Characterization of the Function and Expression of Variants at Potential Rheostat, Toggle, and Neutral Positions in the Na^+/Taurocholate Cotransporting Polypeptide (NTCP)

By

Melissa Jean Ruggiero
B.Sc., Northwest Missouri State University, 2015

Submitted to the graduate degree program in Toxicology and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

_____
Committee Chair: Bruno Hagenbuch, Ph.D.

_____
Tiangang Li, Ph.D.

_____
Kenneth McCarson, Ph.D.

_____
Michele Pritchard, Ph.D.

_____
Liskin Swint-Kruse, Ph.D.

Date Defended: 22 April 2021

The dissertation committee for Melissa Ruggiero certifies that this
is the approved version of the following dissertation:

Characterization of the Function and Expression of Variants at Potential Rheostat, Toggle, and Neutral Positions in the Na^+/Taurocholate Cotransporting Polypeptide (NTCP)

Chair and Graduate Director: Bruno Hagenbuch, Ph.D.

Date Approved: 24 April 2021

Abstract

Mutations and single-nucleotide polymorphisms (SNPs) (mutations that occur in more than one percent of a population) can occur throughout the human genome. The study of polymorphisms, single-nucleotide mutations, and predictive pharmacogenomics has become a focus for both researchers and physicians in recent years. However, knowing about the existence of such variants is only part of the challenge. Mutations in DNA can impact the function of the gene, protein, enzyme, etc. that is encoded by that region of DNA to varying degrees. For decades, computer algorithms have attempted to predict outcomes of amino acid substitutions based on assumptions about important amino acids that were derived from decades of experimentation: 1) most amino acid substitutions result in damaged protein function or structure 2) few substitutions allowed for normal protein function and 3) the same amino acid substitutions will behave similarly in homologous proteins. Thus, there were three types of functional outcomes that positions were generally expected: neutral, deleterious, or catastrophic. A mutation that causes little to no effect is considered a neutral substitution. In contrast, if a mutation causes decreased protein function that results in a clinical phenotype it is considered to be a deleterious substitution. Further, if a mutation results in a complete lack of function and even degradation of the protein, it is classified as a catastrophic substitution.

However, one major flaw of the current algorithms is that they are largely based on data and rules obtained from mutating conserved amino acid positions. Very few studies systematically studied whether these rules apply to nonconserved positions. This leaves mutations occurring in nonconserved locations enigmatic. Recent studies in the LacI/GalR transcriptional regulator family as well as in globular proteins have shown that multiple substitutions at some nonconserved positions result in intermediate outcomes in addition to neutral and catastrophic.

Further, when taken together, multiple substitutions at these amino acid positions result in a continuum of outcomes akin to a dimmer switch, thus they are referred to as rheostat positions. In addition, many of these intermediate substitutions do not always damage the protein, nor do the same amino acid substitution in different homologs always produce the same functional outcomes. Because these substitutions do not follow these canonical assumptions, they are currently unpredictable. Furthermore, the existence and behavior of rheostat positions has yet to be elucidated in other types of proteins, including integral transmembrane proteins. These proteins are particularly important for the disposition and metabolism of therapeutic drugs. Thus, alterations in the function of these proteins are particularly important for the prediction of drug response. Therefore, our aim is to predict and characterize rheostat positions in an integral transmembrane protein to improve computational predictions of rheostat-like functional outcomes. Our original hypothesis was that these positions would be more likely to occur at nonconserved amino acid locations.

We decided to use the Na^+/taurocholate cotransporting polypeptide (NTCP) as a model integral transmembrane protein. Human NTCP is a transporter highly expressed at the basolateral membrane of hepatocytes and is involved in the enterohepatic circulation of bile acids. In addition, NTCP can transport sulfated hormones and certain statins. Given the diversity of these substrates, we decided to utilize three model substrates to study the function of NTCP: taurocholate (a bile acid), estrone-3-sulfate (a sulfated hormone), and rosuvastatin (a statin). We then selected potential rheostat positions within human NTCP using three methods. The first was to select a known polymorphic location shown to have varying levels of functional impact depending on the substrate, position 267. Next, we selected a position, 271, that was hypothesized to be a rheostat based on protein structure modeling and energy calculations.

Finally, we selected two additional positions based on evolutionary differences calculated using a multiple sequence alignment consisting of over 1500 sequences from both pro- and eukaryotes. Using this final method along with other criteria, we selected a highly conserved and a highly nonconserved amino acid positions, 102 and 146, respectively.

Our studies concluded that position 102 was the strongest rheostat for both expression and function, while positions 267 and 271 were neutral for expression but rheostatic for all substrates, except for estrone-3-sulfate transport by N271 variants. Further, position 146 was neutral for most variants for both expression and function. In addition, characterization of select S267 and N271 variants revealed impacts on Michaelis-Menten kinetics to varying degrees. For example, S267 variants demonstrated more alterations in NTCP's capacity while at position 271 most alterations affected the transporter's velocity or turnover. We next related the function and expression of these S267 and N271 variants to simulated structure model energy predictions to assess if they could aid in the prediction of the observed functional outcomes. While there were some correlations between the surface expression and calculated energy scores, more studies are needed to determine if improved computational models could be used to predict complex rheostatic outcomes.

Acknowledgements

I would first like to acknowledge the funding sources that made the work in this dissertation possible: the W.M. Keck Foundation and the National Institutes of Health grants, R01 GM077336, P20 GM103549, and P30 GM118247.

I would next like to thank my mentor, Dr. Bruno Hagenbuch, for all his time, energy and commitment, to not only me but to all the students that he interacts with. Bruno, you constantly go above and beyond for your students, a trait that I have always appreciated and admired. It was an absolute pleasure working with you these last six years as a graduate student, as well as the Summers before that. I will be forever grateful that you took a chance on me as an undergraduate and if it weren't for you, I would have never pursued graduate school. I look up to you in every way and hope to someday be at least half the mentor and scientist you are. Through your guidance, I have accomplished more than I thought possible and I hope to always make you proud with the scientist I become. Thank you for not only being the best mentor but also a good friend.

Next, I would like to thank my committee members, both past and present: Dr. Michele Pritchard, Dr. Kenneth McCarson, Dr. Tiangang Li, and Dr. Clifford Mason, but most of all Dr. Liskin Swint-Kruse. Your guidance, support, and mentorship were pertinent in bringing this project to fruition.

In addition, I would like to show my appreciation for my previous and current lab mates: Yuchen, Wen, Kelli, Miki, Jess, Regina, Bri, Haley, Jon, Matt, Jessica, and Steffie, as well as all the undergraduate and high school students I had the pleasure of mentoring. I am so grateful to

have worked and built friendships with so many wonderful scientists. Your willingness to listen and to help me troubleshoot was imperative to the success of this dissertation.

I would also like to thank the other students, faculty, and staff in our department without whom the Pharm Tox department wouldn't be the same. There are others across the KUMC campus, too many to name, that I had the opportunity to work with on projects and in organizations. To those of you to whom I'm referring, please know you made my experience at KUMC even more memorable and I will never forget you.

Finally, I would like to thank all my family and friends for their support and understanding, especially when I cancelled plans to focused on my studies. You all kept me sane and helped me celebrate every milestone and accomplishment.

Most importantly, I would like to thank my amazing parents and my wonderful husband, Alex. You have always encouraged me to pursue my dreams, reach for higher goals, and continue to do my best even when I doubted myself. Without all your support both emotionally (and financially), I never would have gotten this far. I will be forever indebted to you. You all can now finally stop asking me when I am going to graduate!

Table of Contents

Chapter 1

Introduction

Chapter 2

Materials and Methods

Chapter 3

A clinically-relevant polymorphism in the Na^+/taurocholate cotransporting polypeptide (NTCP) occurs at a rheostat position

Chapter 4

Examination and characterization of predicted rheostat, toggle and neutral positions within the Na^+/taurocholate cotransporting polypeptide (NTCP)

Chapter 5

Chapter 6

List of Figures

List of Tables

List of Abbreviations

ADME	=	Absorption, distribution, metabolism and excretion
ANOVA	=	Analysis of variance
ASBT	=	Apical Sodium-Dependent Bile Acid Transporter
BCA	=	Bicinchoninic acid
BCRA 1/2	=	Breast cancer gene 1/2
BCRP	=	Breast cancer resistance protein
BLAST	=	Basic local alignment search tool
BSA	=	Bovine serum albumin
BSEP	=	Bile Salt Export Pump
CCK	=	Cholecystokinin
cDNA	=	complementary DNA
CYP	=	Cytochrome P450
DMEM	=	Dulbecco's modified eagle medium
DNA	=	Deoxyribose nucleic acid
dNTP	=	Deoxynucleoside triphosphate
E3S	=	Estrone-3-sulfate
E3S	=	Estrone-3-sulfate
EDTA	=	Ethylenediaminetetraacetic acid
EV	=	Empty vector
FBS	=	Fetal bovine serum

FXR	=	Farnesoid X-activated receptor
HBV	=	Hepatitis B virus
HDV	=	Hepatitis D virus
HEK	=	Human embryonic kidney
HEPES	=	25 mM 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
HRP	=	Horseradish peroxidase
IBAT	=	Ileal bile acid transporter
Km	=	Michaelis constant
LacI/GalR	=	Lactose repressor protein/Galactose repressor protein
LB	=	Luria-Bertani
MARS	=	Multiple circular sequence Alignment using Refined Sequences
mRNA	=	messenger ribonucleic acid
MRP	=	Multidrug resistance protein
MSA	=	Multiple sequence alignment
nsSNP	=	Non-synonymous single nucleotide polymorphism
NTCP	=	Na^+/Taurocholate Cotransporting Polypeptide
OATP	=	Organic Anion Transport Polypeptide
OCT	=	Organic Cation Transporter
Opti-MEM	=	Opti-minimal essential medium
PBS	=	Phosphate-buffer saline
PCR	=	Polymerase chain reaction
PMI	=	Precision Medicine Initiative

PTM	=	Post translational modification
PVDF	=	Polyvinylidene difluoride
REU	=	Rosetta energy units
SD	=	Standard deviation
SDS	=	Sodium dodecyl sulfate
SHP	=	Short heterodimer partner
SLC	=	Solute carrier
SNP	=	single-nucleotide polymorphism
SOAT	=	Sodium-dependent organic anion transporter
SOC	=	Super optimal broth with catabolite repression
TBS	=	Tris buffered saline
TBS-T	=	Tris buffered saline - tween 20
TCA	=	Taurocholic acid or taurocholate
TCH	=	Taurocholate
TM	=	Transmembrane domain
tRNA	=	Transfer ribonucleic acid
UGT	=	Uridine diphosphate glucuronosyltransferases
Vmax	=	Maximum transport rate
WT	=	Wildtype

Chapter 1

Introduction

I. Personalized or Precision Medicine

Overview

Since the birth of modern medicine after the industrial revolution in the 18th century, a physician's approach to a patient's treatment has been starting with the standard of care, treating every patient like the "typical patient", followed by trial and error until the desired outcome is achieved (Sankar and Parker, 2017). This guessing game of doses and medications can lead to severe side effects, high medical costs, and even delayed treatment and disease progression (Vogenberg et al., 2010). As medications and treatments became more widely used and accepted, variability in both patient demographics and drug responses increased as well. Some patients may respond to a treatment exactly as expected with minimal side effects, some may not respond to a treatment at all, and others may have severe side effects that cause the discontinuation of treatment regardless if the treatment worked or not (Roden et al., 2019). Similar demographics such as location, age, diet, preexisting conditions, socioeconomic status, and more may lead to similar responses or predisposition to disease (Vogenberg et al., 2010). Whether a patient is predisposed to a condition due to familial inheritance or exposed to environmental factors throughout their life that led to disease, such as alcohol leading to and cirrhosis of the liver, there is one common theme, DNA and genetic variability.

Improvement and Initiatives Impacting Personalized Medicine

The Human Genome Project took 13 years to complete and with it came a wealth of knowledge. There are over three billion nucleotides in the human genome that initially were predicted to encode approximately 100,000 genes (Collins and Fink, 1995). However, additional reports have determined that the actual number of protein-coding genes is closer to 20,000 rather than the initially thought 100,000 (Clamp et al., 2007). In addition, compared to the reference sequences, there are three to five million genetic variants in every human genome (Brittain et al., 2017). Many human disease conditions and disease susceptibilities are related back to genetic variations such as single nucleotide mutations that were either inherited or gained throughout the patient's lifetime (Collins and Fink, 1995). The completion of the Human Genome Project allowed for gene mapping of diseases and therefore the emergence of targeted therapies.

As a result, there has been tremendous progress in the fields of pharmacogenetics and pharmacogenomics. Pharmacogenetics is the study of how genetic variations affect drug response by looking at individual genes, while pharmacogenomics takes into account the entire genome and is the broader study of how genetic variations affect drug disposition (Vogenberg et al., 2010). These two concepts are the basis for an emerging approach to medical treatments, personalized or precision medicine. This approach to medicine is where a physician can examine and take into account a patient's lifestyle, environment, and genetic variability when deciding the best treatment options. This will lead to accelerated diagnoses, optimal treatments and decreased risk of adverse events (Sankar and Parker, 2017; Phillips et al., 2018).

