homologs** and most of the SLC10 family members (see PROMALS3D alignment – chapter 4.5 figure 27). Moreover, a **water molecule** located in the **binding site** could be observed interacting on one side with the oxygen of TCH via a hydrogen bond and with Histidine 294 (His294) on the other side, serving as a putative **linker**. Therefore we concluded that the **water molecule is possibly involved in the binding process of TCH** and other BA substrates. (figure 10)

Next, when inspecting our model of **hASBT** (template ASBT_{NM}) we found Asparagine (Asn) and Threonine (Thr) to be conserved too, but instead of His

a **Glutamine** (Gln287) was present. This Gln is likewise able to form a hydrogen bond with the above-mentioned water molecule. Since it is **conserved** in 5 out of 7 human SLC10 members (PROMALS3D alignment – figure 27) we hypothesized that this **water is necessary for binding the substrate in the inward open conformation** and thus was included in our models.

Provided with this important additional information we started to narrow down the number of homology models to work with **by measuring the distances** of the hydrogen bond forming **water** and the nitrogen or oxygen of **Gln287**. By choosing only models with a distance smaller than 3 Å (Angström) we wanted to ensure that the possibility for hydrogen bond interactions is given. We ended up with eleven models for the “**nitrogen-distance**”, four models for the “**oxygen-distance**” of Gln to the water molecule and five models with the **best DOPE score of hASBT** (modeled from 3ZUY) in the **inward open** conformation (figure 11). For the **outward-open conformation** the most representative model was chosen according to the highest DOPE score ranking, since no ligand was resolved in the bacterial homolog and barely some information regarding protein ligand interactions was available. A detailed information and values about this selection process can be seen in appendix 2.

After the preliminary distance selection of adequate models, the binding **site refinement step via Re-Docking** started. Here we wanted to ensure that the protein’s binding site was prepared in a suitable way where the known binding pose of TCH could be reproduced. Most importantly, hydrogens were added to the protein by using the Protein Preparation Wizard tool from the Schrödinger suite, among other preparations such as optimization and minimization steps.⁴⁹ Their correct orientation was adjusted using the **Interactive H-bond Optimizer** panel. Arranging the hydrogen bond network and therefore providing a suitable environment was a challenging task, but provided the fundament for placing a valid docking grid.

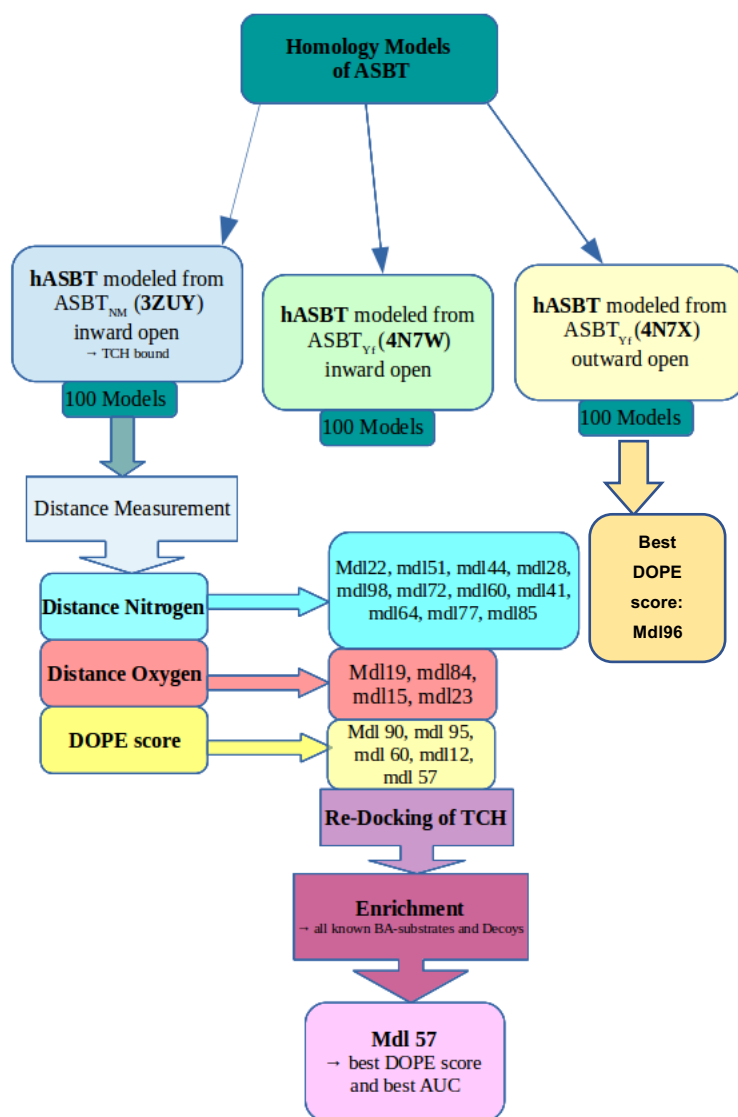


Figure 11:

Created Homology models of hASBT

After measuring the distances only the most accurate models were shortlisted for further investigation and in the end model 57 was chosen as the most representative

homology model for hASBT. For NTCP model 96 was selected according to the best ranking.

NB: The "oxygen-distance" measurement was added afterwards for the reason that as well the nitrogen and the oxygen are capable of interacting with the water molecule. As Gln gets flipped the calculation was added since it was unclear to that point what conformation was prevailing.

After many misleading attempts the **hydrogen bond interaction network** was finally defined by the following way: One hydroxyl-group of **TCH** (labeled as **R₁**; attached to C₃) is interacting with Thr110 by contributing its Hydrogen, therefore serving as a **Hydrogen Bond Donor**. TCH's hydroxyl-group labeled as **R₂** (attached to C₇) is interacting with the water molecule by receiving the hydrogen, and therefore is a **Hydrogen bond Acceptor** (figure12).

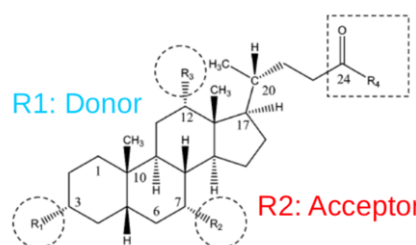


Figure 12: Bile Acid scaffold and function of residues

Moreover, this **water molecule** is donating its hydrogen to **interact with Gln287** to complete the assumed hydrogen bond network (figure 13).

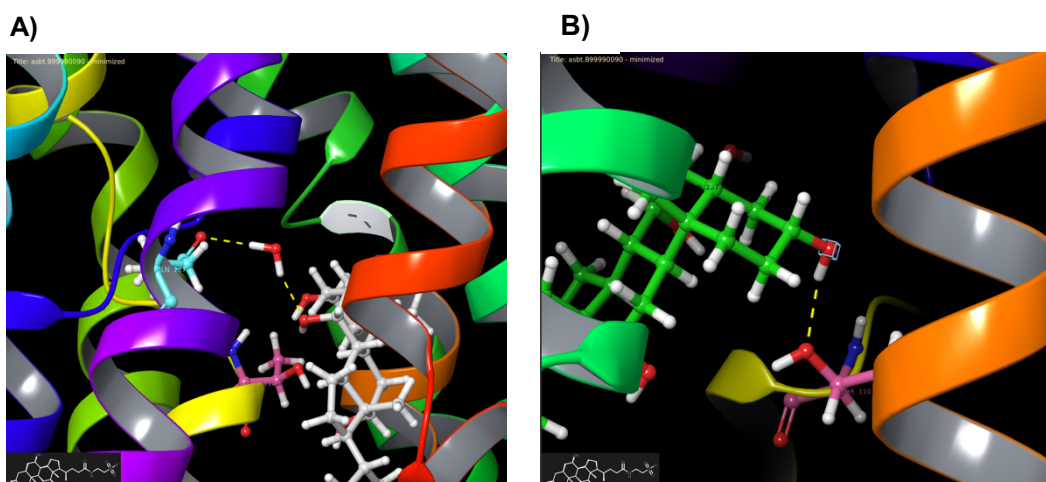


Figure 13: Depiction of hydrogen bond interactions of TCH (grey/green):

A) Hydrogen bond network of Gln287 (blue), water molecule and R₂ of TCH

B) R₁ of TCH interacting as a hydrogen bond donor for Thr110 (pink)

Sometimes, depending on the model, a side chain **rotamer** of the amino acids Thr110 or Gln287 had to be produced by using PyMOL's "Mutagenesis Wizard".⁴⁸ This step was essential to enable hydrogen bond interactions. By means of above-mentioned settings the **grid** for every chosen model was calculated by using the Receptor Grid Generation panel of Schrödinger's Maestro in order to represent the protein's properties needed for docking.

Hydrogen-bond constraints were applied for **re-docking TCH** using the above listed residues: Thr110 via oxygen as a Donor and the water molecule (H₂O) as a Donor via the oxygen (figure 14).

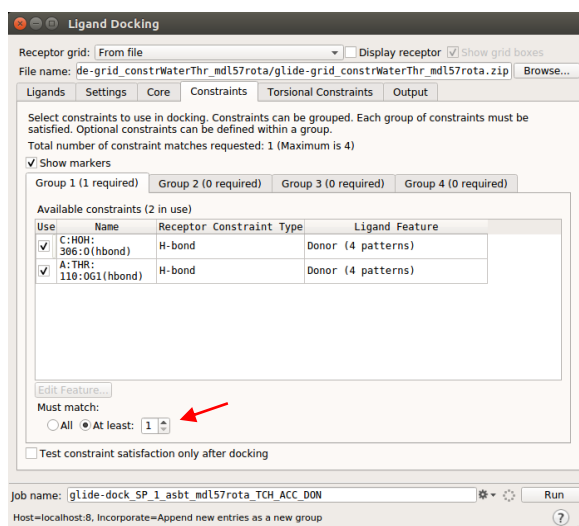


Figure 14: Constraints tab of the Ligand Docking panel: chosen constraints for Re-Docking TCH

After detecting all the basic settings needed, the re-docking of TCH was conducted for every model and then inspected via PyMOL. The constraints and settings for docking got iteratively improved until a valid docking pose was reproduced for every model.

We aimed to generate an environment where choosing at “least one constraint” in the docking panel (figure 14 – red arrow) was enough to accurately dock BAs, because we wanted to bias the docking procedure as little as possible by setting too strict constraints.

For example (figure 15): when **Re-Docking of TCH to Model 98** (Thr-Rotamer) superimposed to the crystal structure binding pose of TCH (grey), only minor differences between the tail positions can be seen, which were considered as nonrelevant at this stage of docking.

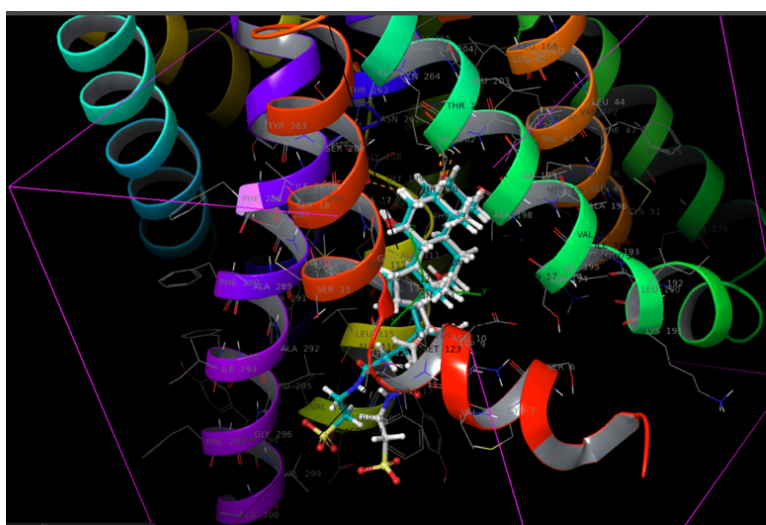


Figure 15: Re-docking of TCH in mdl 98 (inward open)

The re-docked TCH (cyan) is perfectly superimposed to the binding pose of the homologous crystal structure (grey) due to a complex setting of constraints and a profound preparation of our models (*chosen grid space is depicted as a purple square*)

Our conclusion seems to be supported by the QSAR Pharmacophore model created in 1999 by *Werner Kramer* and his group, which contains one hydrogen bond acceptor and one donor feature for rabbit ASBT.⁵⁰

4.1.5 Model validation by Enrichment

As stated before the **enrichment process** is a method for estimating the accuracy of built homology models by assessing the ability to discriminate between known ligands and generated decoys. The calculated enrichment

curve displays the ratio (in %) of ranked known ligands (true actives) in respect to the screened database (decoys + true actives). We docked all known BA-substrates of ASBT according to a review of *J. Geyer* and their originated decoys by DUD E (table 3 – yellow highlighted BA).⁶

The calculations were run with an “in house script” that was firstly automatizing the docking procedure of BA and decoys in Glide, secondly ranked them according to the Glide DOPE score and then calculated the AUC values. Out of the given data **model number 57** performed **best** with an **AUC of 61.607 %** docked with **at least one constraint** (Thr or water) matching (figure16). Interestingly, model 57 holds also the **best ranked normalized DOPE score** of **-0.65323**.

This model was then chosen for further docking studies, since it was the best one to distinguish between true ligands and decoys.

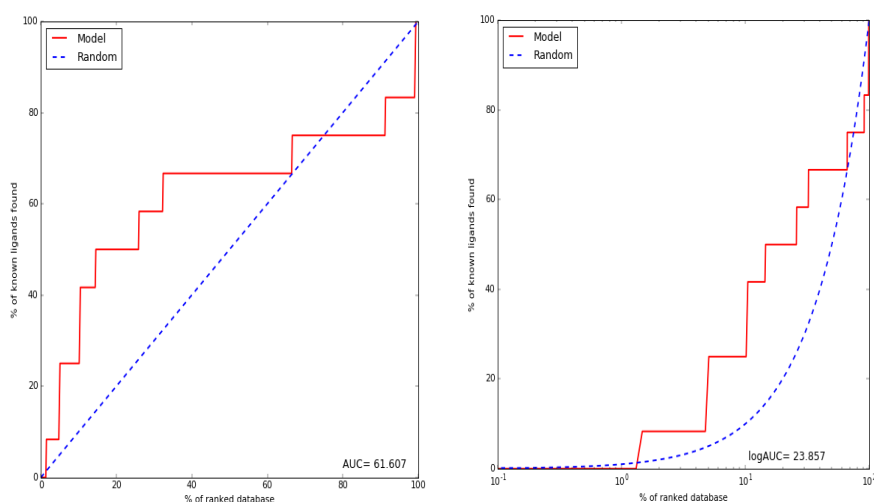


Figure 16: Calculated Enrichment curve of model 57 showing an acceptable result of AUC = 61%, which represents the probability of finding an active better ranked than an inactive.

Moreover our model loses accuracy (worse than random selection) of proper discrimination after screening around 70% of our database.

An AUC value of around 61% was fine for our use, but surely it could be considered to opt for more optimizations of this model, to attain better AUC values. The right panel shows the enrichment curve in a semi-logarithmic scale.

4.1.6 Modeling and preparation of hNTCP

The same steps and procedures of **homology modeling** and **binding site refinement** were executed for **hNTCP**. Likewise one hundred models of each conformation were generated for hNTCP. The **inward open** conformation was as well modeled on the template 3ZUY (ASBT_{NM} with water molecule included), and the **outward open** conformation on the template 4N7X (ASBT_{Yf}).

When inspecting NTCP's **inward open** conformation it was revealed that ASBT's Thr110 is replaced by **Asn103** (Asparagine – N) which is also capable of interacting via hydrogen bonds with the labeled R₁- hydroxyl-group of TCH. The R₂-hydroxyl-group of TCH is, as shown for ASBT, also interacting with the included **water molecule**. Again a hydrogen bond network could be established by connecting TCH's hydroxyl-group (R₂) via the water molecule to NTCP's Gln289.

A necessary reduction of the produced models was done by measuring the **distance of Asn103's nitrogen (hNTCP) to R₁'s oxygen (TCH)**. By setting this threshold we obtained 10 models with a measured distance under 3.2 Å. Furthermore, the **5 best ranked models**, according to the DOPE score, were added to our selection (appendix 2- excel sheet of chosen models).

As done for ASBT the hydrogen bond interactions and grids were prepared as following: The hydroxyl-group of TCH labeled as **R₁** is serving as a hydrogen bond **donor** and TCH's hydroxyl-group labeled as **R₂** (interacting with the water molecule) is a hydrogen bond **acceptor**. When **re-docking TCH** acceptable binding poses could be retrieved in hNTCP (inward open) by using these settings.

4.2 Docking of BA in ASBT inward open conformation

4.2.1 Preparation and Receptor Grids

After the intended accuracy of our models could be observed in the enrichment curve, we started our docking studies with model 57 (inward open hASBT) choosing the same constraints as for the re-docking procedure, the water molecule, and Thr110. Also former protein preparation (Protein Preparation Wizard) and ligand preparation (LigPrep) settings of model 57 were applied. This time **all** known **transported BA** substrates (table 3) with available experimental data from transport assays were selected **as ligands to be docked**.⁶ These calculations were performed in Glide²⁶ with the intention to notice differences in interactions or receive different binding poses which would explain the substrate specificity or selectivity of hASBT.

4.2.2 Docking of ligands – Affinity inward open

Based on our docking results an attempt to explain the experimentally seen **preference of C₇ - α -hydroxyl groups** (labeled as **R₂**) (Cholate, CDC, DC, LC and corresponding conjugates) **over C₇- β -OH** (UDC and conjugates)³ can be made. Taking a closer look at the presence or absence of established hydrogen bonds with this specific water molecule could be a first starting point to understand the **variety and difference in affinity of primary and secondary BAs** (table 3 - experimental K_i/K_m values).⁶

The difference of primary and secondary BAs is most importantly defined by the distinction of a α -hydroxyl group (R₂) attached to C₇ or none. Secondary BAs would either have no hydroxyl-group attached to R₂ (C₇ = H) or UDC and its conjugated derivatives would own a β -hydroxyl group (see figure1) which seems to have a major impact on the transport affinity.³

Table 3: An overview of BA substrates transported by ASBT and NTCP with corresponding K_i or K_m values illustrating differences in specificity between the two transporters and intern substrate affinity
x...no experimental values available

Abbreviation	Substrates	ASBT	NTCP
	Primary Bile Acids		
C	Cholate	$K_m = 33 - 37 \mu M$	$K_m = 6 - 34 \mu M$
GC	Glycocholate	x	$K_m = 27 \mu M$
TC	Taurocholate	$K_m = 12-18 \mu M$	$K_m = 6 - 34 \mu M$
CDC	Chenodeoxycholate	$K_i = 3.3 \mu M$	x
GCDC	Glycochenodeoxycholate	$K_i = 5.7 \mu M$	x
TCDC	Taurochenodeoxycholate	$K_i = 6.1 \mu M$	$K_m = 5 \mu M$
	secondary BA		
DC	Deoxycholate	$K_i = 6.3 \mu M$	x
GDC	Glycodeoxycholate	$K_m = 2 \mu M$	x
TDC	Taurodeoxycholate	$K_i = 17.2 \mu M$	$K_m = 7.4 \mu M$
LC	Lithocholate	n.t.	n.t.
GLC	Glycolithocholate	x	x
TLC	Taurolithocholate	x	x
UDC	Ursodeoxycholate	$K_i = 75 \mu M$	x
GUDC	Glycoursodeoxycholate	$K_m = 24.1 \mu M$	x
TUDC	Tauroursodeoxycholate	$K_i = 28 \mu M$	$K_m = 14 \mu M$
	Bile acid sulfates		
CDC3S	Chenodeoxycholate-3-sulfate	not transported (n.t.)	substrate
TLC3S	Taurolithocholate-3-sulfate	rabbit Asbt $IC_{50} = 9.1 \mu M$	rabbit Ntcp $IC_{50} = 0.8 \mu M$
	Steroid sulfates		
Oest3S	Oestrone-3-sulfate	n.t.	$K_m = 27-60 \mu M$
DHEAS	DHEAS	n.t.	transported

For example, when docking TC, CDC, UDC and LC in model 57 (hASBT: inward open) with mentioned donor(R₁)-acceptor(R₂) settings, and constraints only set on Thr110, the following differences of interaction can be seen (figure 17).

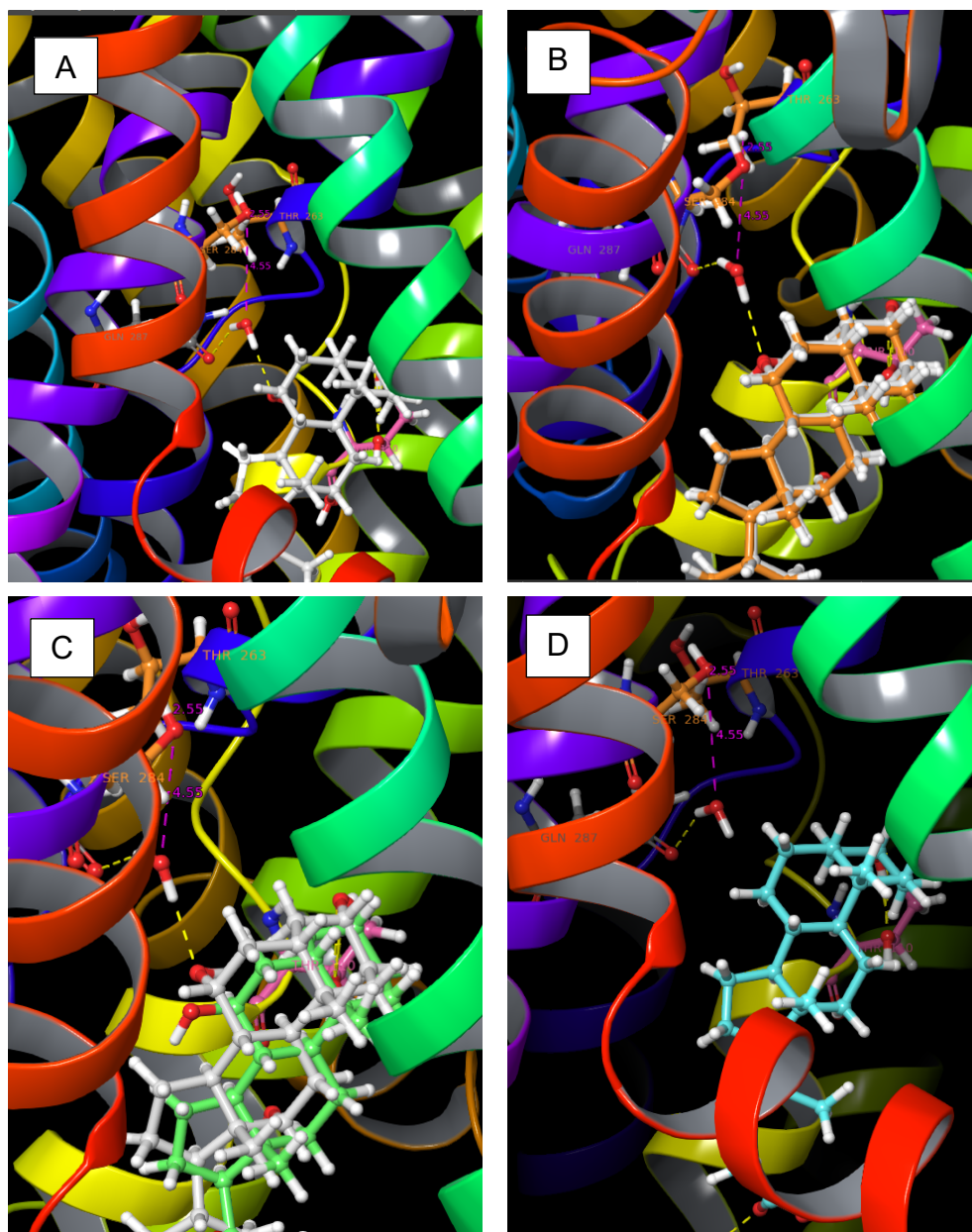


Figure 17: Different interactions with water molecule:

- A)** TC (grey, α -OH at R₂) establishing a hydrogen bond with water molecule
- B)** CDC (orange, α -OH at R₂ - superimposed to TC) forming a hydrogen bond
- C)** UDC (green, β -OH at R₂ - superimposed to TC) not interacting with the water molecule due to β -configuration of hydroxyl-group
- D)** LC (blue, no OH at R₂) also no interaction

TC ($R_1=R_2=R_3= \text{OH}$) forms a hydrogen bond with the water molecule having an $\alpha\text{-OH}$ (R_2) as an acceptor, which could be linked to the experimentally observed high affinity and good transport rate.

CDC ($R_1=R_2=\text{OH}$, $R_3=\text{H}$) is also showing an interaction of $\alpha\text{-R}_2\text{-OH}$ with the water, as well as a low K_i indicating to be a good inhibitor and has a lower transport rate. **UDC** is not able to interact with the binding site water when being docked with the same constraints, attributed to the different steric orientation (tilted away) of the $\beta\text{-OH}$ (R_2). This observation reinforces our hypothesis that an absence of water interaction, in this case due to the $\beta\text{-OH}$ orientation, is decreasing the affinity, as observed from the experimentally measured weak inhibition (high K_i).

LC ($R_1=\text{OH}$, $R_2=R_3=\text{H}$) is as well not establishing hydrogen bonds with the water molecule, because of the absence of the R_2 hydroxyl-group. Interestingly, no transport or inhibition of LC is measured in ASBT.

Together with other steric reasons the discovery of these stated hydrogen bond network interactions could play a crucial role in defining the difference of substrate affinity for ASBT.

Drawing a connection to the observation made by *Polli et al.*³ that CDC exhibits a greater inhibition potential than UDC and considering the experimental values of transport assays for TC, CDC and UDC and our findings could be a starting point of **explaining a difference in substrate affinity by presence or absence of C7- $\alpha\text{-OH}$.**

In the end we could observe that for hASBT the attained binding poses of all BA were found to set up **analog interactions** and therefore created **similar binding poses**. This was also the case for the resulting docking poses of hNTCP. Since no unique or additional interactions could be seen, we concluded that the **substrate selectivity** is very likely to be made in the **outward open conformation**. This assumption would go along with the fact, that molecules are first exposed to the extracellular side of a protein and have to match the needed properties in order to be transported as a substrate.

4.2.3 “Structural water” Hypothesis

Overall, our results show that the hydrogen bond networked formed between the previously mentioned water, the ligands and the binding is very likely to play an important role in the distinct BA affinities. We might even consider it as a **structural water**, meaning it is located inside hydrophobic pockets aiding to stabilize the protein's structure via strong hydrogen bonds.⁵¹ To proof our hypothesis it would be necessary to perform a water analysis through **molecular dynamic (MD) simulations** to distinguish structural from bulk water. Different fluctuation patterns and calculated binding energies would then allow a prediction of structural water, buried in the binding site, or bulk water just located on the protein's surface.^{52,53}

4.2.4 Comparison for NTCP

As apparent from table 3, **NTCP** has a **broader substrate specificity**³ including bile acid sulfates and steroid sulfates (Oestrone-3-sulfate, DHEAS). Since this circumstance was adding more complexity to the transport pattern, slightly different adaptations in the grid preparation for the re-docking had to be applied for hNTCP.