In 2015, President Obama announced the Precision Medicine Initiative (PMI) in his State of the Union address. This initiative was developed to create an infrastructure to introduce and

implement precision medicine into the masses. In addition, PMI established a cohort study of a million volunteers from all walks of life to gather more information about different patient demographics, lifestyle, environment, and genetic variability (Sankar and Parker, 2017). However, there were limitations to this initiative. First, for this gathered information to be useful, numerous legal, ethical, and social issues needed to be addressed. For example, once the genetic testing of a patient was completed, how and what personal and medical information could be shared without violating the patient's privacy. Next, oversampling and targeting of certain populations needed to be avoided. In addition, the information gathered needs to be beneficial for all populations, not just a few. Regardless of these limitations, there has still been a tremendous amount of progress made in genetic testing and precision medicine.

Current Genetic Testing Landscape

Three types of genetic testing are: targeted or static gene panel sequencing, whole exome sequencing, and whole genome sequencing. The most focused type of genetic sequencing is targeted or gene panel sequencing which examines a specific gene or a set of genes for mutations. This is currently the most common use of genetic testing in medicine and typically looks for mutations that cause well known genetic diseases, such as certain cancers. Next, instead of focusing on single target genes, whole exome sequencing focuses on the DNA coding regions, the exons. This includes sequencing of more genes to give a more complete picture, however, it does not take into account the whole genetic information that may impact pharmacological responses, i.e., the role of introns. Finally, whole genome sequencing, as the name indicates, is sequencing of the entire genome giving a complete comprehensive picture of the genetic make-up of the patient. However, there needs to be a system in place to interpret the

information and therefore predict the impact of any genetic variations discovered (Brittain et al., 2017).

As of 2018, there were more than 75,000 genetic tests on the market. The majority of these tests are geared towards prenatal genetic testing and predisposition to hereditary cancers and less on pharmacogenetic tests. Over the course of two years, a cohort of six clinical domains reported that less than five percent of spending for genetic testing was used for pharmacogenetics. In contrast, hereditary cancers and prenatal genetic testing each accounted for over 30 percent of the total genetic testing costs (Phillips et al., 2018). One of the most common examples is genetic testing for mutations in the BRCA 1/2 genes to determine predisposition to breast cancer. Following genetic testing patients that have mutations in the BRCA genes may decide to undergo prophylactic mastectomies to decrease their risk of breast cancer. While these genetic tests are a step in the right direction, there is considerable room for improvement. Some of the areas that need improving include: data interpretation, impact on the patient as a whole, and the correlation between mutation and drug response.

Limitations of Genetic Testing and Personalized Medicine

Personalized or precision medicine has the power to bridge the gap between molecular information and clinical manifestations. Instead of treating symptoms, physicians could address the underlying genetic and molecular causes of the disease (Brittain et al., 2017). Furthermore, this approach has the potential to shift medical treatment from treating every patient the same to individualizing a patient's treatment regimen. As mentioned above, the current genetic testing landscape is focused on targeted genes and even whole exome sequencing but not for a lack of

interest in whole genome sequencing. A lot of information can be gained by whole genome and exome sequencing but only if researchers and physicians have the capacity and technology to interpret the data. More genetic information is being gathered every day but, with the inability to interpret it, the data are meaningless.

However, to predict a patient's pharmacological response and therefore tailor their treatment to meet their specific medical needs, a physician or researcher would, at the very least, need to know three important things. First, it is crucial to know the patient's physical fitness and demographics such as their age, diet, medical history, and even environmental factors that may impact their health. Next, as a result of whole genome sequencing, the physician would know every genetic variation in the patient's genome. Finally, the physician would need to be able to interpret the genetic data to know how those genetic differences impact the patient's health and response to treatment. Unfortunately, the technology for the final data interpretation is not yet available. Thus, the protein function prediction technology needs to be improved in order for the expense and use of whole genome genetic tests to be beneficial (Brittain et al., 2017).

II. **Protein Mutations and Functional Outcome Prediction**

Protein Overview

Proteins are complex and dynamic macromolecules that are a crucial element of all biological processes (Alberts et al., 2002). They are extremely versatile macromolecules responsible for everything from hemoglobin carrying oxygen to actin within muscles responsible for movement and everything in between (Berg et al., 2002). What begins as a string of DNA in a single cell eventually forms a fully functional protein that is essential for carrying out a biological process.

DNA Mutations

DNA mutation refers to any alteration in the DNA sequence that can greatly impact a protein's fate and function. These mutations can occur anywhere within the DNA and for a variety of reasons, including environmental exposures, chemical causes, errors during DNA replication and even random, spontaneous mutations that can become inherited genetic variations (Clancy, 2008). These factors can result in the following types of point mutations at the DNA level: substitutions, insertions, and deletions of nucleotides.

For example, exposure to ultraviolet radiation causes the formation of a bond between two thymine (T) nucleotides that are next to each other in a DNA sequence. This bond, a thymine dimer, results in two T's being read as a single nucleotide causing the second thymine to be skipped during DNA replication or transcription. This causes one fewer nucleotide to be available for codon translation and therefore the next nucleotide is read instead, ultimately shifting each subsequent codon by one nucleotide resulting in different amino acids being added to the protein. This is referred to as a frame shift mutation (Clancy, 2008). Most of the time this mutation is caught during DNA replication, but if it is not noticed by the replication machinery then it is thought to be one of the main causes of skin cancer (Pinak, 2006).

Single-Nucleotide Polymorphisms

On the other hand, if a mutation that alters a single base pair is present in more than one percent of a population's genome then it is referred to as a single-nucleotide polymorphism (SNP). In addition, if a SNP leads to a change in the amino acid sequence it is referred to as a non-synonymous single-nucleotide polymorphism (nsSNP)(Yates and Sternberg, 2013). For instance,

if a single adenine (A) in the codon GAG (guanine, adenine, guanine) is replaced with a thymine (T), the codon now reads GTG, and a valine is inserted into the protein instead of a glutamic acid. This is a simple substitution mutation in hemoglobin that can result in rod-shaped blood cells that are ineffective in transporting oxygen, which is characteristic of sickle-cell anemia (Clancy, 2008).

Amino Acid Biochemistry

Without knowing that the single nucleotide substitution in the example above manifests into sickle-cell anemia, it may not seem like a big deal. However, basic amino acid biochemistry knowledge is needed to fully understand why this single change made such a large impact in a patient's health. There are four main categories amino acids fall into based on their side chain or R-group: negatively charged, positively charged, polar, and non-polar. Table 1-1 below further categorizes the amino acids.

Table 1-1. Categorization of Amino Acids

Category	**Subcategory**	**Amino Acids**
Polar Charged	Negative	Aspartate (Asp, D) Glutamate (Glu, E)
	Positive	Lysine (Lys, K) Arginine (Arg, R) Histidine (His, H)

Polar Uncharged		Serine (Ser, S) Threonine (Thr, T) Asparagine (Asn, N) Glutamine (Gln, Q)
Hydrophobic	Aromatic	Phenylalanine (Phe, F) Tryptophan (Trp,W) Tyrosine (Tyr, Y)
	Non-Aromatic	Alanine (Ala, A) Isoleucine (Ile, I) Leucine (Leu, L) Methionine (Met, M) Valine (Val, V)
Special Cases		Cysteine (Cys, C) Glycine (Gly, G) Proline (Pro, P)

The above classification and additional factors, such as the size of the R-group or substituent, impact the proteins folding and post-translational modifications. As a result, the frequency for where each amino acid is likely to fall within the folded protein can be calculated. For example, methionine and glutamic acid are most likely to be found in an alpha helix, whereas valine and

isoleucine will likely be in beta sheets, and finally proline and glycine are most likely to be found in loops or turns (Berg et al., 2002).

Amino Acids and Protein Formation

The first step in the process of making a protein is the transcription of DNA into messenger RNA (mRNA). The mRNA may then undergo splicing, or the removal of extra information (introns), before it gets transported out of the nucleus. With the help of ribosomes and transfer RNA (tRNA), each set of three nucleotides in the mRNA is read as a codon which corresponds to one of the twenty amino acids. This process continues until each codon is read and the mRNA is fully translated into a string of amino acids to make up a single protein. But the process is not complete. As translation progresses, the protein gets folded into secondary (α helix, β sheets) and tertiary (combination of secondary structures connected by loops and turns) structures due to hydrogen bonding and interactions between the amino acid side chains, followed by quaternary structures if the protein is made up of more than one amino acid chain. In addition to protein folding, post-translational modifications (PTM) can occur simultaneously after translation.

Post-translational modifications chemically modify various amino acid positions on the protein to further diversify the protein by extending the number of the 20 natural amino acids to over 140 possible residues (Uversky, 2013). Chemical modifications of select amino acids in a protein help determine the location, activity, additional protein interactions and even the protein's fate by covalently adding sugars, peptides, lipids and more or cleaving peptide bonds (Mann and Jensen, 2003; Uversky, 2013). There are numerous types of PTMs from ubiquitination, phosphorylation and glycosylation to proteolysis, all of which help to differentiate the fate and

function of a protein. Furthermore, these modifications can impact the protein's folding. Post-translational modifications can occur at any point following translation, meaning that they can impact protein folding, the overall structure of the protein, and therefore the protein's function.

Implications of Amino Acid Substitutions

Consider the implications of a single mutation on the process of forming a complex, functional protein. If a single nucleotide is altered, the protein may be changed by one amino acid. What are the potential consequences? At best, the single amino acid substitution does not impact the protein's structure or function. Next, the mutation may not alter protein stability but may still affect the function. This outcome could be due to a protein's ability to reorganize or repack itself to accommodate the change in the amino acid. While this may not impact downstream processes, such as post-translation modifications, the protein's function may be altered. However, the mutation may not cause an effect that reaches biological relevance. Thus, the mistake goes unnoticed, and the patient's health is not negatively impacted.

On the other hand, an amino acid substitution may alter their response to drugs or cause long-lasting detrimental effects leading to diseases like cancer. This may be due to altered protein structure causing protein instability leading to 1) degradation and amino acid recycling or 2) protein dysfunction and improper regulation. One example is if a glutamic acid with a negatively charged side chain is replaced with an uncharged, non-polar, hydrophobic valine. This single substitution disrupts the side chain interactions of the replacement amino acid with its neighboring amino acids. This may disrupt hydrogen bonding thus impacting the protein's folding. As a result, the protein may not be able to interact with additional amino acid chains

thus the protein folding may be disrupted. Further, this mutation may occur at a site of a crucial post-translational modification. All these things can add up to protein instability, improper trafficking to its correct location, and ultimately dysregulated function.

A third option is that mutations can impact function but not structure. This is discussed more extensively in the next section.

Protein Function as a Result of Amino Acid Substitutions

The functional outcomes of amino acid substitutions can be divided into three classes (Meinhardt et al., 2013). The first is considered a "neutral" mutation (Figure 1-1A). This is when a single amino acid substitution results in statistically identical function to the non-mutated, or wildtype, protein. If numerous amino acids at the same location result in similar outcomes to wildtype then the amino acid position is also considered neutral (Figure 1-2A) (Fenton et al., 2020).

Next are detrimental mutations. These mutations result in a dramatic decrease or total loss of protein function and can be divided into two different classes: deleterious and catastrophic. The use of these two terms depend on the field of research. Deleterious mutations occur when the single substitution results in decreased or diminished protein function that reaches a biologically and/or environmentally determined threshold (Figure 1-1B). This deleterious terminology is generally used by Biologist or scientists in similar fields that look specifically for this threshold to be reached. On the other hand, the term catastrophic mutations is mainly used in the field of Biochemistry to describe a loss in protein function that is accompanied by protein misfolding or instability leading to protein degradation (Figure 1-1C). The substitution can also have a catastrophic effect on the function, such as when an enzyme active site is altered. If multiple

amino acid substitutions at a single position all result in either deleterious or catastrophic outcomes, then the amino acid position is considered a toggle position (Figure 1-2B). These positions would be similar to a light switch, where the fully functional wildtype protein is the "on" position and the "off" is the mutated, non-functional or degraded protein. This is the typical behavior that researchers would expect to observe at positions where changes in the protein do not often occur, namely conserved positions (Fenton et al., 2020).

Finally, the third class of amino acid substitutions results in intermediate function (Figure 1-1D). This is described as a mutation that results in a protein whose function is statistically different from the wildtype protein but is still functional, meaning there is an intermediate alteration in function. Further, if the location that results in this intermediate function is mutated to all other amino acids, the resulting substitutions' function can be arranged into a continuum of functional outcomes from detrimental to wildtype and even greater than wildtype (Figure 1-2C) (Meinhardt et al., 2013; Fenton et al., 2020). These "rheostat" positions are analogous to a dimmer switch. One example of a study that showed this rheostat-like alteration in function was demonstrated by the random mutagenesis of leucine 545 in organic anion transporting polypeptide (OATP) 1B1. This randomized mutagenesis resulted in 6 OATP1B1 variants each with a stepwise decrease in function ranging from the wildtype at 100% down to approximately 10% (Ohnishi et al., 2014). In addition, rheostat positions were found and examined in many other proteins including pyruvate kinase, LacI GalR transcriptional regulators, angiotensin-converting enzyme, and β-lactamase inhibitory protein (Meinhardt et al., 2013; Adamski and Palzkill, 2017; Hodges et al., 2018; Wu et al., 2019; Procko, 2020). The presence of rheostat positions in a variety of proteins further emphasizes the need to accurately predict their functional outcomes.