Preparing the hydrogen bonds to obtain a single suitable grid for the “nitrogen distance models” was impossible, so we excluded them and started to refine the remaining ten models with the best scores. A big obstacle was the matter of fact that, whenever using the same settings as for ASBT (TCH: R₁ = hydrogen bond donor, R₂= acceptor) we could, as expected, **successfully dock** the already known **bile acids**, but were **not** able to retrieve reasonable binding poses for **sulfated BAs or steroid sulfates**. This indicates that **two grids are necessary** to acquire valid binding poses for each species of substrate, which we interpret as a non-ideal condition. Due to a limitation of time and available information about hNTCP we paused the enrichment plans of NTCP's inward open models. Moreover, we decided to carry out our docking studies for NTCP (inward open) with the **best ranked DOPE score model 96**.

4.3 Induced Fit Docking of BA in ASBT (outward open)

As previously mentioned due to identical docking results and interactions made in the inward open conformation of hASBT we further hypothesized that the **selectivity has to be defined in the outward open conformation**. As there is no structure of a template in complex with a ligand available in this conformation, Induced Fit Docking was performed to **characterize binding modes** for primary and secondary BAs. Furthermore, due to the flexible approach of both the ligand and protein, in this special docking method we aimed to obtain more precise docking results.

We built our models using the homolog template ASBT_{Yf} (PDB ID 4N7X) crystalized without any ligand bound (apo form). Relying on accessibility data investigating the possibility of a **theoretical horizontal binding site** by Zhou *et al.*⁴¹ we run a binding site search ourselves for additional information (figure 18A). To predict the ligand binding site **FTSite**, an application of the FTMap server, was used to determine and rank the possible areas.⁵⁴ A combination of the pink and green pocket seems to reinforce the above stated hypothesis of a horizontal binding pocket (figure 18B), as there is evidently enough space.

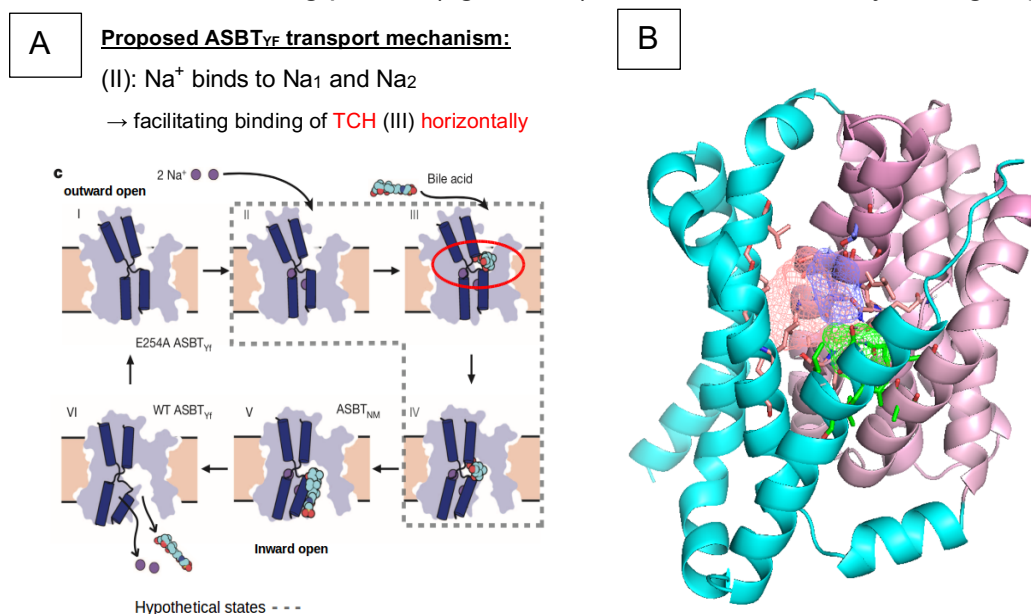


Figure 18: Inspecting the outward open conformation

A) Proposed transport mechanism by Zhou *et al.* suggesting horizontal binding of TCH

B) FTSite binding site prediction of ASBT_{Yf} (4N7X): Three predicted pockets in pink, blue and green with the interacting residues shown as sticks in corresponding colors. Combining the green and pink pocket would support a horizontal binding mode.

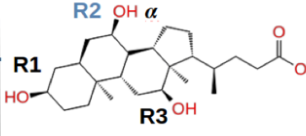
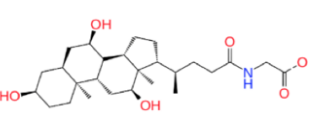
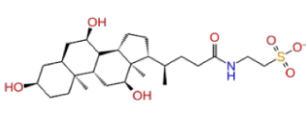
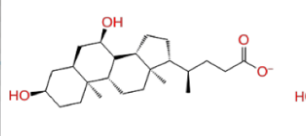
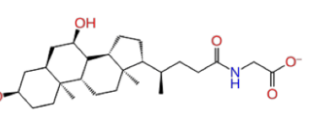
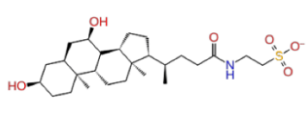
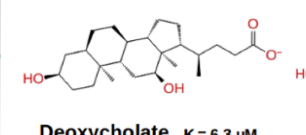
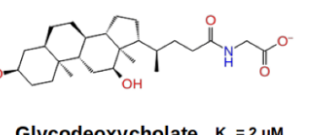
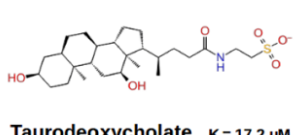
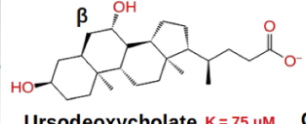
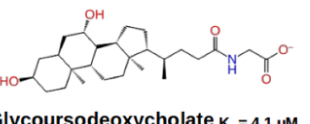
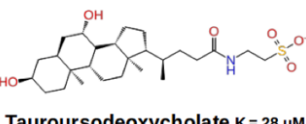
pink: transport domain; cyan: scaffold domain

As mentioned previously, we could not perform enrichment calculations in this conformation and therefore we chose the best ranked model according to the Z-DOPE score.

The grid box for ligand placement was set according to our latest knowledge by choosing the three mutated residues of the template homolog 4N7X (Asn266, Gln287, Thr110) and an **implicit membrane** was added too. This step is intending to simulate the hydrophobic environment on the protein and improves the accuracy of calculations for transmembrane proteins such as ASBT.

We divided the known transported BAs into **four groups** according to their **common hydroxylation patterns and substitution profile** (table 4). This was done as a preparation for the following clustering step of the IDF docked output poses of the substrates. Aiming to compare the existing properties and interactions of every group's best populated cluster among each other.

Table 4: Grouping the Bile Acids

Group 1	 Cholate $K_m = 33 - 37 \mu M$	 Glycocholate	 Taurocholate $K_m = 12-18 \mu M$
	 Chenodeoxycholate $K_i = 3.3 \mu M$	 Glycochenodeoxycholate $K_i = 5.7 \mu M$	 Taurochenodeoxycholate $K_i = 6.1 \mu M$
	 Deoxycholate $K_i = 6.3 \mu M$	 Glycodeoxycholate $K_m = 2 \mu M$	 Taurodeoxycholate $K_i = 17.2 \mu M$
Group 2B	 Ursodeoxycholate $K_i = 75 \mu M$	 Glycoursodeoxycholate $K_m = 4.1 \mu M$	 Tauroursodeoxycholate $K_i = 28 \mu M$

Grouping the bile acids according to their hydroxylation pattern as a prearrangement for the clustering process:

Group 1 ($R_1, R_2, R_3 = OH$): Chol, GC, TC

Group 2 ($R_1, R_2 = OH$): CDC, GCDC, TCDC

Group 3 ($R_1, R_3 = OH$): DC, GDC, TDC

Group 2B ($R_1 = \alpha-OH, R_2 = \beta-OH$): UDC, GUDC, TUDC

Thereby we expected to see differences of interactions for each group and **elucidate the internal difference of substrate selectivity for ASBT and differences of specificity between ASBT and NTCP.**

Additionally, the fact that despite the similar scaffold and properties a noticeable variety of affinities are given within the different BA species, is very interesting (see K_i/K_m in table 4).

4.3.1 Clustering based on Volume Overlap

After the IFD was performed the complexes were clustered based on the **volume overlap** (using the "Clustering Based on Volume Overlap" panel in Maestro).³³ The calculation of the overlapping volume matrix was based on SMARTS[‡] (Simplified Molecular Input Line Entry System) of the common atoms (scaffold of BA and additional common carbon atoms) using single linkage with a fixed atom radius of 0.5 Å. We **clustered** groups 1&2, 1&3 and 1&2B according to their **common binding pose** to compare the impact of different hydroxylation profiles on binding.

From each most populated cluster **one representative binding pose** was selected that met the requirements of being the most common pose and having the best possible ranked score.

Our final aim was to use these selected clustered poses of each group to build **structure-based pharmacophores** in order to depict and rationalize the substrate specificity and to screen databases for new compounds in the future. The procedure of pharmacophore building will be the topic of the next chapter.

4.3.2 Trend in orientation of clustered poses

After the clustering on common binding poses was finished, it was revealed that **individual binding poses** could be observed for primary and secondary BAs. In particular two yet unnoticed amino acid residues were brought into our focus: It could be observed that **Thr267** and **Ser294** are establishing

[‡] A language for describing molecular patterns

hydrogen bonds with hydroxyl groups of **R₁ (C3)** and **R₂ (C7)** of **primary BAs** leading to an “**upstanding**” position of the steroid scaffold. For **secondary BAs** only **one hydrogen bond** was established with **R₂ (C7)** inducing a “**downward bending**” of the steroid scaffold (figure 19) .

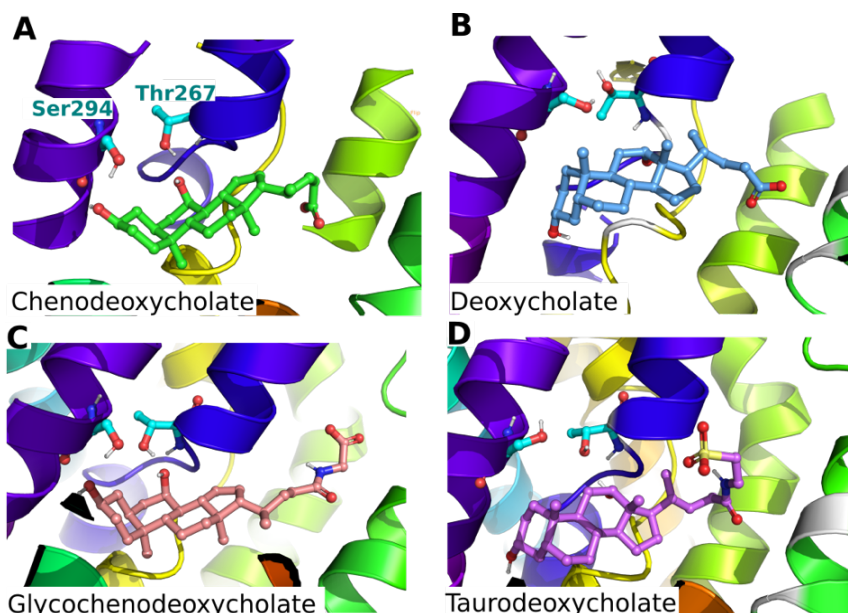


Figure 19: Clustered poses of IDF docking: primary (A,C) and secondary (B,D) bile acids docked in hASBT outward open conformation. A clear trend of orientation can be seen either being tilted “up” or “down” depending on established hydrogen bonds.

Taking a look at the **inward open** conformation of hASBT we saw that **Thr267** is **interacting with Ser290** via hydrogen bond, assuming to act as a kind of **lock mechanism** to fasten the inward open state and possibly aid the release of the substrates.

4.3.2.1 Selectivity: “Ser-Thr-Lock Theory”

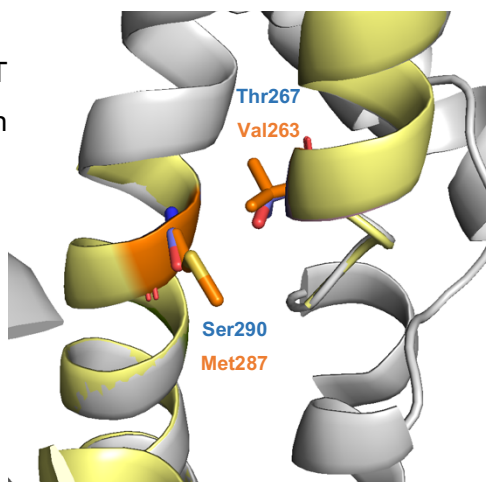
The homolog templates contain as well two hydrogen bond forming Ser residues located exactly at these positions. This lead to our initial hypothesis of the “**Ser-Thr-Locking mechanism**”. We are trying to correlate the internal selectivity profile of hASBT with this dynamic mechanism of BAs interacting with Ser and/or Thr in accordance with the assumed binding position. Interestingly, when paying attention to the conservation of these residues within the SLC10 family many substitutions can be seen: **Thr267** is replaced

by Val263 (for hNTCP) and Ile (hSOAT). **Ser290** is substituted by Met287 (for hNTCP) and a Gly (for hSOAT SLC10a6– not shown) (figure 20).

We first suspected that the variation of these residues and their attributed properties (hydrophobic, polar...) could be, among other mentioned reasons, a trigger for the different selectivity between ASBT (SLC10A2), NTCP (SLC10A1) and SOAT (SLC10A6).

Figure 20: Comparing residues of hASBT (grey) and hNTCP (yellow) in outward open position:

<u>ASBT</u>	→	<u>NTCP</u>
Threonine 267	→	Valine 263
Serine 290	→	Methionine 287



However, this theory needs to be further investigated, because when trying to transfer our plausible docking poses from hASBT (figure 19) to hNTCP the bulky Met287 was obstructing the binding site. BAs could not identically bind to the assumed horizontal binding pocket, requiring further comprehensive analyses.

As well we hypothesize that the **substrate specificity** of ASBT and NTCP is ascribed to an interplay of **BAs configuration** (quantity of OH; α or β position), **interactions** with **transporter specific amino acid residues** essential for substrate recognition and resulting **spatial limitations**. A combination of these factors are, from our point of view, likely to determinate whether bile acids, sulfated BAs or steroids are being transported or not.

4.3.2.2. Affinity calculations of clustered poses

Due to the fact that our IFD poses for the outward-open conformation could not be validated by enrichment yet, we wanted to see if the calculated binding

affinities of our selected IFD poses were correlating with available literature data and therefore would reinforce our docking output. In fact we tried to confirm the placement of our poses with a **correlation of calculated binding affinities** between receptor and ligand (ΔG - *delta G*) and empirical **K_i/K_m values**. To add higher confidence to our estimation the **binding free energy** calculations were run with three different software packages. We used MM-GBSA calculation of Schrödinger's Prime⁵⁵ via command line, LigandScout's Interactive Binding Affinity Estimation⁵⁶ panel (binding affinity surface score) and BioSolveIT's SeeSAR⁵⁷ calculated affinity values for our chosen poses. Prime uses for calculating free **binding affinities** (ΔG - *delta G*) the well-established method of **MM-GBSA** (Molecular mechanics with generalized Born and surface area solvation) calculation. The estimation of binding affinities (in kcal/mol) is based on the interaction properties of the ligand and protein by subtracting the sum of calculated energy of the ligand and receptor from the calculated energy of the complex.⁵⁸

Subsequently, through comparing the obtained data we wanted to be able to see if the calculated binding energies go along with the measured values, which would then confirm our docking poses and thus reinforce our theory of substrate affinity.

Overall our computed ranking of the rated BAs is not entirely corresponding to the known affinity values of transport, but LigandScout's binding affinity score was so far the most reliable parameter for ranking the substrate's affinity as shown in table 5. The obtained MM-GBSA values are fitting into the expected threshold of inaccuracy, but SeeSAR's ΔG values seem to be distributed randomly. For example GCDC, a secondary BA, having the second lowest K_i in this table (5.7 μM), is ranked third place by MM-GBSA calculation (-77,242 kcal/mol) and got accurately classified to a "mM to μM "-affinity-range by LigandScout's binding affinity score (-22,74) (figure 21). But SeeSAR's calculation of a K_i around 6068 μM differs greatly from the experimentally evaluated data.

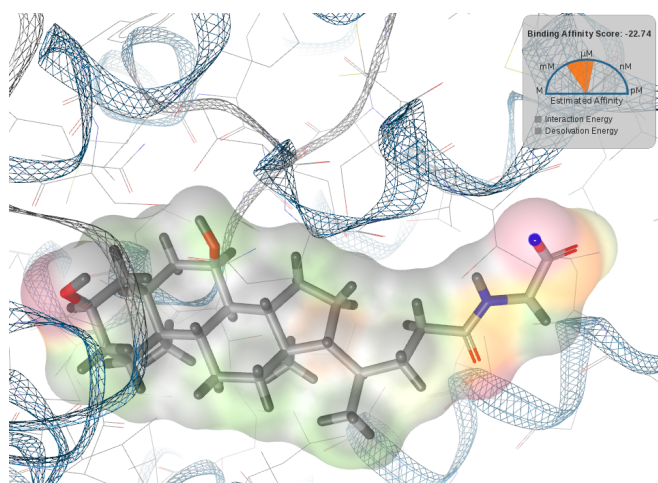


Figure 21: Interactive Binding Affinity Estimation of GCDC. LigandScout was used to depict the binding affinity surface of the IFD pose of GCDC in hASBT outward open and calculate a binding affinity score (see grey square). The estimated affinity is also accompanied with a barometer indicating the range of affinity (molar to picomolar).

Table 5: Merged output of affinity calculations of ASBT (outward open, with implicit membrane) compared to experimental values of transported BA substrates. The ranking of the BA is made according to their measured K_i/K_m values

Red value ...lowest affinity; green value...highest affinity

Primary BA are marked with a purple background color, secondary BA with cyan blue color

NB: indicating the smaller the value, the more affine; K_m and K_i values have been separated since they cannot be compared.

ASBT outward open (membrane)						
Maestro -Prime clustered poses	Experimental K_i (μ M)	MMGBSA ΔG	Ligand Scout Binding affinity score	Range of unit	GLIDE gscore	SeeSAR Mean (μ M)
CDC_gr1_2	$K_i = 3.3$	-73,713	-10,42	M to mM	-11,597	4557,7
GCDC_01_gr1_2	$K_i = 5.7$	-77,242	-22,74	mM to μ M (best)	-11,576	6086,1
TCDC_Gr1_2	$K_i = 6.1$	-52,466	-20,49	(mM to μ M)	-12,222	6827700,7
DC_new_01_gr1_3	$K_i = 6.3$	-77,696	-10,11	mM	-10,457	209356,7
TDC_01_gr1_3	$K_i = 17.2$	-92,758	-5,26	M (to mM)	-12,253	3241,1
TUDC_02_gr2B	$K_i = 28$	-84,002	-0,51	M (worst)	-10,545	4171102,6
UDC_new_gr2B	$K_i = 75$	-73,718	-2,67	M (second worst)	-11,175	23387,2
GDC_01_gr1_3	$K_m = 2$	-83,228	-8,95	M to mM	-12,217	79,7
TC_01_gr1_2	$K_m = 12-18$	-94,080	-9,93	M to mM	-12,182	862981
TC_01 Gr1_3	$K_m = 12-18$	-94,080	-9,93	M to mM	-12,182	862981
TC_01 Gr2B	$K_m = 12-18$	-94,080	-9,93	M to mM	-12,182	862981
GUDC_01_gr2B	$K_m = 24.1$	-91,615	-9,08	M to mM	-12,936	43036705,6

However it must be pointed out that it is very complicated to draw the connection between assessed K_i/K_m values and computer calculated binding energies especially with the aim of further application as a prediction method. It has been shown that for example MMGBSA values are able to retrace and comparatively rank affinities of congeneric molecules, but are not able to calculate absolute true values.⁵⁹ As well it has to be mentioned that surely one limiting factor here and for other predictions is the accuracy of our developed homology models, as all our conclusions are strictly depended on their qualities and correctness.

Finally our hope was to establish a generalizable protocol or finding a suitable prediction tool to confidently assess binding poses e.g. for inhibitors or yet unknown substrates. Creating this workflow couldn't be fully accomplished and will surely require more time, rigorous improvement and probably adaptations when transferring it to hNTCP. Once achieved this will be a handy method for affinity ranking to confirm predicted binding poses and ultimately will allow a better classification of new substrates or inhibitors.

4.3.3 Heat map and PLIF

As prior mentioned we used the PLIF tool to transform 3D protein-ligand interaction data into a "structural interaction profile", and then finally to summarize these results via a heat map.^{35,60} In this heat map a color code - increasing in intensity - represents the corresponding quantity of ligands involved for this residue specific interaction. This was done to point out the most important residues for binding and to visualize the prevailing kind of interactions.

For generating PLIFs we used all IFD protein-ligand complexes calculated from our ASBT outward open model, not only our selected clustered poses (mentioned in 4.3.1 Clustering based on Volume Overlap). By doing so we wanted to avoid biasing the outcome of preferred interaction patterns through pre-selecting poses.

Figure 22 shows the **polar interaction matrix** of hASBT depicting the count of ligand and residue interactions. Likewise the interaction matrices of hydrophobic, hydrogen bond donor and acceptor interactions have been closely inspected to ensure the integrity of the heat map.

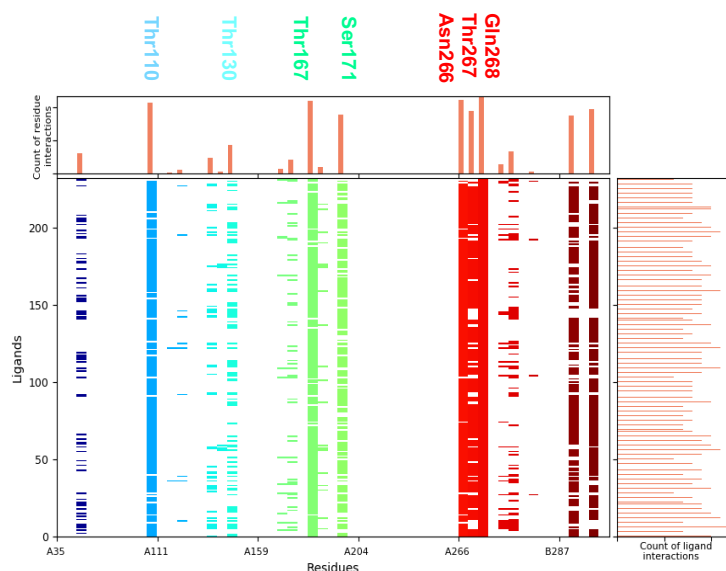


Figure 22: PLIF of Polar Interaction matrix

Only common interactions displayed (of the binding site) above the average count number have been taken into account for visual inspection, but all visible interactions have been included for the heat map creation. As striking interactions classified would be Thr130, Thr167, Ser171, Asn266, Thr267, Gln268.

NB: Thr110 got excluded, since it was used as a constraint for docking and as well aa residues over count 280, since they were not part of our model.

According to the frequency of common interactions stated in this heat map (figure 23) we tried to trace back the function of extensively mentioned residues by comparing them to the primary sequences of our PROMALS3D alignment (figure 27) to enhance our structural information.

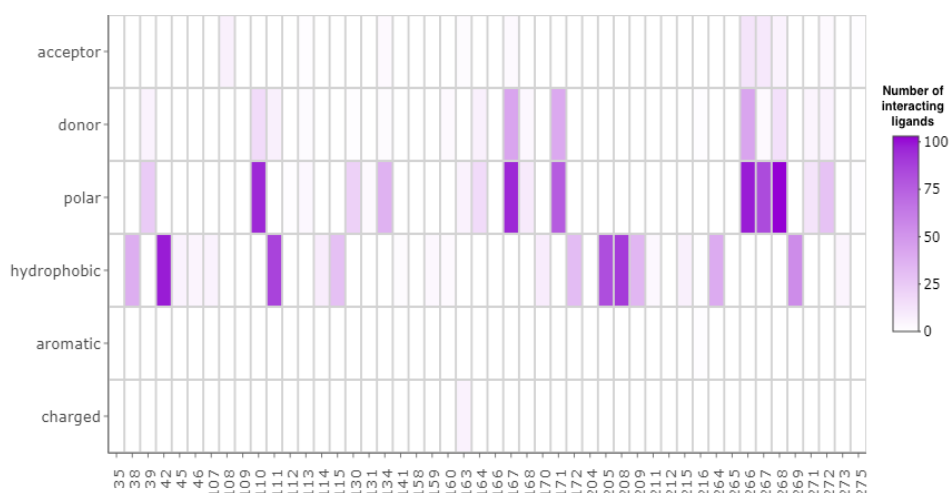


Figure 23: Heat map of ASBT (outward open) illustrating the common protein-ligand interactions. The x-axis shows the residue number of the amino acids and the y-axis displays the kind of interaction stated. Residue Thr267 e.g. states polar interactions with over 100 ligands of our IDF docked protein-ligand-complexes, pointing out that this specific residue could be of great interest in respect to specificity.

The greater the count of ligand involved in interactions, the darker violet the color scale.

Following residues were particularly frequently represented in the heat map and should be mentioned (table 6 A and B):

Table 6 A and B: Summary of important residues of heat map and known related function or further information. Residues in **bold** establish a great amount of interactions and need to be closely investigated.

* suitable for mutation to proof importance for interaction

Table 6 A		Acceptor	
		Gly108 crossover region, highly conserved	Asn266 crossover region, highly conserved
		Thr267	-
		Donor	
		Thr167 * interacting with tail !	Ser171* interacting with tail → not conserved ! (Leu for NTCP; Cyst for SOAT)
		Asn266	Gln268

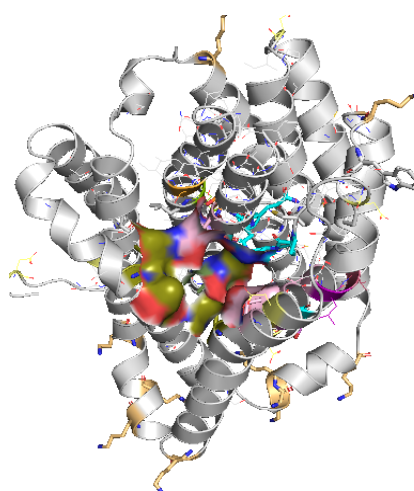
Polar		Hydrophobic	
Thr39	Thr130	Leu38	Leu42 conserved (not SOAT)
Thr134	Asn164	Ala111	Leu115
Thr167 interacting with tail	Ser171* Interacting with tail	Leu172	Ile205
Asn266 Conserved strictly	Thr267* Involved in Ser/Thr binding theory	Ile208	Ala209
Gln268	Thr272 conserved	Met264	Leu269 conserved (not SOAT)

Table 6 B

Concluding it can be said that, when looking at these interacting residues mentioned in table 6 in PyMOL (depicted as a surface), we get a rough first impression of the binding pocket's properties. The possible horizontal binding site could be divided in a **polar region**, where hydroxyl-groups of BAs interact and a **lipophilic region**, where the conjugated C₂₄-tail region establishes interactions (figure24).

Figure 24: PyMOL session of major interacting residues – ASBT outward open (grey):

Indicating a polar interaction area (*olive green*) for the “head region” (hydroxylated BA-scaffold) and a lipophilic area (*pink*) interacting with the C₂₄-tail hydrocarbon chain of the BAs “tail region”.