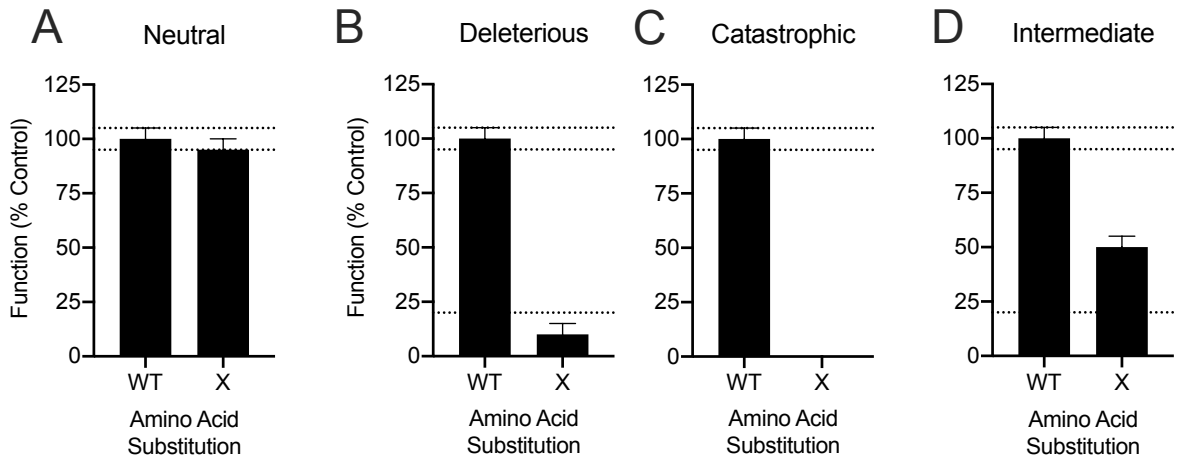


Figure 1-1. Simulated example of functional outcomes for single amino acid substitutions.

Simulated data were analyzed using GraphPad Prism to depict the described mutational outcomes: (A) neutral, (B) deleterious, (C) catastrophic, and (D) intermediate. Wildtype (WT) or unmutated protein is shown at 100% with an example standard deviation of 5. "X" indicates an arbitrary amino acid substitution. Upper horizontal lines denote the upper and lower limits of the artificial wildtype standard deviation to show differences from wildtype. The lower horizontal lines on panel B and D indicate a biologically relevant threshold.

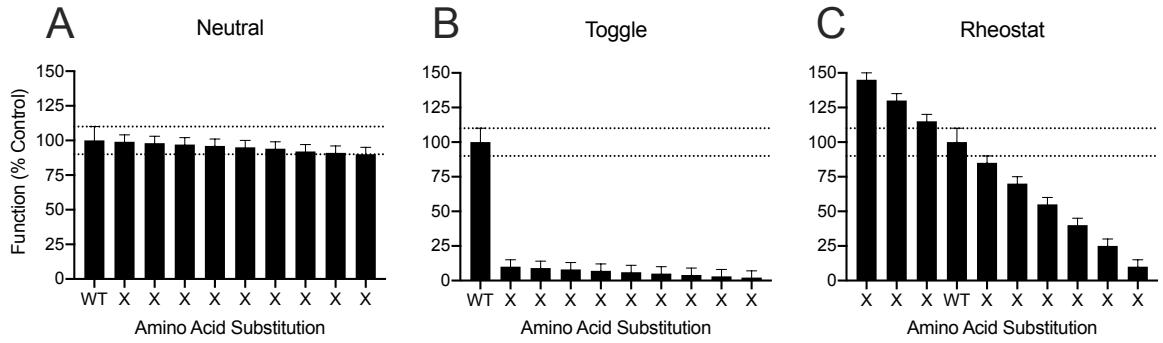


Figure 1-2. Simulated depiction of possible functional outcomes for an amino acid position.

Simulated data were analyzed using GraphPad Prism to depict the described amino acid position outcomes: (A) neutral, (B) toggle, and (C) rheostat. Wildtype (WT) or unmutated protein is shown at 100% with an example standard deviation of 5. "X" indicates an arbitrary amino acid substitution. Horizontal lines denote the upper and lower limits of the artificial wildtype standard deviation to show differences from wildtype.

Of these three classes of amino acid substitutions, intermediate are by far the newest and least understood and characterized. While the knowledge of rheostat positions is still in its infancy, they should be considered when predicting functional outcomes of mutations as they probably occur in many proteins. Consequently, computational prediction of the functional outcome of a mutation at a rheostat location is often inaccurate and mistaken for a neutral or detrimental mutation. Further, we are particularly interested in rheostat positions because of the potential implications for drug metabolism and drug response.

Computational Methods for the Prediction of Protein Function

As of 2013, there were 54 function-annotation algorithms available for the prediction of protein functional outcomes. Since then, numerous additional algorithms have been developed (Miller et al., 2017). These methods based their predictions on one or more of the following: amino acid sequence, evolutionary conservation of amino acid and their genomic context, protein structure and interaction with other proteins, and microarray data (Radivojac et al., 2013). Further, these programs or algorithms aimed to predict functional, structural, fitness, and pathogenesis outcomes of non-synonymous single-nucleotide polymorphisms. However, they only considered a few constraints and categories such as biological principles and pattern recognition techniques in addition to mathematical and computational approaches (Miller et al., 2017). As a result, predictions are limited to the biochemical structure of the replacement amino acids and the data collected from other species or studies, all while assuming they all cause the same or similar effects. While these computer algorithms took multiple sequence alignments into account, they never discriminated between the conserved and nonconserved locations. Further, they were only able to accurately predict functional outcomes at toggle positions (Miller et al., 2017). In

addition, recent studies on LacI/GalR homologs showed that nonconserved variants that are functionally important do not follow the same constraints or behaviors, either biochemically or otherwise, as mutations at conserved locations (Meinhardt et al., 2013). Moreover, mutations at nonconserved locations seem to have a remarkably diverse range of functional consequences. This variability in outcomes indicates that the binary, allowed versus detrimental, assumptions current computational programs are based on are invalid. In addition, there seems to be no commonality or consensus among the algorithms on the classification when a non-synonymous single-nucleotide polymorphism is neutral versus non-neutral. For instance, while one study deems a mutation neutral, another group may classify it as non-neutral. Furthermore, this "binary" type sorting has been found to be inappropriate in the case of rheostat positions (Miller et al., 2017).

III. **Liver Physiology and the Enterohepatic Circulation**

Overview of Liver Physiology

The liver, a major vital organ, is responsible for several essential functions including: the production of bile and plasma proteins like albumin; the metabolism of drugs, vitamins, hormones, and bilirubin; and the storage of vitamins (Kalra et al., 2020). To execute all these functions, the liver is made up of a complex network of cells, blood vessels, and bile ducts. Disruptions of any of these functions can lead to liver injury and toxic amounts of endogenous and exogenous compounds in circulation causing further bodily harm. Thus, complex networks and precise regulation is needed to maintain appropriate liver function.

Overview of Drug Absorption, Distribution, Metabolism, and Excretion

After a patient is administered a medication whether it be orally, topically, intravenously, etc., that pharmacological compound must undergo absorption, distribution, metabolism, and finally excretion (ADME). When a patient is given an oral medication, for example the cholesterol-lowering drug rosuvastatin, the drug tablet or capsule is first dissolved in the stomach and then released into the intestine. Once in the intestine, the drug is transported from the lumen into the enterocyte, the major cell type of the intestine. However, unlike bile acids, the drug is transported by a drug transporter such as OATP2B1. Rosuvastatin can then either be released back into the lumen of the intestine to be excreted or it may be secreted into the portal blood where it is then carried to the liver, rosuvastatin's major site of action (Nezasa et al., 2002).

Once at the liver, rosuvastatin is transported into the hepatocyte by active transport by several transporters like OATP1B1, 1B3, 2B1 and NTCP (Ho et al., 2006; Kalliokoski and Niemi, 2009). Inside the hepatocyte, rosuvastatin will be minimally metabolized by CYP P450 enzymes and uridine diphosphate glucuronosyltransferases (UGT) (McTaggart et al., 2001; Prueksaritanont et al., 2002; Schirris et al., 2015). It can then be secreted into the bile canaliculi via the breast cancer resistance protein (BCRP) or the multidrug resistance protein (MRP) 2 or it can be secreted back into the blood by MRP4 (Huang et al., 2006; Bowman et al., 2020). The majority of rosuvastatin is secreted into bile to then be eliminated from the body in feces. However, a small amount may be secreted back into the blood, to be eliminated via urinary.

Enterohepatic Circulation of Bile Acids

Bile acids are endogenous substances strongly associated with the liver and consequently xenobiotic metabolism. They are synthesized within hepatocytes from cholesterol via multistep reactions catalyzed by multiple enzymes beginning with either Cytochrome P450 (CYP) 7A1 (primary pathway) or CYP27A1 (an alternative pathway). The newly-formed bile acids are then conjugated to the hydrophilic amino acids taurine and glycine to decrease their toxicity and increase their solubility (Chiang, 2009; Chiang, 2013). After their synthesis, bile acids are transported out of the hepatocyte and into the bile canaliculi via transporters such as the bile salt export pump (BSEP) (Stieger et al., 2007). Next, they follow the biliary tree into the hepatic bile duct. From here they are stored and concentrated in the gallbladder until their release is stimulated after a meal by cholecystokinin (CCK) (Chiang, 2013). They then continue down the common bile duct through the sphincter of Oddi and are dumped into the duodenum of the small intestine where they aid in the digestion of fat and absorption of fat-soluble vitamins (Anwer and Stieger, 2014; Li and Chiang, 2014).

As they move through the small intestine, bile acids are either reabsorbed in the ileum by the apical sodium-dependent bile salt transporter (ASBT, *SLC10A2*; also known as ileal bile acid transporter or IBAT) or they proceed into the colon where they are deconjugated into secondary bile acids by the gut microbiota before being reabsorbed (Di Ciaula et al., 2017). Approximately five percent of bile acids are lost daily, however, most of the bile acids (~95%) are reabsorbed into the portal circulation and are then carried back to the liver. They then are transported into the hepatocytes primarily by the sodium-taurocholate cotransporting polypeptide (NTCP, *SLCO10A1*) (Claro da Silva et al., 2013) and secondarily by organic anion transporting

polypeptides (OATPs). This entire circulatory process is known as the enterohepatic circulation of bile acids (Li and Chiang, 2014).

Enterohepatic Circulation of Xenobiotics

Enterohepatic circulation exists for xenobiotics as well, with slight variations. An example of a drug that undergoes extensive enterohepatic circulation is ezetimibe. Like rosuvastatin, ezetimibe is an oral medication that is prescribed for hypercholesterolemia (Kosoglou et al., 2005). However, its' site of action is in the intestine, rather than the liver. In particular, ezetimibe inhibits the sterol transporter, the Niemann Pick C1-Like 1 protein (NPC1L1) (Levy et al., 2007). Further, once ezetimibe is absorbed into the intestinal mucosa, the majority of the drug (>95%) undergoes glucuronidation first-pass metabolism via UGT isoenzymes (Patrick et al., 2002; Soulele and Karalis, 2019). Parental ezetimibe and ezetimibe-glucuronide are then secreted into the portal blood and carried to the liver where they are further glucuronidated before being secreted into bile. After secretion into bile, ezetimibe-glucuronide moves through the biliary tree, the gallbladder, and eventually is dumped back into the lumen of the small intestine (de Waart et al., 2009). Here, ezetimibe-glucuronide is hydrolyzed back to ezetimibe, allowing for reabsorption back into the intestinal mucosa followed by glucuronidation, completing the enterohepatic circulation (Kosoglou et al., 2005). This process is repeated until the drug is eliminated through feces and urine (Malik et al., 2016).

Xenobiotics and other exogenous compounds get trapped in the enterohepatic circulation, resulting in longer half-lives. For example, a drug that does not undergo enterohepatic circulation, such as acetaminophen, may have a half-life of only a few hours (Prescott, 1980).

While ezetimibe has a half-life of approximately 22 hours (Patrick et al., 2002). Additional compounds may get trapped in the enterohepatic circulation for much longer. For example, perfluorinated compounds, which are environmental pollutants found in things such as stain repellants and non-stick cooking surfaces, can have half-lives of several years depending on their chain length (Genuis et al., 2010). As a result, it is important to know whether or not exogenous compounds, and especially xenobiotics, get trapped in the enterohepatic circulation due to the implications on dose frequency and toxicity.

Disruption of Enterohepatic Circulation

If enterohepatic circulation is disrupted, for instance by dysfunctional transporters like BSEP, it can lead to disease such as cholestasis (Stieger, 2010). Cholestasis increases bile acids in urine and blood and can cause more detrimental conditions such as hepatotoxicity, fibrosis, and/or cirrhosis (Claro da Silva et al., 2013). The transport of endo- and xenobiotics, via hepatic transporters has been extensively characterized. Furthermore, evidence suggests that, when at least one of these transporters contains a mutation and/or a polymorphism, transport function is altered. For example, specific OATP1B1 variants increase statin-induced myopathy susceptibility.