Interestingly, interactions of previously mentioned crucial residues could be observed: For example **frequent interactions** could be detected in our heat map for **Thr267**, which we are speculating to be involved in our “**Ser-Thr-Lock Theory**” of **ASBT** (chapter 4.3.2.1), giving us another evidence of the significant position of this amino acid. Therefore we would suggest a mutation

of the latter mentioned residue and investigation of the influences on binding, in order to proof our concept of the “**Locking mechanism**”. Also mutations of Ser171 and Thr167 would provide useful information, since these residues are interacting with BAs tail regions and we further hypothesize that here is also a mechanism of determining selectivity and affinity.

Still an ongoing process is the generation, evaluation and comparison of established linkages of IFD docked BA substrates for NTCP in the outward open conformation. Later the same procedure could be applied for known inhibitors to unravel the structure-activity relationship. By carving out the difference of common interactions between BA and inhibitors we would gain further information about the mechanism of affinity and selectivity for hASBT and hNTCP.

4.4 Ligand-based Pharmacophores of known ASBT substrates

As stated before, generating a **ligand-based Pharmacophore** is a handy way of identifying and characterizing the ligand’s steric and physicochemical features necessary to interact with the target protein and cause pharmacological effects. Thereby we expect to **understand the crucial features** needed for **interaction** on those models, that could be further used for **virtual screening** to discover and rank **new ASBT-substrates**.

Our **training set** consists of the selected **representative pose** for every group out of the **clustered IFD Docking** complexes for model 53 outward open (mentioned in 4.3.1 Clustering based on Volume Overlap). These receptor-ligand complexes were uploaded to Schrödinger’s application Phase.³⁹ Then it was assured that all ligands were **aligned** according to their **maximum common structure** by using the “Superposition panel”. For every member of each single group one structure-based Pharmacophore was created using the “Develop Pharmacophore Model” panel (Receptor- Ligand complex). Afterwards those selected features were merged into **one general Pharmacophore**, representing each group’s shared attributes. One big advantage of using receptor-ligand complexes was to obtain so called “**excluded volumes**” outlining the protein-occupied space, where

consequently no features can be placed. By doing so we expect our Pharmacophore models to gain selectivity.

Our aim was to **rationalize substrate specificities of hASBT by using Pharmacophores** and to further clarify the differences of affinity within the individual BAs by comparing their associated “group-pharmacophores”, assuming that exposed differences or similarities of chemical and steric features could lead to complementary perception of interaction. Likewise, ligand-based pharmacophores could be built for hNTCP’s known substrates and then representative features of both transporters could be compared in order to rationalize substrate specificity.

For each group a so called “**ePharmacophore** model”³⁸, an automated method to determine which features contribute essentially to the binding process, and one with **manually** chosen features was generated to compare the influences of different chosen chemical attributes. This double procedure was a necessary step intended as an internal validation as during the process we noticed that sometimes a questionable amount and/or placement of the features was set by the integrated predicting algorithm of the Pharmacophore panel. For example, once only one single feature was chosen to define the whole molecule’s steric space and chemical features, which is understandably too little information to characterize the required 3D pattern responsible for ligand-protein interactions.

Every single developed pharmacophore hypothesis was **validated** via the “hypothesis validation step”, scoring the model’s ability to differentiate between known substrates and generated BA decoys (from DUD-E). More precisely, the *Phase Hypo Score* along with the *BEDROC*, a parameter measuring the model’s performance in discriminating ligands from decoys, was considered to assess the quality of performance.

An example for the evaluation steps (enrichment) of group-2-pharmacophore can be seen in figure 25.

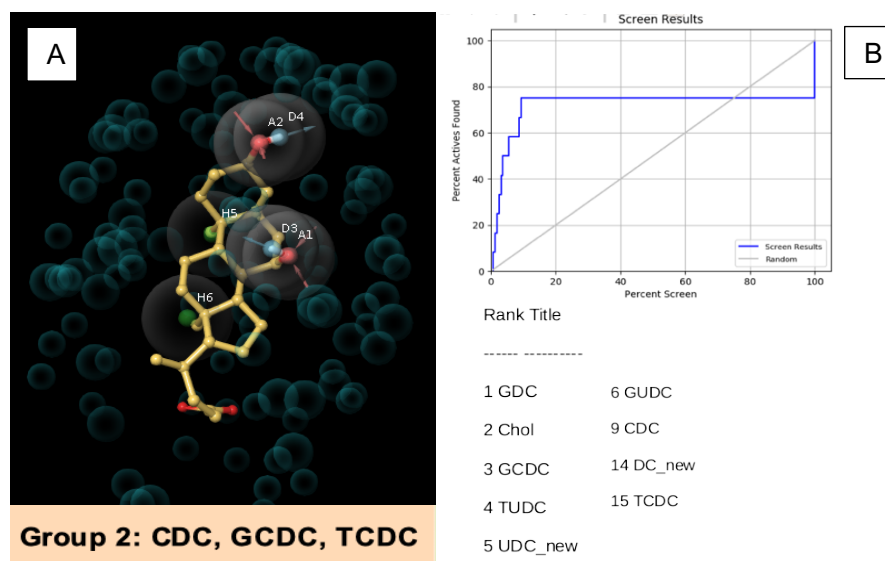


Figure 25: Depiction of Group 2 -assigned Pharmacophore model (manually chosen).

A) Generated Pharmacophore with following chosen features: A/D, A/D, H, H.

Red spheres represent a hydrogen bond acceptor (A), blue spheres symbolize a hydrogen bond donor (D)-feature, green spheres represent hydrophobic features (H). The light blue spheres around the molecule represent the excluded volumes defined by the molecular environment, calculated to prevent steric clashes.

B) The hypothesis validation graph, underneath the list of ranked BAs (rank number) over decoys. It is aimed to score better than the random selection of molecules (vertical grey line).

Figure 26 shows the final version of our four manually generated pharmacophores per group, and the associated features are listed in table 7. We agreed on optimizing and progressing our studies with the manual chosen pharmacophores, since they were more accurate in ranking the BA during the validation process than the automatically generated ePharmacophores (see appendix 3 for the inadequate selected features of the ePharmacophores).

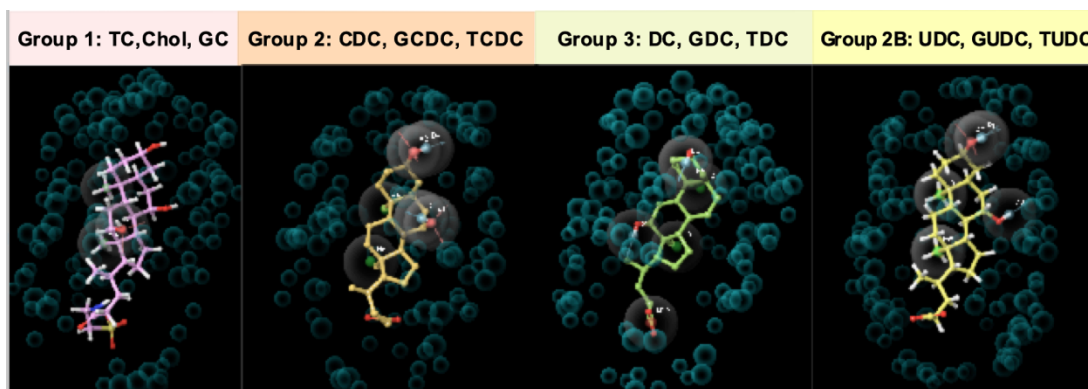


Figure 26: The four finalized pharmacophores depicting the evaluated features needed for the grouped BAs to interact with hASBT.

Table 7: Table of the manually chosen features to depict the steric and chemical features supposed to be needed to interact with hASBT in the outward open conformation.
abbreviations: *Acc...*Acceptor, *Don...*Donor

Manual Pharmacophore	R ₁ (C ₃)	R ₂ (C ₇)	R ₃ (C ₁₂)	Tail (C ₂₄)	note
Group 1: (Cholate, GC) TC	x	x	Acc	x	x
Group2: CDC, GCDC, TCDC	Acc&Don	Acc&Don	x	x	2 hydrophobic features
Group 3: DC, GDC, TDC	Don	x	Don	x	2 hydrophobic features
Group 2B: UDC, GUDC, TUDC	Acc&Don	Don	x	x	2 hydrophobic features

After all proposed hypotheses have been validated and the enrichment ranking of the known BAs was observed, concerns were raised whether these generated pharmacophores were able to represent the 3D patterns needed to cause a pharmacological effect in a correct way. Therefore we suspect them of **being not suitable for screening** new ASBT-substrates yet. Firstly we reckon these models of being too general, since they were wrongly ranking “non-group-BA-members” as hits during the validation process. Furthermore, a majority of the pharmacophores could not reliably assign and rank substrate members of their associated group. This can be partly explained with the remarkably similarity of the chosen BA molecules for the training set, as only minor differences causing them to be divided into different groups. Although

we are strongly convinced that it is possible to characterize and rationalize the determining factor responsible for substrate specificity by using a ligand-based pharmacophore approach, we suggest that our procedure needs to be reconsidered and maybe a modified approach by for example using a different software. Moreover, the grouping of the active ligands (training set) should be reconsidered or adjusted, in order to simplify the selection of features.

4.5 Phylogenetic relationships within the SLC10 family

Phylogenetic analyses have shown to be a powerful method of gaining overall information of evolutionary related proteins. Therefore a multiple sequence analysis of different homolog proteins is performed. Structural information can be derived from distinct patterns as strictly **conserved residues**, rather required e.g. for structural stability or a certain functional role (ligand interaction sites, catalytic site). Often these residues are involved in ligand-protein interactions. Since nature is known for re-using patterns that have proven their functionality, various information can be received from detecting similarities of structure or sequence of related proteins. Highly **variable residues** can be a sign of protein specific properties, causing the difference of substrate specificity for instance evolved through various evolutionary differentiation processes.^{61,62}

With respect to the sequence analyses and phylogenetic relationship studies of Geyer *et al.*⁶ and Ming Zhou *et al.*⁴¹ we started to analyze our performed PROMALS3D alignment in more details, looking for regions ascribed with distinct functions. The sequence alignment was conducted containing the primary sequences of the two homologs ASBT_{NM}, ASBT_{Yf}, hASBT (uniport ID: Q12908), hNTCP (uniport ID: Q14973), hSOAT (SLC10A6, uniport ID: Q3KNW5) and their complementary mammal species (rat, mouse, pig, rabbit, horse). The idea was to **track the conservation** of residues and patterns within different species and thereby to **identify residues involved** in the **substrate recognition and translocation process**.

ASBT and SOAT are the two most homologous members of the SLC10 family (sisters within clade I)⁶ having a sequence similarity of about 70%. Therefore we hoped to find out the reasons why SOAT is able to additionally transport steroid sulfates, while ASBT is strictly bound to non-sulfated BAs.

NTCP, a more distinct relative to ASBT (still in clade I), with about 63% sequence similarity, has a broader spectrum of transported substrates (sulfated) than ASBT.⁶ As stated many times before in this thesis, NTCP is another helpful target for comparing the differences in structure related substrate specificity.

A great number of mutations (SNPs) and structural important residues mentioned in different sources have been collected and added to our generated PROMALS alignment¹⁶ (figure27 page 1 and 2) as a starting point of the sequence analysis. Then systematically all the gained information was used to draw structural and functional conclusions out of the given alignment and our previous findings.

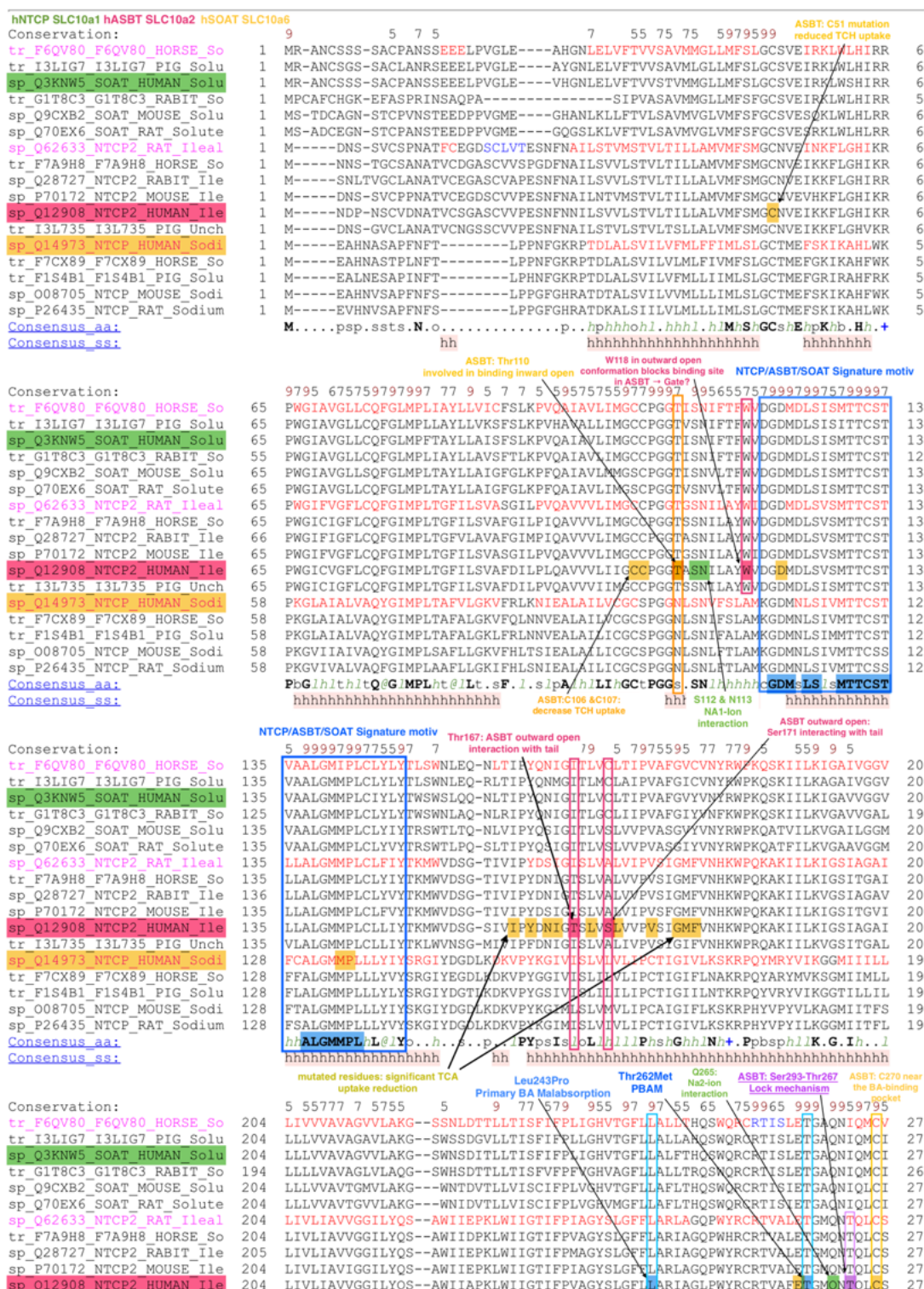
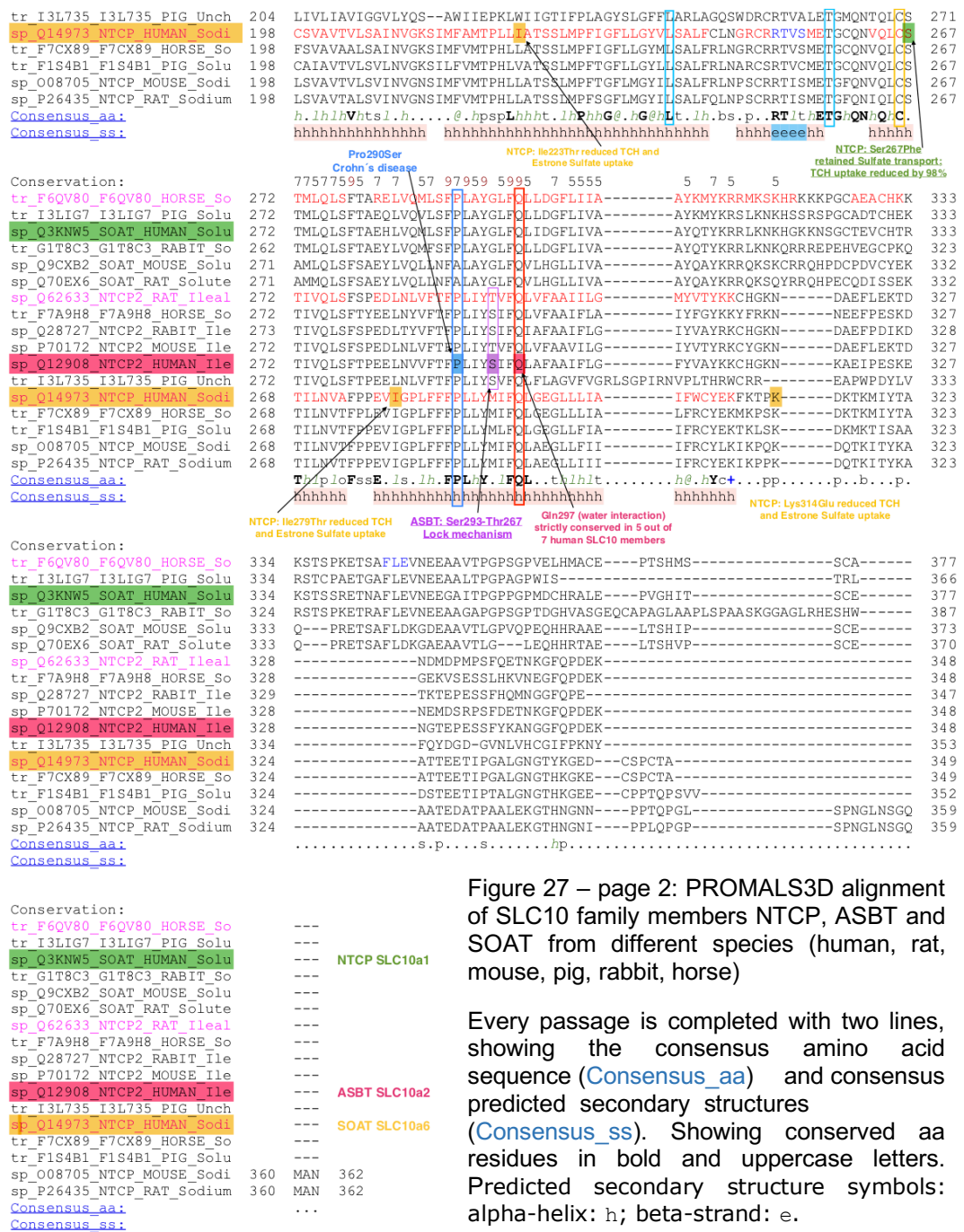


Figure 27– page 1: PROMALS3D alignment of SLC10 family members NTCP, ASBT and SOAT from different species (human, rat, mouse, pig, rabbit, horse).

According to the manual: Representative sequences are marked with **magenta** names and are colored according to predicted secondary structures (**red**: alpha-helix, **blue**: beta-strand). If the sequences are listed in aligned order, other sequences underneath a representative sequence are part of the same pre-aligned group.

NB: residues mentioned in previous chapters may differ in their numbering from the seen residues here this is the case, because during the refinement process some adaptations were made and residues got cut. Therefore in this alignment e.g. Gln297 (uncut version) equals Gln287 (cut version) mentioned in chapters above.



The following observations could be made:

Since the option “aligned” was chosen for the input sequences, closely related sequences will be placed next to each other, confirming the above-mentioned close relation of ASBT and SOAT.

Overall, mutations in strictly conserved regions are generally leading to severe consequences for the affected individual. For example, a single

punctual mutation (SNPs) observed in ASBT's sequence (Leu243Pro or Thr262Met) is associated to **Primary Bile Acid Malabsorption (PBAM)**. PBAM is a disease leading to major gastrointestinal issues and an increased fecal BA excretion due to the fact that ASBT's transport function is impaired.⁴ Furthermore, a "**signature motif (ALGMMPL)**" (aa position 137-143) is **strictly conserved** in ASBT, NTCP and SOAT through the species. Its function remains undefined so far, but it is speculated that specific residues obtain a key function for transport and membrane expression. More specifically cysteine mutagenesis for **Met141** and **Pro142** resulted in a decreased uptake of TCH, indicating **a crucial role for the functionality of ASBT**.⁶

Moreover it was revealed that the membrane-bound P142 is likely to pair with G139 (grouping GxxP as a motif; x...any aa), probably to initiate a Proline-induced structural helix-change (helix packing) (figure 28 A). Pro142 is conserved within many species from SLC10A1 to SLC10A6, hence an important role of latter residues concerning the transport cycle of ASBT cannot be denied. Additionally it is worth to mention the negatively charged Asp122, preserved within the species of SLC10 family and putatively interacting with sodium. Asp124 is conserved within ASBT and SOAT and could affect substrate (BA) binding. All these mentioned structural predictions by *Swaan et al.*⁶³ are based on the seven transmembrane (TM) topology for ASBT and could slightly differ from our ten TM homology model based on latest available topological information of ASBT_{YF} or ASBT_{NM} provided by *Zhou et al.*⁴¹

Trp118 is a residue that attracted our attention because of its interesting position change during ASBT's conformational change. Inspecting the outward open position, a state where the transported substrate approaches ASBT from the extracellular side, Trp118 is tilted forwards and seems to block the substrate's exit route with its **bulky, hydrophobic indole-side chain** (figure 28 B). Tryptophan is generally infrequently seen in proteins, because of its unique and energetic demanding properties. However, It is acknowledged that Trp plays a determining role in transmembrane proteins, regarding protein stability (hydrophobic mismatch) and conformational change.⁶⁴

Provided with this information it would be within the realms of possibility that this notable and widely conserved aa in ASBT and SOAT's sequences contributes to the transport cycle, either functioning as a **gate** or allowing to establish aromatic as well as pi-interactions with ligands.⁶⁵

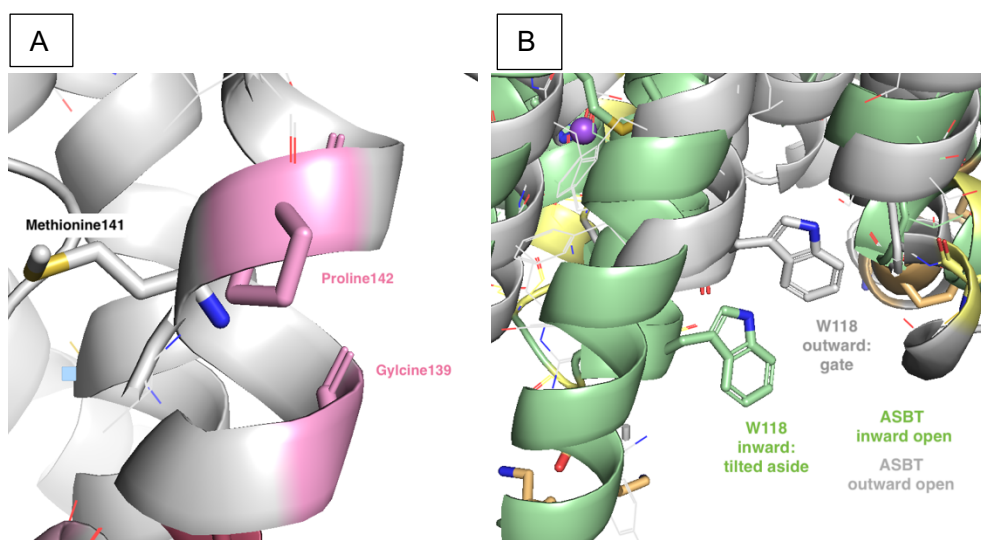


Figure 28: A) Depicting the strictly conserved “signature motif”, a possible interaction between Pro142 and Gly139 is conceivable and assumed to have structural consequences on the protein structure.

B) Showing ASBT outward (grey) and inward (green) open conformation superimposed. Trp118 (W) seems to block the exit route of ASBT's substrate and therefore maybe play a yet undiscovered role as a gate for the transport cycle.

The polar amino acids **Thr167** and **Ser171** were discovered through our IFD studies, showing preferred interactions with docked **BA's tail region**. We suspect them, because of their **non conservation**, to affect the differences of substrate specificity between hASBT, hNTCP and hSOAT (see chapter 4.3.3 “Heat map and PLIF”).

Noteworthy is also the expected variability of amino acids **Thr267** and **Ser293** which play a key role in our postulated “**Ser-Thr-Lock Theory**” for hASBT (figure 20, chapter 4.3.2.1). Here we suspect the diverging exhibited amino acids among ASBT, NTCP and SOAT to be the reason of significance. The almost complete trans-species conservation (ASBT: Ser-Thr, SOAT: Ile-Gly, NTCP: Val-Met) for each transporter itself, but not within even close related ones, seems to support our hypothesis of causing the **differences in**

substrate specificity. No functional involvement in conformational change can be expected from those residues. Furthermore, a local closeness to important residues can be seen (Thr262, Gln297) and could additionally be interpreted as a sign of relevance.

Thr110 and **Gln297** should also be mentioned, since those residues have been part of our “structural water hypothesis” (see chapter 4.2.3). Here we speculate, on the basis of mutation studies⁴¹, that the putative **structural water is necessary for binding the substrate** in the **inward open conformation**, when **interacting with Gln297** via a **hydrogen bond network**. The continuous conservation in 5 out of 7 human SLC10 members (except P5 and P3 – data not shown) is from our point of view a strong evidence of relevance and involvement in substrate binding. Thr110’s rather moderate conservation (exception: all NTCP species – Asn) could be a hint of different interaction patterns with hydroxylated-BA residue-R₁(C₃) and hence resulting differences in affinity between ASBT and NTCP. Finally, one more interesting residue among the other marked should be mentioned, a punctual mutation from **Ser267** to **Phe preserved NTCP’s estrone sulfate uptake**, but reduced TCH and Cholate’s (= BAs) uptake. Because of the immediate vicinity to mutations causing PBAM in ASBT (Leu243; Thr262) and the strict conservation of this region, *Kim et al.*⁸ conclude that this Ser is located in an area essential for interaction with BA-substrates.

Definitely further investigation, for example mutation studies, will be needed to verify important residues to be able to entirely explain ligand interactions for every transporter.

5. Conclusions and Outlook

As mentioned in the beginning, the elucidation of a protein's transport mechanism including a definition of crucial features associated with ligand-interactions, is never an easy task. It is hard to trace back the inner selectivity- or affinity-deciding determinant only based on a homology model due to the complex nature of proteins.

Owing to the fact that only a very limited amount of structural information was available for hASBT and hNTCP (no crystal structure), we had to build valid **homology models** first. After rigorous investigation and according to the latest state of knowledge we decided to base our model-calculations on two promising homologs called ASBT_{Yf} and ASBT_{NM}. When appropriate homology models for ASBT and NTCP were ensured, we started our docking studies with the aim of a basic understanding of established interactions between different bile acid species in **ASBT's inward open conformation**. Thereby we were able to state our hypothesis of a putative **structural water** involved in binding (chapter 4.2.3) and could get a hint what defines the variations in **affinity** between primary and secondary BAs (chapter 4.2.2).

When thinking about the transport process itself we adapted our approach and focused on the transporter's **outward open conformation**. It is generally assumed that the initial protein-ligand contact is made in this conformation.⁴¹ For this reason we suspect at this point a first differentiation whether a substrate is suitable to be transported or not (**specificity**).