IV. **The Sodium Taurocholate Cotransporting Polypeptide**

Overview of the Sodium Taurocholate Cotransporting Polypeptide

The Na^+/Taurocholate Cotransporting Polypeptide (NTCP) is a sodium dependent transporter encoded by the solute carrier family 10 member 1 gene (*SLC10A1*) (Meier and Stieger, 2002; Anwer and Stieger, 2014). The SLC10 family has six additional members. These include the apical sodium-dependent bile acid transporter (also known as the apical bile salt transporter ASBT or ileal bile salt transporter IBAT, encoded by *SLC10A2*), the steroid sulfate carrier (SOAT, *SLC10A6*), and four orphan transporters: SLC10A3, SLC10A4, SLC10A5, and SLC10A7 (Doring et al., 2012). Since its cloning in 1994, much knowledge has been gained about NTCP. Human NTCP is a 56kDa glycosylated transporter protein made up of 349 amino acids configured into 9 transmembrane domains. It is located on the basolateral membrane in human hepatocytes as well as in rat pancreatic acini (Hallen et al., 2002; Hagenbuch and Dawson, 2004). Furthermore, NTCP has been identified and examined in a variety of species such as humans, rabbits, rats, mice, and hamsters. When compared to humans, rodents' NTCP have a 77 percent amino acid identity. Additionally, mice have two NTCP splice variants (Claro da Silva et al., 2013). While rodents are a widely utilized model for NTCP studies, these variations lead to species differences which should be taken into careful consideration (Hagenbuch and Dawson, 2004).

NTCP Homologs

Thus far, human NTCP has not been crystalized. Therefore, the majority of what is known about its structure has been discovered through experimental methods like hydrophobicity analysis,

alanine insertion scanning, and glycosylation site mutagenesis (Hagenbuch and Meier, 1994; Hallen et al., 2002). However, there are two published bacterial homologs of another SLC10A family member, ASBT, that have been crystalized allowing for a better understanding of the structure of the members of the SLC10A family. These homologs are from bacterial strains *Neisseria meningitis* and *Yersinia frederiksenii* each of which have two models that have been crystalized. *N. meningitis* was the first of the two to be crystalized and its models, 3ZUX and 3ZUY, only vary in that 3ZUY is a mutant of 3ZUX due to the addition of 8 amino acids in the C-terminus (Claro da Silva et al., 2013). The 3ZUX model shows the binding of two sodium ions along with a molecule of taurocholic acid in an inward-open conformation (Zhou et al., 2014). When compared to human ASBT, this model shares a 26% identity and 54% similarity and has an additional transmembrane domain (Claro da Silva et al., 2013; Anwer and Stieger, 2014). Its transport abilities when transformed into *Escherichia coli* cells showed similar capacity to transport taurocholate as ASBT and NTCP but with lower affinity than ASBT and comparable affinity to NTCP (Doring et al., 2012).

The other homolog, more recently crystalized from *Y. frederiksenii,* is very similar to *N. meningitis* in that it also shows ten transmembrane domains. However, what makes these structures stand out is that they have been crystallized in both the outward open and inward open conformations, 4N7X and 4N7Y respectively. Having both of these conformations crystalized allows for careful scrutiny of the differences between them and gives rise to a more complete visualization of how sodium and bile acids are translocated across the membrane via SLC10A transport proteins. These differences show discontinuous helices that make up a crossover region allowing for alternating access from either side of the membrane and thus the location and

orientation of taurocholic acid and potentially other bile acids as they are transported across the membrane (Zhou et al., 2014).

Transport of NTCP Substrates

While most substrates transported by NTCP are bile acids, it is considered a multi-specific transporter due to its ability to transport substrates such as sulfated steroids and thyroid hormones, as well as some xenobiotics such as statins. Additionally, NTCP was recently identified as a receptor for the Hepatitis B/D virus (Yan et al., 2012). NTCP-mediated transport is strictly sodium dependent and moves two sodium ions with each bile acid molecule. The sodium-dependency indicates that NTCP is a secondary active transporter that uses the electrochemical gradient required to move the negatively charged bile acids against their concentration gradient (Anwer and Stieger, 2014).

As indicated by its name, taurocholic acid or taurocholate is one of the main bile acids transported by NTCP and the substrate used in the initial cloning of NTCP. Taurocholic acid is used in many studies to quantify the function of NTCP and has a higher affinity for transfected human NTCP than NTCP in human hepatocytes (Doring et al., 2012). One sulfated steroid hormone transported by NTCP, also used to determine its function, is estrone-3-sulfate (E3S) (Claro da Silva et al., 2013). In addition, rosuvastatin, a cholesterol-lowering drug is commonly used as a xenobiotic representative for evaluation of NTCP function. Interestingly, up to thirty-five percent of rosuvastatin is transported into hepatocytes by human NTCP. This is another area where there is a species variation as this substrate is not transported by rat NTCP (Ho et al.,

2006) and interestingly when NTCP contains a specific polymorphism, rosuvastatin's uptake is increased (Pan et al., 2011).

Consequences of NTCP Disruption

NTCP accounts for over 80 percent of the transport of conjugated bile acids into hepatocytes across the basolateral membrane (Eloranta et al., 2006). Therefore, the regulation of NTCP transport and expression is crucial for the maintenance of the enterohepatic circulation of bile acids as well as for cholesterol homeostasis (Eloranta et al., 2006). If this homeostasis or enterohepatic circulation is interrupted, then bile acids (which can be cytotoxic) may begin to accumulate in hepatocytes leading to cholestasis. In humans, late-stage primary biliary cirrhosis, cholestatic alcoholic hepatitis, and drug-induced cholestasis lead to decreased NTCP expression via reduced mRNA or disrupted protein post-translational modifications (Stieger, 2011). To maintain nontoxic bile acid and cholesterol levels, transcription of transporters as well as enzyme levels require strict regulation.

Regulation of NTCP

Transcriptional regulation can be a slow and ongoing process mediated by receptor activation, signaling, cofactors, hormones and transcription factors. Receptors, mainly FXR and SHP, are activated by the binding of bile acids and, in turn, this downregulates bile acid transporters such as NTCP to prevent bile acid accumulation within hepatocytes (Denson et al., 2001; Wolters et al., 2002). Such binding causes a repression of bile acid influx transporters like NTCP while upregulating efflux transporters such as BSEP (Stieger, 2010). This process prevents the

intracellular accumulation of bile acids and helps to regulate the enterohepatic circulation of bile acids. Conversely, the glucocorticoid receptor can transactivate the expression of transporters such as NTCP (Claro da Silva et al., 2013).

Post-translational modifications are utilized when quick changes in the transporter membrane expression are necessary. This is accomplished by significantly altering the expression and transport capacity of NTCP by decreasing its translocation to the membrane, where it has been shown to localize within membrane rafts (Molina et al., 2008). NTCP is a phosphoprotein and needs to be dephosphorylated before it is translocated to the plasma membrane. While the mechanism in humans is not entirely clear, in rats, cyclic AMP drives an intracellular increase in calcium levels. The elevated calcium then initiates a signaling cascade ultimately leading to the dephosphorylation of NTCP and stimulation of vesicular trafficking to the membrane(Anwer and Stieger, 2014). As a result, the transport of bile acids into the cell increases. In addition to dephosphorylation, rodent NTCP is N-glycosylated at two asparagine sites (Hagenbuch, 1997). This glycosylation aids trafficking of NTCP to the membrane. When glycosylation is disrupted, so is NTCP's receptor activity for the Hepatitis B/D virus, indicating that glycosylation is necessary for the trafficking of NTCP to the cell membrane where it then acts as the receptor for the Hepatitis B/D virus (Appelman et al., 2017).

Conversely, to prevent the bioaccumulation of bile acids, a system of checks and balances facilitates retrieval of NTCP from the plasma membrane and decreases the transport of bile acids into hepatocytes. This retrieval is regulated by phosphorylation, which is stimulated by cholestatic bile acids, but the full mechanism is unknown (Anwer and Stieger, 2014). In addition to translocation, physical blockade of bile acid transport into the cell can be facilitated by nitric oxide binding to cysteine residues in a process known as S-nitrosylation (Schonhoff et al., 2011).

Most of what is known about post-translational modifications of NTCP was discovered through studies in rats. And while the human modulation of NTCP is assumed to be similar, much still needs to be evaluated to determine if this assumption is correct (Anwer and Stieger, 2014).

NTCP Polymorphisms

As with any DNA sequence, and consequently proteins, NTCP exhibits genetic variability. A few studies have investigated some of these genetic variants, all of which result in a disruption of NTCP function. The first and most common is a polymorphism found in Chinese populations where a serine at position 267 is replaced with a phenylalanine (c.800C>T, S267F) (Ho et al., 2004). This variant is known as NTCP*2. *In vitro*, this mutation results in reduced taurocholate transport efficiency but does not affect the transport of estrone-3-sulfate and, furthermore, increases the uptake of rosuvastatin (Ho et al., 2004; Choi et al., 2011). Clinically, in children homozygous for this mutation hypercholanemia was observed while in adults only a slight increase in serum bile acid levels was detected (Ho et al., 2004; Deng et al., 2016). It has also been suggested that S267F can lead to intrahepatic cholestasis during pregnancy (Chen et al., 2019). In addition, this polymorphism decreases the infectivity of the Hepatitis B/D virus and ultimately slows down the progression of HBV and HDV infection to liver cirrhosis and hepatocellular carcinoma (Binh et al., 2019).

The second most common mutation leading to NTCP deficiency is an isoleucine to threonine at position 88 (c.263T>C, I88T). This variant has an allele frequency of 0.67%, falling just short of the single-nucleotide polymorphism definition. Interestingly, isoleucine 88 is located at a relatively conserved amino acid position, and it was predicted to alter the structure of NTCP and

was thus expected to be pathogenic. In agreement with the predictions, the I88T mutation was discovered in neonatal patients and presents as elevated serum bile acids and increased indirect bilirubin, indicating a dysfunctional NTCP (Qiu et al., 2017).

Another missense variant, serine 199 to arginine (c. 595A>C, S199R) was found in four heterozygous patients in combination with the heterozygosity for the S267F polymorphism. This combination of mutations was found in two unrelated children and adults with mixed clinical outcomes. While all showed varying amounts of elevated serum bile acids, the pediatric patients showed a trend towards decreased total bile acids while the adults showed relatively normal total bile acids (Li et al., 2019). A fourth missense mutation was found in multiple genetic studies of children presenting with severe elevations of conjugated serum bile acids. It was demonstrated that a homozygous mutation led to the replacement of arginine at position 252 to a histidine (c. 755G>A, R252H) (Erlinger, 2015; Vaz et al., 2015). Interestingly, neither of these cases resulted in clinical manifestations commonly associated with hypercholanemia such as pruritis, liver abnormalities, or systemic cholestasis (Erlinger, 2015).

When taken together, these variants project a spectrum of NTCP deficiencies ranging from severe hypercholanemia to almost normal total bile acids. These few variants that result in NTCP deficiency have neither been well characterized nor predicted prior to clinical observation of elevated serum bile acids. In addition, most of them have not been characterized outside of their clinical relevance. Research associated with these case studies solely looked at the variations in the transport of select endo- and xenobiotics by the mutant NTCP but were not comprehensive. In addition, they did not consider whether these mutations resulted in differences of surface expression and/or transport kinetics.

V. **Concluding remarks for the introduction and study justification**

There is an obvious discrepancy in the prediction of mutations at rheostat positions and understanding their impact on personalized medicine. If personalized medicine is going to become a more widely-used and accepted approach to medical treatment, this discrepancy needs to be addressed and prediction algorithms need to be improved. In order to aid in the improvement of mutational outcome prediction, we need to understand rheostat positions by fully characterizing them. Thus, we chose to start with an important type of protein that is essential in drug metabolism: integral transmembrane transporter proteins. As a model transporter, we chose the Na^+/taurocholate cotransporting polypeptide to address three specific questions for this dissertation: **1) Do rheostat positions exist in integral transmembrane proteins like NTCP? 2) If they do exist, can we predict which locations are likely to be rheostat positions? 3) Further, can we determine which locations will not result in rheostatic outcomes?**

Specific Aims of this Dissertation

The overall goal of this dissertation research was to predict and characterize rheostat positions in an integral transmembrane protein to improve computational prediction of rheostat-like functional outcomes. Thus, this dissertation defends two specific aims.

Specific aim one tests the hypothesis that **rheostat locations exist in the integral transmembrane protein, the Na^+/taurocholate cotransporting polypeptide**. To address this aim, a polymorphic location within NTCP, position 267, that had previously shown substrate dependent alterations in the function of NTCP when serine was replaced by phenylalanine, was

selected. Site-directed mutagenesis was used to mutate serine 267 within the cDNA of NTCP to all other nineteen amino acids. Once mutated, the plasmids were transfected into HEK293 cells and the function and expression of the variants were characterized using radiolabeled uptake and surface biotinylation experiments. Select variants were then further evaluated using concentration-dependent uptake experiments to determine differences in Michaelis-Menten kinetic parameters. Finally, function and expression differences were compared to energy scores calculated based on homology models, to determine if these calculations could be used for future predictions of rheostat positions. The purpose of this study was to determine if a known polymorphism in NTCP was also a rheostat position, thus confirming their existence in an integral transmembrane protein. Furthermore, if the position acted as a rheostat, to characterize the alterations in NTCP caused by amino acid substitution.

Specific aim two was divided into two sub-aims. The first sub aim tests the hypothesis that the **simulated energy scores based on a structure model could predict protein stability and mutational tolerance indicating a potential rheostatic position**. To address this aim, we selected position 271 within NTCP for a few distinct reasons. First, and most importantly, calculated energy scores for variants at position 271 resembled a rheostat. However, the energy scores also predicted that multiple variants would be similar to wildtype. Next, this position was located close to S267, but it was further away from the substrate binding site. Lastly, multiple sequence alignments and ConSurf calculations revealed lower evolutionary conservation, indicating mutational tolerance. Once selected, we followed the same protocols as for aim one to evaluate the effect of amino acid replacement at position 271 on NTCP transport and expression.