Our induced fit docking studies were carried out for hASBT with the result of a so far unrevealed trend of binding, including a **distinct orientation for primary and secondary BAs** (chapter 4.3.2).

Further an approach was made to define hASBT's **substrate specificity**. We assume it to be a combination of transporter specific amino acid residues, essentially for substrate recognition (4.3.2.1 "Ser-Thr-Lock Theory"), and the substrate's configuration. These calculations have been supplemented by interaction patterns derived from a **heat map**.

All these considerations were made with respect to hNTCP which helped to gain useful information.

Ligand-based Pharmacophores (chapter 4.4) were built to aid the validation of our postulated hypotheses, which are unfortunately currently not capable to provide a convincing explanation.

Finally, investigating the **phylogenetic relationship** (chapter 4.5) within the SLC10 family gave a profound overview of stated patterns and aided evolutionary retracement.

In conclusion, it can be said that this diploma thesis is combining **structure based** (homology modeling, docking) as well as **ligand-based approaches** (pharmacophore modeling) to unravel the reasons for affinity, selectivity and specificity from different perspectives. This allowed to combine the knowledge gained from ligand-displayed features and the consideration of steric binding site characteristics into one comprehensive theory.

This piece of work could be used as a starting point for further research, and could pave the way for a complete characterization of hASBT's and hNTCP's transport cycle. In the future this could lead to **new drugs** specifically aiming to **inhibit ASBT**, offering people suffering e.g. from **Hypercholesteremia**, a simple but yet effective therapeutic approach. Therefore, a heat map of established interactions and pharmacophores based on already known Inhibitors could be envisioned.

Knowing the exact determinants defining substrate transport could also help to design **BA-linked drugs** with enhanced bioavailability.³

To reach this goal still some working steps lie ahead, such as verification of our predicted noteworthy residues via a combination of **mutation studies** and **transport uptake assays** to proof our concept of binding.

As mentioned above our ligand-based **pharmacophores** should be adjusted to obtain more precise screening results when searching for new compounds. Generally it can be said that the drug target ASBT has for sure a wide area of therapeutic application, which can be expanded the more information is gathered.

The process of determining hNTCP's transport cycle could be tackled in the same way as performed for ASBT in this thesis.

Currently the assumptions stated in this thesis can be seen as hypothetical, since we have only limited options to proof our rigorous elaborated, but still theoretical concept so far. Maybe in the near future more information or even the crystal structures of hASBT and hNTCP will be available, providing insight in the true mechanism of transport.

6. References

1. Dawson, P. A. Role of the Intestinal Bile Acid Transporters in Bile Acid and Drug Disposition. in *Drug Transporters* (eds. Fromm, M. F. & Kim, R. B.) vol. 201 169–203 (Springer Berlin Heidelberg, 2011).
2. Monte, M. J., Marin, J. J., Antelo, A. & Vazquez-Tato, J. Bile acids: Chemistry, physiology, and pathophysiology. *World J. Gastroenterol.* **15**, 804 (2009).
3. Balakrishnan, A. & Polli, J. E. Apical Sodium Dependent Bile Acid Transporter (ASBT, SLC10A2): A Potential Prodrug Target. *Mol. Pharm.* **3**, 223–230 (2006).
4. Claro da Silva, T., Polli, J. E. & Swaan, P. W. The solute carrier family 10 (SLC10): Beyond bile acid transport. *Mol. Aspects Med.* **34**, 252–269 (2013).
5. Hu, N.-J., Iwata, S., Cameron, A. D. & Drew, D. Crystal structure of a bacterial homologue of the bile acid sodium symporter ASBT. *Nature* **478**, 408–411 (2011).
6. Geyer, J., Wilke, T. & Petzinger, E. The solute carrier family SLC10: more than a family of bile acid transporters regarding function and phylogenetic relationships. *Naunyn. Schmiedebergs Arch. Pharmacol.* **372**, 413–431 (2006).
7. Roberts, M. S., Magnusson, B. M., Burczynski, F. J. & Weiss, M. Enterohepatic Circulation. *Clin. Pharmacokinet.* **41**, 751–790 (2002).
8. Ho, R. H., Leake, B. F., Roberts, R. L., Lee, W. & Kim, R. B. Ethnicity-dependent Polymorphism in Na⁺-taurocholate Cotransporting Polypeptide (SLC10A1) Reveals a Domain Critical for Bile Acid Substrate Recognition. *J. Biol. Chem.* **279**, 7213–7222 (2004).
9. Venclovas, Č. & Margelevičius, M. Comparative modeling in CASP6 using consensus approach to template selection, sequence-structure alignment, and structure assessment. *Proteins Struct. Funct. Bioinforma.* **61**, 99–105 (2005).
10. Colas, C., Ung, P. M.-U. & Schlessinger, A. SLC transporters: structure, function, and drug discovery. *MedChemComm* **7**, 1069–1081 (2016).
11. Muhammed, M. T. & Aki-Yalcin, E. Homology modeling in drug discovery: Overview, current applications, and future perspectives. *Chem. Biol. Drug Des.* **93**, 12–20 (2019).

12. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
13. BLAST: Basic Local Alignment Search Tool.
<https://blast.ncbi.nlm.nih.gov/Blast.cgi>.
14. HHpred | Bioinformatics Toolkit.
<https://toolkit.tuebingen.mpg.de/tools/hhpred#>.
15. Sali, A., Potterton, L., Yuan, F., van Vlijmen, H. & Karplus, M. Evaluation of comparative protein modeling by MODELLER. *Proteins* **23**, 318–326 (1995).
16. PROMALS3D Documentation.
http://prodata.swmed.edu/promals3d/info/promals3d_help.html.
17. Pei, J., Kim, B.-H. & Grishin, N. V. PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Res.* **36**, 2295–2300 (2008).
18. Webb, B. & Sali, A. Comparative Protein Structure Modeling Using MODELLER. *Curr. Protoc. Bioinforma. Ed. Board Andreas Baxevanis Al* **54**, 5.6.1-5.6.37 (2016).
19. Sali, A. & Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815 (1993).
20. Colas, C. *et al.* Ligand Discovery for the Alanine-Serine-Cysteine Transporter (ASCT2, SLC1A5) from Homology Modeling and Virtual Screening. *PLOS Comput. Biol.* **11**, e1004477 (2015).
21. Mysinger, M. M., Carchia, M., Irwin, John. J. & Shoichet, B. K. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* **55**, 6582–6594 (2012).
22. Kirchweber, B. *et al.* In Silico Workflow for the Discovery of Natural Products Activating the G Protein-Coupled Bile Acid Receptor 1. *Front. Chem.* **6**, (2018).
23. Comparative Protein Structure Modeling of Genes and Genomes | Annual Review of Biophysics. https://www-annualreviews-org.uaccess.univie.ac.at/doi/full/10.1146/annurev.biophys.29.1.291?url_ver=Z39.88-2003&rfr_id=ori%3Arid%3Acrossref.org&rfr_dat=cr_pub%3Dpubmed.
24. Meng, X.-Y., Zhang, H.-X., Mezei, M. & Cui, M. Molecular Docking: A Powerful Approach for Structure-Based Drug Discovery. *Curr. Comput. Aided-Drug Des.* **7**, 146–157 (2011).
25. Pagadala, N. S., Syed, K. & Tuszynski, J. Software for molecular docking: a

review. *Biophys. Rev.* **9**, 91–102 (2017).

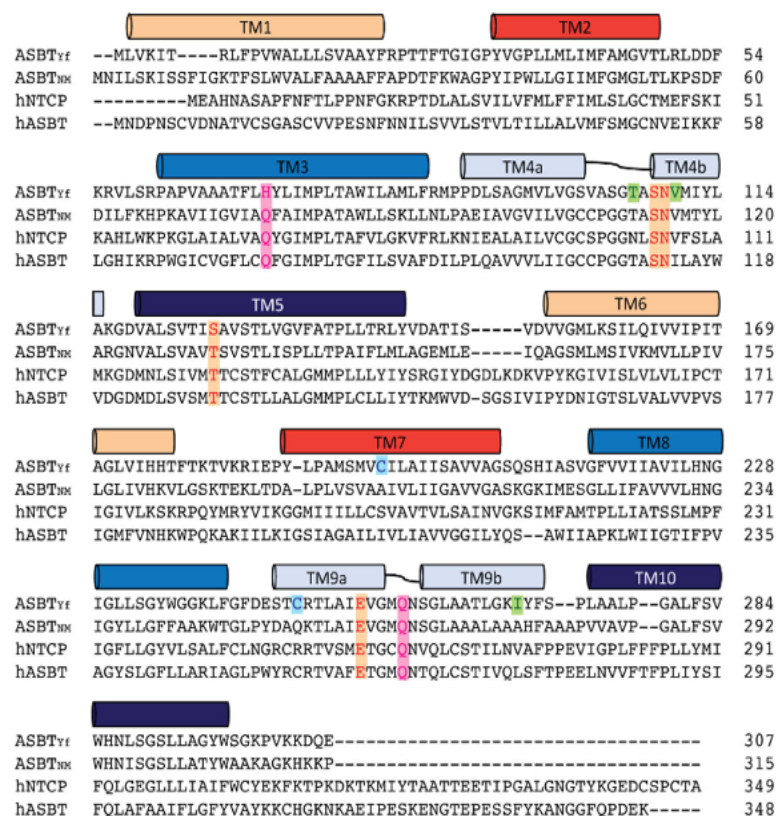
26. *Schrödinger Release 2019-1: Glide, Schrödinger, LLC, New York, NY, 2019.*
27. *Schrödinger Suite 2012 Protein Preparation Wizard; Epik version 2.3, Schrödinger, LLC, New York, NY, 2012; Impact version 5.8, Schrödinger, LLC, New York, NY, 2012; Prime version 3.1, Schrödinger, LLC, New York, NY, 2012.*
28. Multiple grid arrangement improves ligand docking with unknown binding sites: Application to the inverse docking problem | Elsevier Enhanced Reader.
<https://reader.elsevier.com/reader/sd/pii/S1476927117303821?token=A5559B4430330B0D3299F4E94A53E5C51CC517DB259940DE2583589CE59B820B880ED37A54F4319D14A392AF4C25B328> doi:10.1016/j.compbiolchem.2018.02.008.
29. Receptor Grid Generation Panel.
http://gohom.win/ManualHom/Schrodinger/Schrodinger_2015-2_docs/maestro/help_Maestro/glide/receptor_grid_generation.html.
30. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy | Journal of Medicinal Chemistry.
<https://pubs-acsc-org.uaccess.univie.ac.at/doi/full/10.1021/jm0306430>.
31. What are the advantages and disadvantages of Glide regular docking and induced fit docking? | Schrödinger. <https://www.schrodinger.com/kb/739>.
32. Koshland, D. E. The active site and enzyme action. *Adv. Enzymol. Relat. Subj. Biochem.* **22**, 45–97 (1960).
33. *Schrödinger Suite 2009 Induced Fit Docking protocol; Glide version 5.5, Schrödinger, LLC, New York, NY, 2009; Prime version 2.1, Schrödinger, LLC, New York, NY, 2009.*
34. Karthikeyan, M. & Vyas, R. *Practical Chemoinformatics*. (Springer, 2014).
35. Jain, S., Grandits, M., Richter, L. & Ecker, G. F. Structure based classification for bile salt export pump (BSEP) inhibitors using comparative structural modeling of human BSEP. *J. Comput. Aided Mol. Des.* **31**, 507–521 (2017).
36. Kutlushina, A., Khakimova, A., Madzhidov, T. & Polishchuk, P. Ligand-Based Pharmacophore Modeling Using Novel 3D Pharmacophore Signatures. *Molecules* **23**, (2018).
37. Yang, S.-Y. Pharmacophore modeling and applications in drug discovery: challenges and recent advances. *Drug Discov. Today* **15**, 444–450 (2010).
38. Phase User Manual. 246.