The second sub-aim tests the hypothesis that **amino acid substitutions will not result in rheostatic outcomes at all locations within NTCP.** To confirm that not every amino acid

position in NTCP would result in rheostat-like behavior, we selected a highly conserved amino acid position, 102, where the native amino acid is a glycine in most species. The natural conservation and lack of variation at this position suggested that mutations would be detrimental, and we hypothesized that substitutions at position 102 would result in a toggle-like outcome. In addition, we selected position 146, an evolutionarily nonconserved position that is believed to be in an extracellular loop. Given the flexibility in extracellular loops and the ability for position 146 to accept numerous amino acids at this location naturally, we believed substitution with all 19 amino acids would result in an overall neutral outcome. Two examine these two positions, we mutated and transfected them into HEK293 cells as described in aim one and then completed transport and expression studies.

Chapter 2

Materials and Methods

The following methods were written by Shipra Malhotra and John Karanicolas: homology model for human NTCP and modeling human NTCP variant.

I. Materials

Materials for Transport Studies

Radiolabeled [^{3}H]-taurocholate (6.5 Ci/mmol) was purchased from PerkinElmer (Boston, Massachusetts). [^{3}H]-Estrone-3-sulfate (50 Ci/mmol) and [^{3}H]-rosuvastatin (10 Ci/mmol) were purchased from American Radiolabeled Chemicals (Saint Louis, Missouri). Taurocholic acid sodium salt (97% pure) and estrone-3-sulfate sodium salt (containing 35% Tris stabilizer) were purchased from Sigma Aldrich (Saint Louis, Missouri). Rosuvastatin (98% pure) was purchased from Cayman Chemicals (Ann Arbor, Michigan). Pierce™ BCA Protein Assay Kit was purchased from Thermo Fisher Scientific (Waltham, Massachusetts). Optiphase HiSafe 3 Scintillation Cocktail was also purchased from PerkinElmer.

Materials for Mutagenesis and Plasmid Isolation

QuikChange Lightening Site-Directed Mutagenesis kits were purchased from Agilent (Santa Clara, California). HiSpeed Plasmid Midi and Mini Kits were purchased from Qiagen (Hilden, Germany). One Shot® TOP10 Competent Cells were purchased from Thermo Fisher Scientific.

Materials for Cell Culture and Transporter Expression

Human Embryonic Kidney 293T/17 (HEK293) cells and Dulbecco's Modified Eagle's Medium were obtained from American Type Culture Collection (Manassas, Virginia). Hyclone Fetal Bovine Serum (Hyclone, Logan, Utah), Penicillin- Streptomycin (5,000 U/mL) and Opti-MEM™ Reduced Serum Medium were purchased from Thermo Fisher Scientific. FuGENE® HD Transfection Reagents were purchased from Active Motif (Carlsbad, California).

Materials for Surface Biotinylation and Western Blotting

EZ-link™ Sulfo-NHS-SS-Biotin and NeutrAvidin Agarose Resin beads were purchased from Thermo Fisher Scientific. cOmplete™ Protease Inhibitor Cocktail was obtained from Roche Diagnostics Corporation (Indianapolis, Indiana). 2-Mercaptoethanol and 4-20% Mini-PROTEAN® TGX™ Precast Protein Gels, 12-well, 20 µL were purchased from Bio-Rad Laboratories (Hercules, California). Anti-alpha 1 Sodium Potassium ATPase antibody was purchased from Abcam (Cat. No. ab7671; Cambridge, Massachusetts). Tetra-His Antibody, BSA-free was purchased from Qiagen (Cat. No. 34670). HRP conjugated goat anti-mouse secondary antibody (Cat. No. 31430) and SuperSignal West Pico Chemiluminescent Substrate were obtained from Thermo Fisher Scientific.

Additional analytical grade chemicals and reagents were purchased from commercial sources.

II. **Methods**

Mutagenesis Primer Design

Primers were designed to flank the codon of interest in human NTCP (i.e., Serine 267, Asparagine 271, Glycine 102 and Tyrosine 146) by approximately twenty nucleotides and were ordered from Invitrogen (Carlsbad, CA). The length of the primers was dependent on GC content and calculated melting temperature which was not to exceed 75°C. For the first round of insertions, the codon of interest was replaced by "NNN", meaning the codon's nucleotides were randomly inserted into the primer by Invitrogen and thus a "pool" of primers for random amino acid replacement was used. Following the mutagenesis reactions, resulting plasmid DNA was sequenced and it was found that this process typically yielded ten to twelve of the nineteen amino acids correctly inserted into the position of interest. Specific primers were then designed for the remaining seven to nine missing amino acids. If the same amino acid was randomly inserted using different nucleotides, the codon most commonly found in humans was preferentially selected (Castro-Chavez 2011). This was also true for selective primer design. All selected plasmids were sequenced by GENEWIZ (South Plainfield, New Jersey) prior to experimental use.

Site-directed mutagenesis

Mutagenesis reactions were completed using the QuikChange Lightning Multi Site-Directed Mutagenesis Kit. We used a previously-cloned His-tagged human NTCP in the pcDNA5/FRT vector for the mutagenesis as template. Reaction solutions were made on ice per the QuikChange Lightening Multi Site-Directed Mutagenesis Kit Instruction Manual; 2.5 µL 10x QuikChange

Lightening Multi reaction buffer, 100 ng NTCP cDNA, 100 ng mutagenic primers, 1 µL dNTPs, 1µL QuikChange Lightening Multi enzyme blend, and distilled H_2O to bring the final volume to 25 µL (note: QuikSolution was not used, and enzyme blend was added to the solution last). Solutions were then placed in a thermocycler and PCR amplified following the protocol in Table 2-1 below.

Table 2-1. Mutagenesis Thermocycling Parameters

# of Cycles	Temp in °C	Length of Cycle (Time)
1	95	2 minutes
30	95	20 seconds
	55	30 seconds
	65	2 minutes 50 seconds (30sec/kb of plasmid - NTCP ~ 5.6 kb)
1	65	5 minutes
1	4	∞ hold

After PCR cycles were complete, parental, non-mutated, DNA was digested using 1 µL *DpnI* restriction enzyme at 37°C for five minutes. Five microliters of mutated cDNA were then transformed into One Shot TOP10 cells following the chemical transformation procedure. Five microliters of mutant cDNA were added to 50 µL of thawed competent One Shot® TOP10 *E. coli* cells and incubated on ice for 30 minutes before heat shocking the reaction at 42°C for 30 seconds. Prewarmed SOC medium was then added to each reaction and incubated at 37°C while shaking at 100 rpm for one hour. Various amounts (5, 50, and 200 µL) of culture were then

added to individual prewarmed Luria-Bertani (LB) agar plates containing 100 µg/mL ampicillin (Sigma, Saint Louis, Missouri). Bacteria were then grown inverted overnight at 37°C.
The next morning bacterial colonies were randomly selected for amplification in liquid LB broth and plates with remaining colonies were wrapped in parafilm and stored at 4°C. Bacterial cultures were then grown overnight at 37°C while shaking. The cDNA was isolated from liquid cultures using QIAGEN mini prep kits and sequenced. Constructs containing the appropriate amino acid replacements were transfected into HEK293 cells for functional or expression studies as described below. Similar processes were followed to obtain more cDNA of variants when needed except for the use of Midi instead of Mini Prep Kits.

Cell culture

HEK293T/17 cells were cultured and maintained on 10 cm cell culture plates and grown at 37°C with 5% CO_2 prior to plating for experimental use. For plating, cells were first washed with 1xPBS (phosphate buffered saline) and then removed from the plate with 1x trypsin in PBS. Trypsinization was stopped by the addition of culture media, cells were removed from the plate by pipetting and then counted. The cells were then plated on poly-D-lysine coated flat bottom multi-well cell culture plates at the cell densities and volumes listed in Table 2-2. Cells were cultured in Dulbecco's Modified Eagle's Medium containing 100 U/ml penicillin, 100 µg/ml streptomycin and 10% fetal bovine serum for 24 hours prior to transfection. Experiments were then completed 48 hours post transfection.

Table 2-2. Cell Culture Plating Parameters

Plate	Cell density (cells/well)	Total volume per well
96-well	35,000	100 µL
48-well	100,000	300 µL
6-well	800,000	3 mL

Transient transfection of HEK293 cells

Empty vector, wildtype NTCP, and variant NTCP plasmids were transfected into HEK293 cells 24 hours after plating or once the cell density reached about 60-80 percent confluency. Transfections followed the FuGENE HD protocol obtained from Promega. All plasmids were transfected in triplicates for functional studies on 48-well plates and quadruplicate for 96-well plate experiments. For surface biotinylation experiments, single wells in 6-well plates were transfected. The amount of DNA transfected per well on 96-, 48- and 6-well plates was 100 ng, 280 ng, and 3 µg, respectively.

Initial uptake experiments

The uptake procedure has been optimized and is well established in our laboratory for functional studies (Zhao et al. 2015). Initial uptake experiments were completed on 48-well plates for positions S267, N271, and Y146 and 96-well plates for G102. Uptake buffer containing sodium (142 mM NaCl, 5 mM KCl, 1 mM KH_2PO_4, 1.2 mM $MgSO_4$, 1.5 mM $CaCl_2$, 5 mM glucose, and 12.5 mM HEPES, pH 7.4) was used for all washes and to prepare uptake solutions. Cells were washed three times with pre-warmed uptake buffer and then incubated for five minutes on a

heating block at 37°C with 0.3 μCi/mL of either [^{3}H]-taurocholate or [^{3}H]-estrone-3-sulfate, or 0.5 μCi/mL [^{3}H]-rosuvastatin which is equivalent to 30 nM, 5.8 nM, and 50 nM, respectively. Uptake was then terminated by washing with ice-cold uptake buffer. Cells were next solubilized in a 1% TX-100 solution in PBS. Radioactivity was measured using a liquid scintillation counter and protein concentrations were determined using Pierce BCA protein assays. Results were calculated by correcting for total protein, subtracting background uptake by cells expressing empty vector, setting wildtype NTCP as 100%, and then comparing all variants to wildtype as percent of control.

Surface biotinylation

HEK293 cells were plated in 6-well plates and transfected as described above. Forty-eight hours later, cells were incubated for 15 minutes on ice, and all solutions and buffers used were prechilled. Each well was washed twice with PBS and then incubated for one hour at 4°C while rocking with 1 mg/mL EZ-Link NHS-SS-Biotin in PBS. Cells were then washed twice with PBS and incubated for twenty minutes at 4°C with 100 mM glycine in PBS while rocking. Cells were then washed a final two times with PBS before being lysed using a lysis buffer (10 mM Tris, 150 mM NaCl, 1 mM EDTA, 0.1% SDS, 1% Triton X-100, in H_2O pH 7.5, and 1 X protease inhibitors (cOmplete protease inhibitor cocktail, Sigma-Aldrich)). The lysed cells were then removed from the plate using a pipette and collected into 1 mL Eppendorf tubes. The lysates were centrifuged at 10,000 x g for two minutes to remove cellular debris. One hundred microliters of the supernatants were transferred to PCR tubes to be used for total protein analysis and the remaining sample was transferred to 1 mL Eppendorf tubes containing 150 μL of NeutrAvidin Agarose Resin bead slurry, prewashed with lysis buffer. Tubes were incubated for

one hour at room temperature using end-over-end rotation. Then, the beads were washed with lysis buffer three times for five minutes each with end-over-end rotation. Captured proteins were eluted from the beads using 1 X SDS sample buffer (diluted from 5 X SDS sample buffer- 10% w/v sodium dodecyl sulfate, 20% v/v glycerol, 0.2 M Tris-HCl pH 6.8, 0.05% w/v bromophenol blue) containing 5% β-mercaptoethanol and 1 X protease inhibitors at 70°C for 10 minutes and collected by centrifugation at 850 x g for five minutes. Samples were then either immediately used for western blotting or stored at -80°C.

Western blotting

Surface biotinylation samples were heated to 50°C for 10 minutes before separation using 4-20% polyacrylamide gradient gels. If total protein samples were analyzed, they were diluted with 5 X SDS sample buffer containing 5% β-mercaptoethanol and then also heated to 50°C for 10 minutes before separation using 4-20% polyacrylamide gradient gels. Samples were separated using a constant voltage of 150 V for one hour or until the dye front ran out of the gel. After separation, proteins were transferred to nitrocellulose or PVDF membranes using Invitrogen's Power Blotter System. Blots were blocked with 5% non-fat milk in Tris-buffered saline containing 0.1% Tween 20 (TBS-T) for one hour at room temperature while rocking. Following blocking, blots were incubated overnight at 4°C with a combination of a mouse antibody against the alpha subunit of Na^+/K^+-ATPase and a mouse antibody against the His-tag found on NTCP (both at a 1:2,000 dilution) in blocking solution on a rocker. The next day blots were washed three times with TBS-T and once with TBS before incubation with an HRP conjugated goat anti-mouse secondary antibody at 1:10,000 in 2.5% milk in TBS. After one hour at room temperature while rocking, blots were washed three times with TBS and incubated for five minutes (not

rocking) with SuperSignal West Pico Chemiluminescent substrate. Blots were visualized using a LI-COR Odyssey Fc (LI-COR, Lincoln, NE) and bands were quantified using their Image Studio Lite Quantification Software.