39. Schrödinger Release 2019-1: Phase, Schrödinger, LLC, New York, NY, 2019.
40. Dixon, S. L. *et al.* 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.* **20**, 647–671 (2006).
41. Zhou, X. *et al.* Structural basis of the alternating-access mechanism in a bile acid transporter. *Nature* **505**, 569–573 (2014).
42. Döring, B., Lütke, T., Geyer, J. & Petzinger, E. Chapter Four - The SLC10 Carrier Family: Transport Functions and Molecular Structure. in *Current Topics in Membranes* (ed. Bevensee, M. O.) vol. 70 105–168 (Academic Press, 2012).
43. Zhang, E. Y. *et al.* Topology Scanning and Putative Three-Dimensional Structure of the Extracellular Binding Domains of the Apical Sodium-Dependent Bile Acid Transporter (SLC10A2). *Biochemistry* **43**, 11380–11392 (2004).
44. Yan, C. & Luo, J. An analysis of reentrant loops. *Protein J.* **29**, 350–354 (2010).
45. Müller, S. F., König, A., Döring, B., Glebe, D. & Geyer, J. Characterisation of the hepatitis B virus cross-species transmission pattern via Na⁺/taurocholate co-transporting polypeptides from 11 New World and Old World primate species. *PLoS ONE* **13**, (2018).
46. Drew, D. & Boudker, O. Shared Molecular Mechanisms of Membrane Transporters. *Annu. Rev. Biochem.* **85**, 543–572 (2016).
47. UniProt. <https://www.uniprot.org/>.
48. PyMOL The PyMOL Molecular Graphics System, Version 2.2 Schrödinger, LLC.
49. Protein Preparation Wizard; Epik, Schrödinger, LLC, New York, NY, 2016; Impact, Schrödinger, LLC, New York, NY, 2016; Prime, Schrödinger, LLC, New York, NY, 2019.
50. Baringhaus, K.-H., Matter, H., Stengelin, S. & Kramer, W. Substrate specificity of the ileal and the hepatic Na⁺/bile acid cotransporters of the rabbit. II. A reliable 3D QSAR pharmacophore model for the ileal Na⁺/bile acid cotransporter. *J. Lipid Res.* **40**, 2158–2168 (1999).
51. Ball, P. Water as an Active Constituent in Cell Biology. *Chem. Rev.* **108**, 74–108 (2008).
52. Henchman, R. H., Tai, K., Shen, T. & McCammon, J. A. Properties of water

- molecules in the active site gorge of acetylcholinesterase from computer simulation. *Biophys. J.* **82**, 2671–2682 (2002).
53. Michel, J., Tirado-Rives, J. & Jorgensen, W. L. Prediction of the Water Content in Protein Binding Sites. *J. Phys. Chem. B* **113**, 13337–13346 (2009).
 54. Kozakov, D. *et al.* The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins. *Nat. Protoc.* **10**, 733–755 (2015).
 55. *Prime, Schrödinger, LLC, New York, NY, 2019.* (2019).
 56. *LigandScout 4.4.1.* (Inte:Ligand Software-Entwicklungs und Consulting GmbH).
 57. *SeeSAR version 9.0; BioSolveIT GmbH, Sankt Augustin, Germany, 2019, www.biosolveit.de/SeeSAR.*
 58. Mulakala, C. & Viswanadhan, V. N. Could MM-GBSA be accurate enough for calculation of absolute protein/ligand binding free energies? *J. Mol. Graph. Model.* **46**, 41–51 (2013).
 59. Can I relate MM-GBSA energies to binding affinity? | Schrödinger. <https://www.schrodinger.com/kb/1647>.
 60. Deng, Z., Chuaqui, C. & Singh, J. Structural Interaction Fingerprint (SIFt): A Novel Method for Analyzing Three-Dimensional Protein–Ligand Binding Interactions. *J. Med. Chem.* **47**, 337–344 (2004).
 61. Engelen, S., Trojan, L. A., Sacquin-Mora, S., Lavery, R. & Carbone, A. Joint Evolutionary Trees: A Large-Scale Method To Predict Protein Interfaces Based on Sequence Sampling. *PLoS Comput. Biol.* **5**, (2009).
 62. Liu, Y. & Bahar, I. Sequence Evolution Correlates with Structural Dynamics. *Mol. Biol. Evol.* **29**, 2253–2263 (2012).
 63. Hussainzada, N., Da Silva, T. C. & Swaan, P. W. Cytosolic Half of Helix III Forms Substrate Exit Route during Permeation Events of the Sodium/Bile Acid Cotransporter ASBT. *Biochemistry* **48**, 8528–8539 (2009).
 64. de Jesus, A. J. & Allen, T. W. The role of tryptophan side chains in membrane protein anchoring and hydrophobic mismatch. *Biochim. Biophys. Acta BBA - Biomembr.* **1828**, 864–876 (2013).
 65. Seidel, T., Schuetz, D. A., Garon, A. & Langer, T. The Pharmacophore Concept and Its Applications in Computer-Aided Drug Design. in *Progress in the Chemistry of Organic Natural Products 110: Cheminformatics in Natural Product Research* (eds. Kinghorn, A. D. *et al.*) 99–141 (Springer International Publishing, 2019). doi:10.1007/978-3-030-14632-0_4.

7. Appendix

7.1 Supplemental Material



Appendix 1: Sequence alignment of bacterial homologs ASBT_{Yf} and ASBT_{NM}, human ASBT and NTCP used by Zhou *et al.*³⁹ to build their model. This has been a useful template for us to prepare our homology models. Orange (Na₁) and pink (Na₂) marked residues are interacting with the sodium ions and the colored rectangles indicate with their numbering the position of transmembrane helices.

ASBT					Enrichment at least 1	
					AUC	LogAUC
Distance (Nitrogen)	asbt Model	Model number	rank	DOPE score		
2,952	dist_asbt.B99990022.txt	22	#66	-0,47392	41,107	8,867
2,945	dist_asbt.B99990051.txt	51	#74	-0,45965	32,129	7,329
2,936	dist_asbt.B99990044.txt	44	#16	-0,55287	41,432	12,72
2,927	dist_asbt.B99990028.txt	28	#47	-0,50764	15,176	3,233
2,921	dist_asbt.B99990098.txt	98	#14	-0,55976	40,355	8,424
2,898	dist_asbt.B99990072.txt	72	#22	-0,53756	39,529	8,495
2,893	dist_asbt.B99990060.txt	60	#3	-0,60248	25,122	5,337
2,879	dist_asbt.B99990041.txt	41	#75	-0,45952	54,269	11,831
2,828	dist_asbt.B99990064.txt	64	#90	-0,40733	53,853	13,078
2,775	dist_asbt.B99990077.txt	77	#6	-0,58015	41,27	10,578
2,735	dist_asbt.B99990085.txt	85	#17	-0,55261	50,802	14,113
Distance (Oxygen)	asbt Model	Model number	rank	DOPE score		
2,847	dist_asbt.B99990019.txt	19	#7	-0,5772	60,816	21,748
2,816	dist_asbt.B99990084.txt	84	#18	-0,55073	61,421	16,159
2,755	dist_asbt.B99990015.txt	15	#82	-0,44511	60,175	15,897
2,626	dist_asbt.B99990023.txt	23	#71	-0,46242		
Scores of 3ZUY mdl_cut from build.log						
normal. DOPE Score	Model number	rank				
asbt.B99990090.pdb	-0,58596	90	#5		37,011	9,77
asbt.B99990095.pdb	-0,58629	95	#4		50,914	13,389
asbt.B99990060.pdb	-0,60248	60	#3		25,122	5,337
asbt.B99990012.pdb	-0,61315	12	#2		44,546	13,81
asbt.B99990057.pdb	-0,65323	57	#1 best	mdl57RotaGln	61,607	23,857
Scores of NTCP (water)						
normal. DOPE Score	Model number	rank				
slc10a1.B99990093.pdb	-0,5592	93	#5			
slc10a1.B99990025.pdb	-0,56131	25	#4			
slc10a1.B99990057.pdb	-0,56265	57	#3			
slc10a1.B99990049.pdb	-0,57303	49	#2			
slc10a1.B99990096.pdb	-0,59384	96	#1 best	not possible	h bonds with Asn as Acceptor = TCH R1 as Donor	
Distance Nitrogen Asn NTCP	slc10a1 Model	Model number	rank			
3,246	dist_slc10a1.B99990016.txt	16	#91			
2,971	dist_slc10a1.B99990069.txt	69	#13			
2,95	dist_slc10a1.B99990095.txt	95	#94			
2,877	dist_slc10a1.B99990080.txt	80	#28			
2,86	dist_slc10a1.B99990021.txt	21	#06			
2,854	dist_slc10a1.B99990001.txt	1	#62			
2,786	dist_slc10a1.B99990053.txt	53	#64			
2,763	dist_slc10a1.B99990031.txt	31	#72			
2,758	dist_slc10a1.B99990081.txt	81	#98			
2,757	dist_slc10a1.B99990061.txt	61	#78			

Appendix 2: Detailed information of all measured distances (oxygen, nitrogen) of our models (inward open) with associated DOPE score and their ranking. This was done to reduce of big number of homology models for the enrichment process. The AUC of each chosen model is noted and as visible model 57 has the best AUC value and is also ranked first according to its DOPE score.

NB: enrichment could not be executed for hNTCP, since we had not enough structural information for the binding site preparation.

ePharmacophore of groups	R ₁	R ₂	R ₃	tail
group 1: (Cholate, GC) TC	x	x	x	negative; Carboxy: Acc
group2: CDC, GCDC, TCDC	Acc	Don	x	x
group 3: DC, GDC, TDC	Don	x	Don	x
group 2B: UDC, GUDC, TUDC	Don	x	x	x

Appendix 3: Summary of picked pharmacophore features by the automatized selection process of PHASE creating a so called “ePharmacophore”. The selection of only few and unspecific features lead to an imprecise definition of necessary steric features and resulted in poor screening results. Therefore these pharmacophores got excluded.

7.2 Abstract

Two prominent members of the SLC10 family, **ASBT** (SLC10A2) and **NTCP** (SLC10A1), play a key role in the **Enterohepatic Circulation** as sodium-dependent co-transporters of bile acid (BA). ASBT's major task covers the initial uptake of BA from the ileum and its transport to the portal blood vein, where it is delivered to the liver via NTCP, located in the hepatocyte membrane. NTCP and ASBT are the leading and rate-limiting mediators of BA uptake and homeostasis in the liver and intestine.⁴ This circumstance enables various possible applications (e.g. Hypercholesterolemia treatment) of drugs acting either as a substrate or inhibitor of ASBT or NTCP.

The aim of this diploma thesis is to unravel the factors determining substrate specificity and gaining insight in the transport mechanism of ASBT and NTCP.

Due to a shortage of structural information initially **homology models** of both human transporters had to be built in two conformations to trace the substrate translocation process. **Docking studies** were conducted with the goal of understanding the basics of substrate interaction established in ASBT's inward open conformation. **Induced fit calculations** for the outward open state enabled us to hypothesize about a "Locking mechanism" of ASBT causing substrate specificity, and to observe diverging binding modes for primary and secondary BAs possibly involved in affinity differences. In addition, a heat map of binding-involved residues was created as visual aids. Moreover, certain amino acid (aa) residues have been pointed out to be strongly involved in binding substrates or causing conformational changes.

To cover the ligand's contribution ligand-based **Pharmacophore models** have been built, which put the focus on important features for binding and enable screening for new drugs in the future.

Last but not least we concentrated on the **phylogenetic relationship** of the SLC10 family, looking at the conservation of particular aa residues within different species. This was done in order to get a profound understanding of established patterns needed for function and facilitate evolutionary retractability.

7.3 Zusammenfassung

Die beiden Natrium-abhängigen Gallensalztransporter **ASBT** (SLC10a2) und **NTCP** (SLC10a1) tragen einen beträchtlichen Anteil zur Regulierung des Enterohepatischen Kreislaufes bei. Erst genannter Transporter ist für die primäre **Gallensalzaufnahme** (GS) in den Ileozyten (Darm) zuständig, Zweiter für die Aufnahme und Rückführung der Gallensalze aus dem Pfortaderblut in die Leber. Anhand dieser Schlüsselposition bei der Aufrechterhaltung der GS-Homöostase können **pathologische Defekte** der beiden Transporter einen gravierenden Einfluss auf physiologische Prozesse haben (z.B.: primäre Gallensäure-Malabsorption PBAM). Nun kann durch eine **gezielte Hemmung des ASBT-Transporters** mit Arzneistoffen eine erhöhte Ausscheidung von Gallensäuren erzielt und somit entgleiste Cholesterinspiegel reguliert werden. Dies wäre ein denkbarer Ansatz zur **Therapie von Hypercholesterinämie**.

Die **Aufklärung** der die **Spezifität bestimmenden Faktoren** und die Erläuterung der **zugrundeliegenden Bindungsmechanismen** des Transportzyklus´ von hASBT und hNTCP waren **Ziel dieser Diplomarbeit**.

Aufgrund mangelnder struktureller Information der humanen Proteine mussten zunächst **Homologie Modelle** beider Transporter in der jeweils „einwärts-geöffneten“ und „auswärts-geöffneten“ Konformation modelliert werden. Eine **Dockingstudie** mit der Innenseite-zugewandten Position des Proteins wurde durchgeführt, um grundlegende Kenntnis über vorhandene Interaktionen zwischen Substraten und dem Transporter zu erlangen. Mittels „**Induced Fit**“-Berechnungen (induzierte Passform) konnte eine Hypothese über eine **vermutete Konformationsänderung** aufgestellt werden, die im Zusammenhang mit der **Substratspezifität** steht. Ebenso konnten beim Versuch, die Affinität nachzuvollziehen, unterschiedliche Bindungsposen für primäre und sekundäre GS entdeckt werden.

Darüber hinaus wurde eine „**Heat map**“ für die bessere Erkennbarkeit von bindungsbeteiligten Aminosäure(AS)-Resten erstellt. Auch konnten an der **Bindung oder Konformationsänderung beteiligte** wichtige **AS** ausfindig gemacht werden. Ergänzend wurden **Pharmakophor-Modelle** erstellt, um die Beteiligung der Liganden an der Interaktion nicht außer Acht zu lassen. Dies ermöglicht das Herausarbeiten der benötigten Moleküleigenschaften und das zukünftige Filtern von Arzneistoffdatenbanken. Abschließend wurde auf die **phylogenetische Verwandtschaft** von verschiedenen Spezies und deren zugehörigen Sequenzen eingegangen, mit dem Ziel, wichtige etablierte strukturelle Muster nachzuvollziehen.

7.4 List of Abbreviations

AA	Amino Acid
AUC	Area under the curve
ASBT	Apical Sodium-dependent Bile acid Transporter
ASBT_{Yf}	Bacterial Homolog of Y. frederiksenii
ASBT_{NM}	Bacterial Homolog of N. meningitidis
BA(s)	Bile Acid(s)
BLAST	Basic Local Alignment Search Tool
DUD E	Database of Useful Decoys: Enhanced
EHC	Enterohepatic Circulation
Glide	Grid-based Ligand Docking with Energetics
hASBT	human ASBT
HPred	Homology detection and structure prediction by HMM-HMM comparison
hNTCP	human NTCP
IDF	Induced Fit Docking
LDL	Low density lipoprotein
MM GBSA	Molecular mechanics with generalised Born and surface area solvation
MOE	Molecular Operating Environment
NTCP	Sodium/Taurocholate Cotransporting Polypeptide
PBAM	Primary Bile Acid Malabsorption
SMARTS	Simplified Molecular Input Line Entry System
SNP	Single Nucleotide Polymorphism
TCH	Taurocholate (a bile acid)
TMD	Transmembrane Domains

All used abbreviations for bile acids can be seen in table 3