Time dependency and kinetics experiments

The basic uptake method outline in "Initial Uptake Experiments" section was followed for time and concentration dependent experiments. Initial linear rates were determined for wildtype and select variants using multiple time points between 10 seconds and 10 minutes at low and high concentrations of each substrate: taurocholate at 0.1 μM and 100 μM; estrone-3-sulfate at 1 μM and 200 μM; and rosuvastatin at 5 μM and 500 μM (Table 2-3). Kinetics for wildtype and select variants were determined using HEK293 cells plated on 48-well plates. Uptakes were performed using either sodium containing or sodium-free uptake buffers (136 mM NaCl was replace by 136 mM choline chloride) and after correction for total protein and surface expression, results were analyzed using GraphPad Prism 8 (Michaelis-Menten kinetics).

Table 2-3. Time points used for kinetic experiments

	Substrate		
	Taurocholate	Estrone-3-Sulfate	Rosuvastatin
NTCP Wildtype	30 sec	1 min	1 min
S267F	30 sec	20 sec	30 sec
S267N	30 sec	10 sec	30 sec
S267W	30 sec	10 sec	30 sec
N271C	30 sec	20 sec	30 sec
N271H	30 sec	20 sec	30 sec
N271L	30 sec	20 sec	30 sec

Homology model for human NTCP

To model the human NTCP sequence (NCBI: NP_003040), structural models were constructed using the SWISS-MODEL automated protein modeling server (https://swissmodel.expasy.org/) (Arnold et al. 2006, Benkert et al. 2011, Biasini et al. 2014). To model the inward-open conformation of NTCP, we used as a template the structure of a bacterial homologue from *Neisseria meningitidis* (PDB 3ZUY, 25% sequence identity with NTCP) (Hu et al. 2011). To model the outward-open conformation of NTCP, we used as a template the structure of a bacterial homologue from *Yersinia frederiksenii* (PDB 4N7X, 26% sequence identity with NTCP) (Zhou et al. 2014). As with the templates upon which the models were based, the topology of each model comprises 9 transmembrane (TM) helices linked by short loops into core (TM 3,4,5,8,9 and 10) and panel domains (TM 2,6 and 7), along with the substrate-binding intracellular crevice (Figure 3-6). The models of NTCP lack the N-terminal sequence

corresponding to TM1 of the bacterial ASBT crystal structures; to keep the helix numbering consistent with the known crystal structures, we have chosen to start the NTCP helix numbering with TM2.

Modeling human NTCP variants

To facilitate structural exploration in response to sequence variants, we began by using the "relax" protocol (Tyka et al. 2011, Nivon et al. 2013, Conway et al. 2014) in the Rosetta macromolecular modeling suite (Leaver-Fay et al. 2011) to generate a close structural ensemble from each of the two homology models provided by SWISS-MODEL. Each starting conformation was used to carry out 1000 independent simulations and the top-scoring 100 output structures were retained as a representative ensemble for the (wildtype) inward-open or outward-open state.

To build a structural model of a given NTCP sequence variant, we used the "ddG" protocol in Rosetta (Kuhlman et al. 2003, Kellogg et al. 2011, Leaver-Fay et al. 2011). With respect to our study, this protocol was used to introduce the desired amino acid substitution at a selected position and then iterated between optimization of the nearby sidechains and optimization of the backbone. We applied this protocol 10 times to each of the 100 members of our (wildtype) structural ensemble to yield 1000 models of the desired sequence variant. To avoid potential sampling artifacts from drawing the conformation/energy from the single lowest energy conformation sampled, we ranked all conformations for a given sequence variant based on Rosetta energy and carried forward the 50th-best conformation (i.e., 95th percentile) as the representative.

The same process was repeated for each of the 20 amino acids, to generate all possible sequence variants at this position. While the starting models already included the correct amino acid at the positions of interest, the same protocol was nonetheless applied to introduce the wildtype amino acid; this ensured that any structural/energetic changes were indeed due to sequence variations and not simply changes relative to the starting structure induced by the modeling protocol. The same analysis was separately completed using the structural ensemble for the inward-open state and the structural ensemble for the outward-open state.

All Rosetta calculations were carried out using git revision 0e7ed9fd3cd610f2a7c9f3bdcaba64a9b11aab0d of the developer master source code. The two homology models originally provided by SWISS-MODEL were also used as input for energy calculations using FOLDX version 4 (Schymkowitz et al. 2005).

Multiple Sequence Alignment

A multiple sequence alignment of NTCP orthologs from a wide variety of species, including both prokaryotes and eukaryotes, was created using the PSI-BLAST (Basic Local Alignment Search Tool) algorithm (Altschul et al. 1997) from the National Center for Biotechnology Information. Eight separate searches were completed in July 2017. The first six searches were completed using the human SLC10A family members (Claro da Silva et al. 2013) as query sequences: NTCP (*SLC10A1*), ASBT (*SLC10A2*), P3 protein (*SLC10A3*), sodium/bile acid cotransporter 4 (*SLC10A4*), P5 (*SLC10A5*), and SOAT (*SLC10A6*). Several homologs were retrieved by two or more of the searches, which facilitated the later combination of these 6 alignments into one (below). The remaining BLAST searches were completed using two NTCP

bacterial homologs, one from *Neisseria meningitidis* (pdb 3ZUY; (Hu et al. 2011)) and one from *Yersinia frederiksenii* (pdb 4N7X; (Zhou et al. 2014)) as queries, each of which generated thousands of search hits with sequence identities spanning 39%-99%. Following each of the 8 BLAST searches, the resulting sequences were aligned using Clustal Omega (Madeira et al. 2019).

Next, to relate the eukaryotic and prokaryotic sequence alignments, the structural model built for human NTCP was super-imposed onto the 4N7X crystal structure of the bacterial homolog using UCSF Chimera (Pettersen et al. 2004). A structure-based sequence alignment was created from these reference proteins. This alignment was used in combination with PROMALS3D (Pei et al. 2008) to align the human SLC10A family members to the crystal structure. Finally, using sequences that were present in one or more alignments as guide sequences, the algorithm MARS (Parente et al. 2015) was used to combine the eukaryotic and prokaryotic sequences into one file without perturbing the structure-based alignments.

During these processes, sequences were curated by removing (i) duplicate sequences, (ii) particularly long or short sequences (as compared to human NTCP), (iii) sequences with large deletions, or (iv) those with ambivalent amino acid assignments. To prevent over-representation and facilitate proper sampling for subsequent calculations, bacterial sequences were clustered by their sequence identities; clear clusters were evident at the ~80% thresholds. Each of these clusters was randomly sampled to select no more than five sequences for inclusion in the final alignment of 1561 sequences (sequence alignment file can be found in the supporting information section of the publication Ruggiero et al., 2021 or at this link: https://www.jbc.org/cms/10.1074/jbc.RA120.014889/attachment/f7084a93-617b-42f4-8c48-7258a6d855aa/mmc2.txt).

Sequence entropy for the combined, sampled alignment was calculated using BioEdit v 7.0.5.3 software which uses the equation listed below (Tippmann 2004).

$$H(l) = -\sum f(b,l)\ln\big(f(b,l)\big)$$

H(l) is the entropy or uncertainty at an amino acid position, *l*, where the possible residue, *b*, is located. *f(b,l)* represents the frequency that an amino acid, *b*, is located at position, *l*. The sequence entropy calculates the amount of amino acid variability that occurs at each residue within the multiple sequence alignment. The higher the entropy score the greater the variability. For our calculations the highest entropy score was 2.8 and the lowest was zero. The lowest entropy score of zero indicates that there was only one amino acid that was inserted into that position.

We also utilized the ConSurf web server to further analyze our multiple sequence alignment (http://consurf.tau.ac.il) (Ashkenazy et al., 2016). ConSurf considers an additional factor that sequence entropy does not: phylogenetic branching. ConSurf considers the phylogenetic tree branching to analyze evolutionary patterns to determine which amino acids and protein regions are important for protein function and structure. ConSurf then scores each amino acid position on a scale of 1 to 9. The higher the score the greater the correlation to phylogenetic branching and the greater the evolutionary conservation (Ashkenazy et al., 2016).

RheoScale Calculations

To calculate and compare the rheostatic outcomes for each position, we utilized the RheoScale tool described in (Hodges et al. 2018). This tool compares inputted data sets to each other and to

a single wildtype value with its standard deviation and then calculates neutral, rheostat, and toggle scores by separating the data into histogram bins. The number of bins is empirically selected by the user for their data set. In the histogram, one bin includes the wildtype value, and variants that fall into this bin would be functionally similar to wildtype (i.e., neutral); if more than 70% of substitutions fall into this bin, the position is considered to be neutral (Martin et al., 2020). Another bin includes the "dead" (non-functional) variants; if two-thirds of the variants fall into this bin, the position is considered to be a toggle position (Wu et al., 2019). The remaining bins allow for sorting of variants with either intermediate or enhanced outcomes. The more evenly distributed the variants are in the bins, the greater the rheostatic score and the closer to 1. For example, if there are 10 bins selected to evaluate 10 variants and each variant falls into a different bin including one in the wildtype and one in the "dead" bin then this hypothetical position would have a rheostatic score of 1. If the substitutions for a position sample at least half of the accessible range, the position is considered a rheostat position (Hodges et al., 2018).

The number of intermediate bins chosen has a significant effect on the score. The total observed range of outcomes, the error of the data, and the number of variants must be considered. For our analyses, the data sets used were the initial uptake data corrected for surface expression for N271 (Figure 4-4), G102 (Figure 4-11), Y146 (Figure 4-14) and the surface expression data from the previous published position 267 (Chapter 3- Figure 3-3). We compared the function of each position for each substrate. For example, the taurocholate surface corrected initial uptake data from each of the four positions were used and compared for one RheoScale calculation. For these *in vitro* data, the primary limitation was the size of the error bars on the measurements. Therefore, the recommended number of bins ranged from 7 to 11. To be consistent among our comparisons, we selected to override the number of bins and selected 10 bins for all rheostat and

toggle calculations. To accommodate the full distribution of wildtype values obtained over the course of all experiments, neutral scores were determined using 5 bins with a bin size twice as wide (Martin, Wu et al. 2020). We also varied the bin number between 5 and 11 and confirmed that it did not affect the overall conclusions (the same positions fell above the significance thresholds).

However, as mentioned, the RheoScale calculator compares inputted multiple data sets to each other and to a single wildtype value with its standard deviation. Thus, when considering a single data set, for example S267 variant surface expression, the wildtype standard deviation used for the RheoScale calculations may be more or less than the standard deviation for the wildtype for the specific data set in question. Therefore, the RheoScale calculations may improperly include or exclude variants as wildtype, thus giving an inaccurate neutral score. We attempted to combat this error with bin variation. However, to ensure our calculations accurately captured neutral variants, neutral scores were also manually calculated by visual observations of each individual data set. The variants whose average values fell within the wildtype standard deviation were considered the same as wildtype and thus "neutral". An overall neutral score for each position and experimental data set was then calculated by dividing the number of counted "neutral" variants by the total number of viable variants in the data set.

The "dead" value was not overridden unless the minimum data value fell below 2 and the maximum value was only overridden to 400 if the maximal value was greater than 400. In addition, the rheostat values reported in Table 4-2 were weighted, thus bins farther from wildtype and dead got more weight than bins close to those values. All minimum, maximum, "dead" values, and calculator recommended number of bins for each can be found in Supporting Table 4-2.

For the surface expression calculations, all variants were used for each position. If expression was less than one, the value was replaced with 1 and the nonfunctional protein value ("dead") was set to 1. For surface corrected initial uptake data, the values for variants that are not expressed at the surface were excluded from the calculations because "uptake" values are meaningless if the protein is not expressed at the plasma membrane. This was the case for G102I, K, L, R and V, and for Y146N and S.

Statistical analysis

Calculations were performed using GraphPad Prism 8 (GraphPad Software Inc., San Diego, CA). Correlation was evaluated using Pearson and Spearman correlation coefficients. Significance was determined using One-way ANOVA followed by Dunnett's post hoc test for multiple comparisons. Results were considered significantly different at $p < 0.05$.

Data availability Statement

PDB coordinates for the two homology models are deposited and freely available on Mendeley (http://dx.doi.org/10.17632/spt9jkgy2y.1), along with the single lowest-energy Rosetta model for each S267 variant in both the Inward-open and Outward- open conformations.

Chapter 3

A clinically-relevant polymorphism in the Na^+/taurocholate cotransporting polypeptide (NTCP) occurs at a rheostat position

This chapter was previously published as an open access article and is reprinted here with adaptations:

Ruggiero MJ, Malhotra S, Fenton AW, Swint-Kruse L, Karanicolas J, Hagenbuch B. A clinically-relevant polymorphism in the Na^+/taurocholate cotransporting polypeptide (NTCP) occurs at a rheostat position. J Biol Chem. 2020 Nov 9:jbc.RA120.014889. doi: 10.1074/jbc.RA120.014889. Epub ahead of print. PMID: 33168628

Contributions from John Karanicolas, Shipra Malhotra, Liskin Swint-Kruse, Aron Fenton and Bruno Hagenbuch in this chapter include text and figures. Detailed contributions were as follows: computational modeling of NTCP, structure-based Rosetta energy score calculations, FoldX calculations, and corresponding figures were completed by Shipra Malhotra and John Karanicolas. Rosetta energy scores completed by these collaborators were then used for subsequent correlation studies. Figures completed by collaborators are denoted in the figure legends.

I. Introduction

Amino acid substitutions are commonly used to evaluate which amino acids in a protein contribute to its function. Several decades of studies have led to conventional “rules” for mutational outcomes that are now included in many textbooks and are often implicitly or explicitly assumed in the design and interpretation of experimental studies. For instance, at “important” protein positions, only amino acids with biochemical properties similar to the wildtype (WT) are expected to allow function, whereas other amino acid substitutions are

expected to abolish function or structure. However, the mutational studies that gave rise to these rules were primarily focused on evolutionarily conserved amino acid positions (Gray et al., 2012). When substitution studies of less conserved positions were completed, results were seldom consistent with expected outcomes. Instead of an on/off pattern, when nonconserved positions were substituted with a variety of amino acids, each substitution had a different outcome. The fact that one position could be substituted to access a continuum of functional outcomes is analogous to an electronic dimmer switch; therefore, these positions have been labeled as "rheostat" positions (Meinhardt et al., 2013; Hodges et al., 2018; Wu et al., 2019).

To date, biochemical studies of rheostat positions have been limited to a few positions within a few proteins. As of yet, there is insufficient data to demonstrate how widespread such positions are in the protein universe or their general properties. In an effort to expand general knowledge of rheostat positions, the integral membrane transport protein human Na^+/taurocholate cotransporting polypeptide was chosen as a model system, which allowed three specific areas of interest to be addressed.

First, we were curious whether rheostat positions were limited to the soluble-globular class of proteins in which they were discovered (Meinhardt et al., 2013), or if they also exist in transmembrane (TM) proteins. Because soluble and integral membrane proteins evolved under different chemical environments, the properties of one class are not always transferable to the other. In the context of rheostat positions, it helps to know in which types of proteins to expect them. Predictions about substitutions at rheostat positions require different algorithms than predictions at positions that follow textbook substitution rules (Miller et al., 2017; Miller et al., 2019).

Second, if rheostat positions do exist in integral membrane proteins, we wondered whether the functional outcomes arising from various substitutions were dependent on the substrate being transported. That is, we wished to explore the effects of substitutions at rheostat positions on substrate specificity. Although our prior work suggested such substitutions could have complex effects on specificity (Tungtur et al., 2019), that work was carried out with non-natural proteins and ligands. The present study, using a natural protein known to transport multiple substrates, provided opportunity to further document this complex substitution outcome. Amino acid changes that alter substrate specificity are key to the evolution of functional variation and may thus also give rise to different drug sensitivities.

Third, many rheostat positions identified to date are located outside binding sites and do not directly contact ligand or substrate. As such, their molecular mechanisms of action have been difficult to explain (Fenton et al., 2020). Recent studies of a rheostat position in a transcription repressor have indicated that substitutions may alter protein dynamics (Campitelli et al., 2021). Thus, the present study provided opportunity to relate the continuum of functional outcomes of the rheostat variants to the complex conformational changes experienced by an integral membrane protein during transport.

Several features of NTCP facilitated the studies listed above: This protein is expressed at the basolateral membrane of human hepatocytes where it plays an important role in the enterohepatic circulation of bile acids (Hagenbuch and Dawson, 2004). In addition to conjugated bile acids such as TCA, NTCP mediates the uptake of other substrates into hepatocytes, including the hormone metabolite estrone-3-sulfate and several statins such as rosuvastatin (Claro da Silva et al., 2013). Furthermore, several single-nucleotide polymorphisms alter NTCP transport activity

(Ho et al., 2004), which we reasoned might enable the identification of rheostat positions from the 349 positions that comprise this protein.

From various analyses, we identified position 267 as a potential rheostat position in NTCP. Next, we assessed the function of WT NTCP and all 19 amino acid substitutions at position 267 with cellular uptake studies. We also determined whether substitution outcomes were substrate dependent by measuring transport of TCA, estrone-3-sulfate, and rosuvastatin. Additional experiments differentiated the effects of substitutions on protein surface expression and transport kinetics. Finally, we used homology modeling of the available "inward-open" and "outward-open" conformations and energetic calculations to explore the "rheostatic" relationship between protein stability and surface expression.

II. **Results**

Because generalizable features for identifying rheostat positions have not yet been validated, we first faced the challenge of identifying a likely candidate among the 349 amino acid positions of NTCP. Thus, we combined two types of information–functional insights and sequence analysis–to identify a likely rheostat position for the present study.

The functional information was derived from knowledge of NTCP polymorphisms with clinical consequences. The most frequent and best characterized of these is NTCP*2, which leads to the missense amino acid substitution S267F and has an allele frequency of 7.5% in Chinese Americans. Previously published data for S267F indicated reduced transport of TCA, WT-like transport of estrone-3-sulfate, and increased transport of rosuvastatin *in vitro* (Ho et al., 2004; Ho et al., 2006; Pan et al., 2011). Clinically, this mutation results in severe hypercholanemia with

total serum bile acid levels of about 15- to 70-fold above normal in homozygous pediatric patients (Deng et al., 2016; Dong et al., 2019); some of the patients also had elevated liver enzymes, jaundice and gallstones (Dong et al., 2019). In homozygous adult patients, NTCP*2 resulted in total serum bile acid levels two- to fivefold above normal (Liu et al., 2017).

The other prevalent polymorphism is NTCP*3, which is found in 5.5% of African Americans and results in the amino acid substitution I223T. However, the resulting protein expression at the plasma membrane was significantly reduced compared with WT NTCP (Ho et al., 2004). Inadequate plasma membrane expression would make functional studies problematic; thus, this single-nucleotide polymorphism was not further explored in the current work.

Polymorphic positions are, by definition, nonconserved. Likewise, the previously identified rheostat positions in the soluble LacI/GalR homologs were also nonconserved and had moderate to high phylogeny scores (Meinhardt et al., 2013; Tungtur et al., 2019). Thus, to further strengthen our reasoning that NTCP position 267 should be explored as a potential rheostat position before embarking on experiments, we assessed these evolutionary properties. Position 267 was previously reported to be highly conserved among many different animals, including primates, rodents, dogs, cats, horses, chickens, several fishes, and marine chordates (Deng et al., 2016). However, when we expanded the sequence alignment to include 1561 homologs of the solute carrier family 10A (SLC10A) (Claro da Silva et al., 2013) representing all kingdoms of life (see NTCP sequence alignment file in the section of the publication Ruggiero et al., 2021 or at this link: https://www.jbc.org/cms/10.1074/jbc.RA120.014889/attachment/f7084a93-617b-42f4-8c48-7258a6d855aa/mmc2.txt), position 267 had a sequence entropy of 1.54. This intermediate conservation score (overall range of 0.0 to 2.8) suggested that position 267 could

tolerate multiple substitutions without catastrophic outcomes, which is requisite for most amino acid substitutions at rheostat positions.

When we further analyzed the expanded sequence alignment with ConSurf (Ashkenazy et al., 2016), position 267 had a score of 8 (on a scale of 1 to 9), indicating a pattern of change that highly correlated with the branching of the SLC10A phylogenetic tree. This characteristic has been hypothesized to be indicative of positions important for evolving functional variation, (e.g. (Ye et al., 2008; Mazin et al., 2010)) which may be a key biological role of rheostat positions (Meinhardt et al., 2013; Fenton et al., 2020).

Cellular substrate transport by S267 variants

To experimentally ascertain whether position 267 was a rheostat position, we replaced serine with all other 19 amino acids and measured uptake of TCA (Figure 3-1, top), estrone-3-sulfate (Figure 3-1, middle), and rosuvastatin (Figure 3-1, bottom) for each substitution. On the left side of Figure 3-1, uptake is shown with the amino acids in alphabetical order, with WT at far left. The right panels show results ordered from highest to lowest transport, with WT placed within the series.

Notably, for all three substrates, some variants transported substrates better than WT, whereas others had diminished transport. The range of observed changes spanned several orders of magnitude. Thus, position 267 exhibited definitive rheostatic substitution behavior. Furthermore, the rank-order of the amino acid substitutions differed among the three substrates. This is further discussed later, but here, we particularly note that the NTCP*2 polymorphism, S267F, showed a significant decrease in TCA transport, a 2.0-fold increase for estrone-3-sulfate transport, and 2.5-

fold increase for rosuvastatin transport. Previous reports showed decreased TCA transport, estrone-3-sulfate levels similar to WT, and increased rosuvastatin transport (Ho et al., 2004; Ho et al., 2006; Pan et al., 2011). The discrepancy for estrone-3-sulfate could arise from the slightly different uptake conditions including substrate concentrations, incubation time, and/or different cell lines used in the two studies.

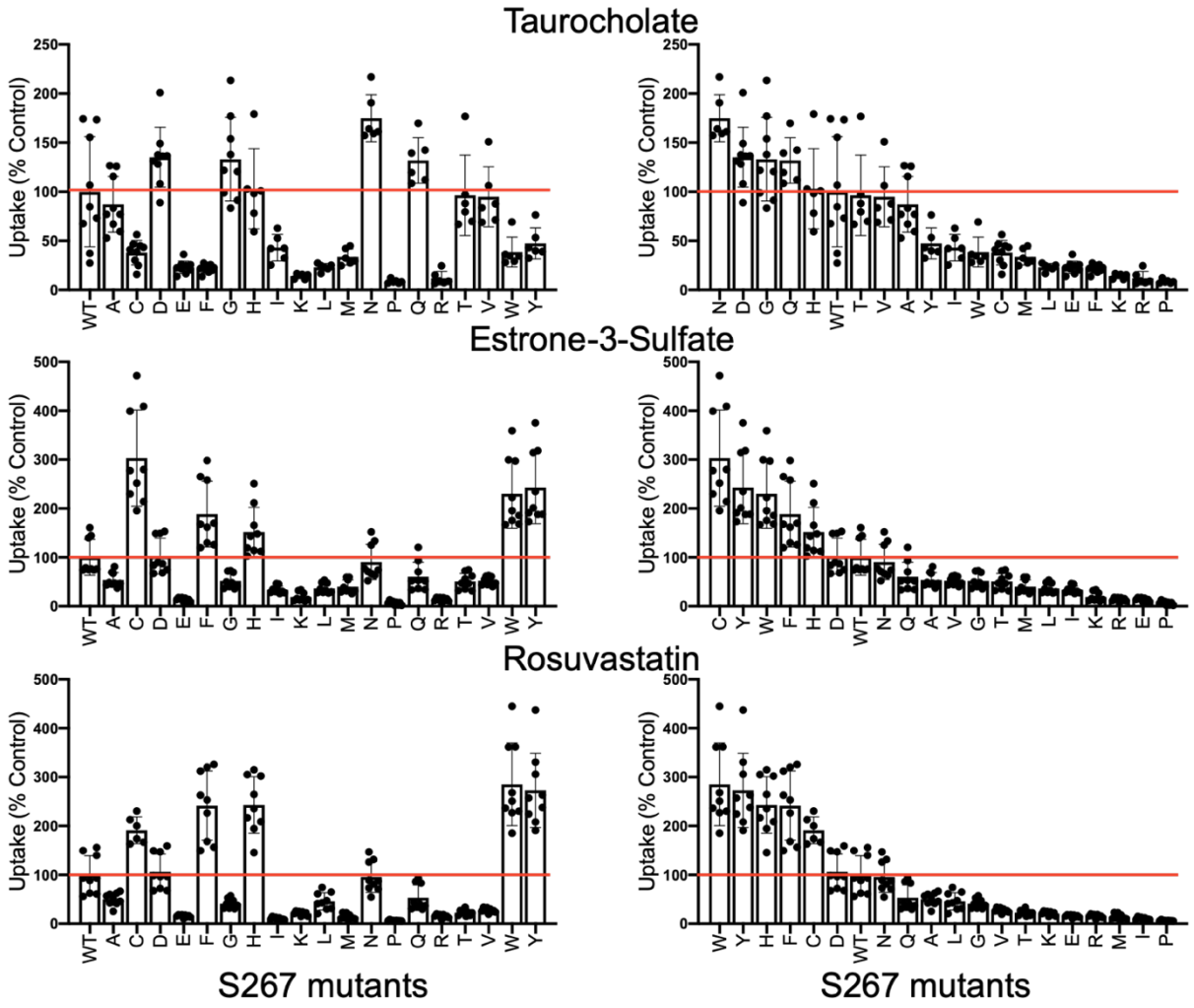


Figure 3-1. Substrate uptake by WT NTCP and S267 variants.

Uptake of ^{3}[H]taurocholate (30 nM), ^{3}[H]estrone-3-sulfate (5.8 nM) and ^{3}[H]rosuvastatin (50 nM) was measured for 5 min at 37°C, 48 hours after transfection of wildtype NTCP, its S267 variants, and empty vector into HEK293 cells. Net uptake was obtained by subtracting the uptake of cells transfected with empty vector from uptake of NTCP-expressing cells. The left-hand side shows the results ordered alphabetically based on the amino acid replacement, and the right-hand side shows the substitutions ordered from highest to lowest transport activity. Results were calculated as percent of WT NTCP. Individual data points as well as the mean ± SD are reported from n=3 biological replicates (each with 2-3 technical replicates) for all but rosuvastatin uptake by S267C for which n=2 biological replicates are shown. Horizontal lines to aid visual inspection correspond to WT values, which were set to 100%; results of statistical analyses are shown in Supporting Table 3-1. NTCP, Na^{+}/taurocholate cotransporting polypeptide; WT, wildtype.

Dissecting the composite cellular outcomes of S267 variants

Protein substitutions can alter substrate transport kinetics, substrate specificity, protein stability, and/or intracellular trafficking to the outer membrane. The cellular uptake assay is sensitive to changes in any of these parameters. Thus, we devised experiments to dissect the S267 variants' functional and structural contributions.

To assess the combined effects of trafficking and stability, we quantified differences in surface expression of NTCP variants using surface biotinylation experiments followed by western blotting (Figure 3-2*A*). When normalized for the loading control Na^+/K^+ATPase, the expression levels varied between 22.5% for S267Q and 153% for S267W, corresponding to an overall variation of about sevenfold. The expression of most of the other variants was similar to WT (Figure 3-2*B*). Thus, NTCP appeared to accommodate different amino acid side chains at position 267 without significantly disrupting the overall structure or trafficking to the cell surface.

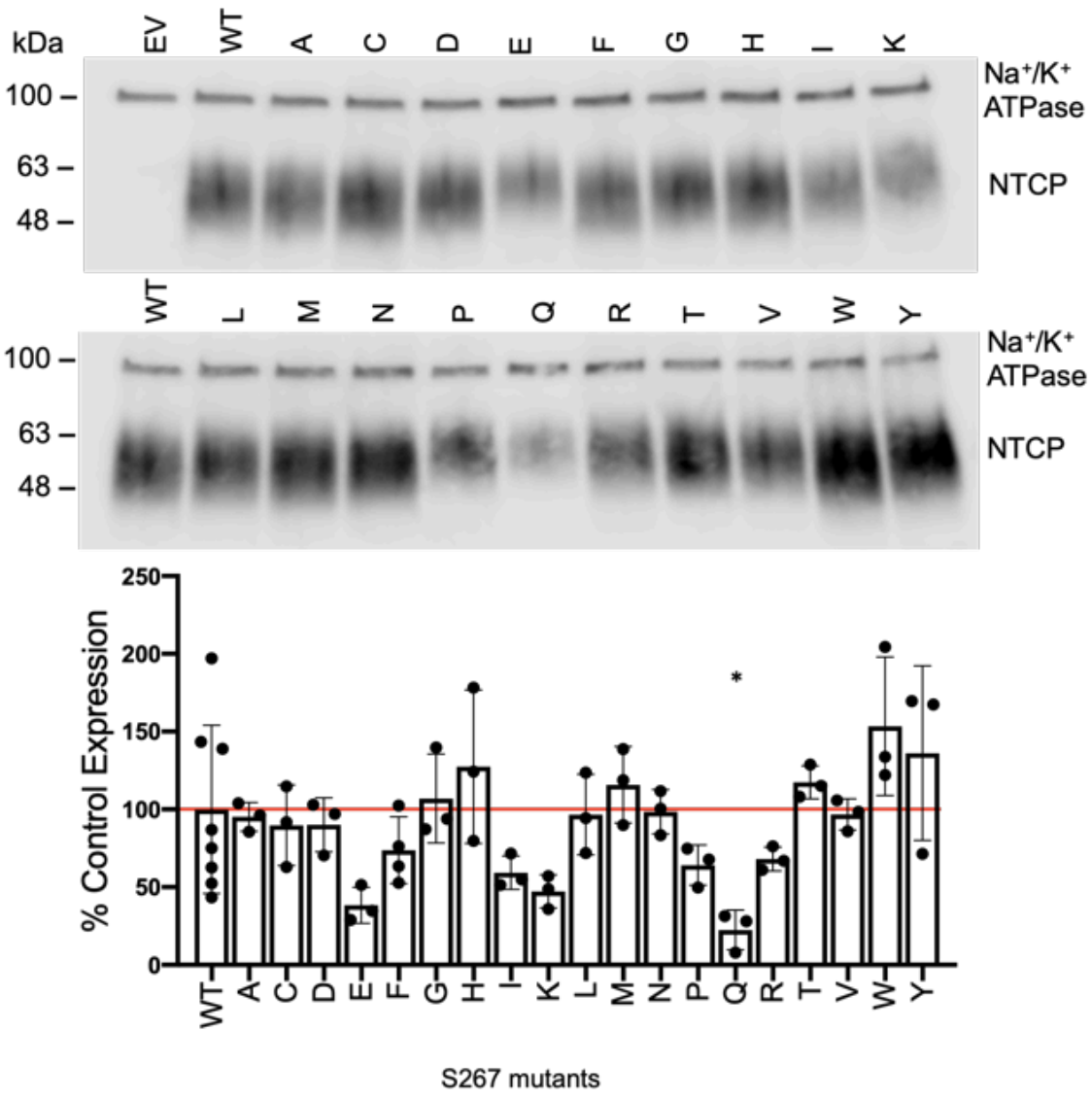


Figure 3-2. Surface expression of WT NTCP and S267 variants.

A, representative Western blot of surface-expressed WT NTCP and its S267 variants in transiently transfected human embryonic kidney 293 cells. EV, WT, and S267 variant proteins were separated on a 4% to 20% gel and then transferred to a nitrocellulose membrane. Blots were probed with a combination of Na^+/K^+-ATPase (loading control at 100 kDa) and tetra-His antibodies (recognizes the His-tagged transporter). *B*, quantification of S267 variants relative to WT NTCP. Expression was quantified using Image Studio Lite, and the bars represent the mean ± SD of three independent experiments; an asterisk indicates a $p < 0.05$ level of significant difference from WT NTCP. The horizontal line indicates WT control, which was set to 100%. EV, empty vector; NTCP, Na^+/taurocholate cotransporting polypeptide; WT, wildtype.

Next, initial uptake experiments from Figure 3-1 were corrected for the surface expression, and results are shown in Figure 3-3. The overall rheostat-like behavior remained, showing that substitutions altered transport function. The rank orders of S267 substitutions were further analyzed using correlation plots to illustrate the effects on substrate specificity (Figure 3-4). Inspection of these plots showed that striking specificity differences were evident for four amino acids (C, F, W, and Y) in the comparison of TCA to estrone-3-sulfate, for five amino acids (C, F, H, W, and Y) for TCA to rosuvastatin, and for one amino acid (C) for estrone-3-sulfate to rosuvastatin. The differential effects of these substitutions – including the clinically relevant polymorphism F – for transporting alternative substrates are a hallmark of altered substrate specificity (Tungtur et al., 2019).

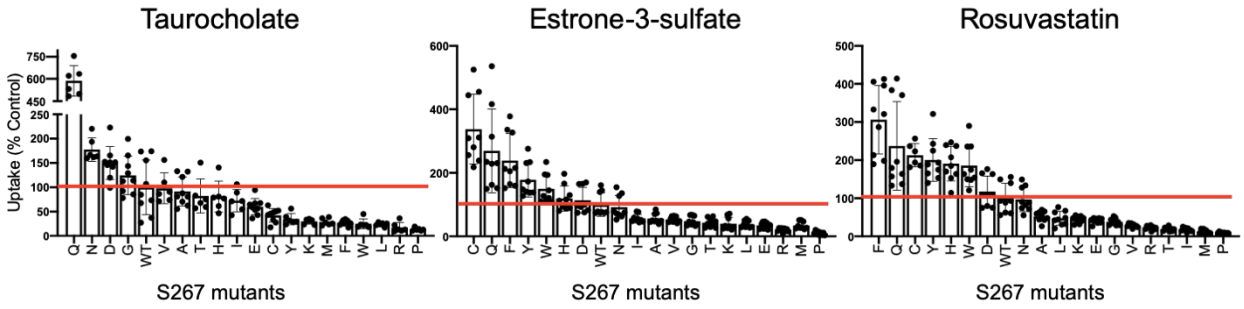


Figure 3-3. Initial substrate uptake normalized for surface expression.

Uptake results (Figure 3-1) were corrected for the surface expression (Figure 3-2*B*) and are presented with the substitutions rank ordered from largest to smallest for each substrate. Horizontal lines indicate wildtype control, which was set to 100%. Error bars represents propagated SD; results of statistical analyses are shown in Supporting Table 3-2.

Of note, correlation with TCA uptake showed similar outliers for estrone-3-sulfate and rosuvastatin while only one outlier was identified when comparing estrone-3-sulfate and rosuvastatin. This suggests that estrone-3-sulfate and rosuvastatin, with the highest Pearson and Spearman values (Figure 3-4*C* and Supporting Table 3-1), share similar modes of interaction with position 267.

When these outlier variants were excluded, the remaining variants showed significant correlation (Figure 3-4, Supplementary Table 3-3), illustrating that the amino acid substitutions had the same direction of change for the alternative substrates. For rosuvastatin and estrone-3-sulfate, the slope of the line was near 1. However, for the other two correlations, the slopes of the lines were less than 1, which is another sign of altered specificity for TCA by these substitutions ((Tungtur et al., 2019) and references therein).

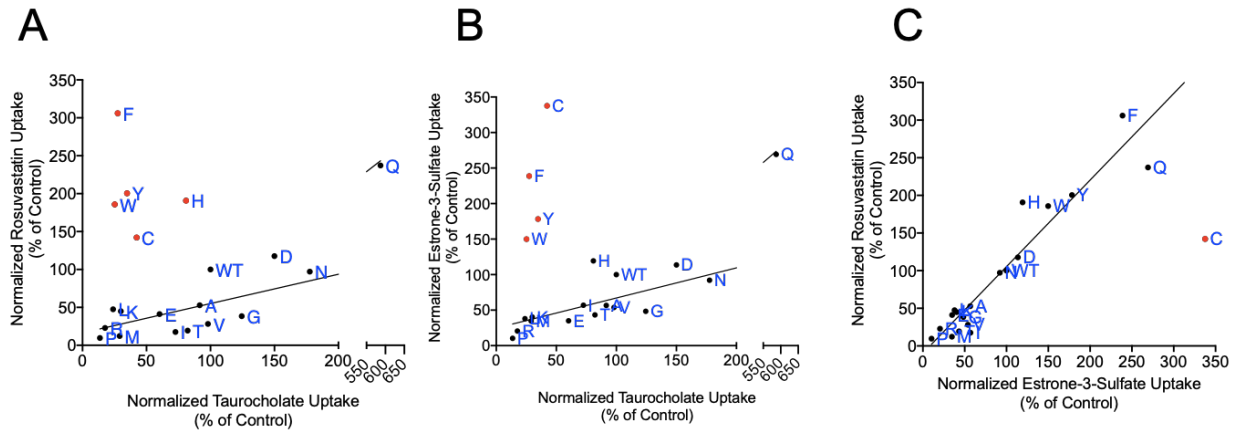


Figure 3-4. Comparison of normalized substrate uptake among the three different substrates.

Normalized uptake values from Figure 3-3 are plotted against each other. Individual points are labeled with letters to indicate their amino acid replacements; wildtype is indicated as WT. *A,* ^{3}[H]taurocholate *versus* ^{3}[H]estrone-3-sulfate, *B,* taurocholate *versus* ^{3}[H]rosuvastatin, and *C,* E3S *versus* ^{3}[H]rosuvastatin. On each panel, several substitutions (*red dots*) were outliers as compared to the others (*black dots*), which is a hallmark of altered substrate specificity. For the nonoutlier substitutions, the Pearson coefficient, which is a measure of the linear correlation, was >0.91 ($p < 0.0001$) for all three substrate comparisons. The Spearman coefficient, which correlates rank order, also showed a strong correlation with values of 0.7 to 0.87 ($p < 0.0001$) (Supporting Table 3-3). For the correlated substitutions, the slopes of the three trendlines were *A,* 0.42 (95% confidence intervals of 0.32-0.52), *B,* 0.39 (0.28 to 0.49), and *C,* 1.15 (0.96-1.34), respectively. Slopes that differ significantly from 1 are another hallmark of altered specificity (Tungtur et al., 2019).

Based on the results presented in Figure 3-3, we performed a full kinetic analysis for select variants: WT, S267F (the NTCP*2 polymorphism), S267N, and S267W. Asparagine at position 267 was chosen because this substitution resulted in similar uptake as WT for estrone-3-sulfate and rosuvastatin but higher uptake for TCA. In contrast, the tryptophan substitution was chosen because it resulted in lower uptake for TCA but higher uptake for estrone-3-sulfate and rosuvastatin. For these four proteins, concentration dependent uptake was assessed under initial linear rate conditions using transiently transfected HEK293 cells. After normalizing for surface expression, we analyzed the results using the Michaelis-Menten equation and calculated K_m and V_{max} values (Figure 3-5, Table 3-1). Both the K_m and V_{max} were frequently altered, indicating altered substrate affinity and transporter turnover.

To summarize these data, we calculated the capacity of each variant to transport the various substrates (V_{max}/K_m) (Table 3-1). For the most part, the capacity of the variants chosen agreed with the rank order shown in Figure 3-3. Of the four variants, WT NTCP had the highest capacity (134 ± 14 μL/mg/min) for TCA (Figure 3-5*A*; Table 3-1) but the lowest capacity for the other two substrates: 15 ± 1.3 μL/mg/min for estrone-3-sulfate and 24 ± 5.1 μL/mg/min for rosuvastatin. The lowest capacity for TCA (11 ± 2.1 μL/mg/min) was determined to be for S267W (Figure 3-5*A*; Table 3-1), in agreement with the single time point single concentration results (Figure 3-3*A*). For estrone-3-sulfate and rosuvastatin (Figure 3-5, *B*-*C* and Table 3-1), S267F showed the highest capacity with 99 ± 11 μL/mg/min and 112 ± 23 μL/mg/min, respectively, which confirms the results presented in Figure 3-3, *B*-*C*. In summary, amino acid substitutions at rheostat position 267 differentially altered both kinetic parameters for substrate transport.

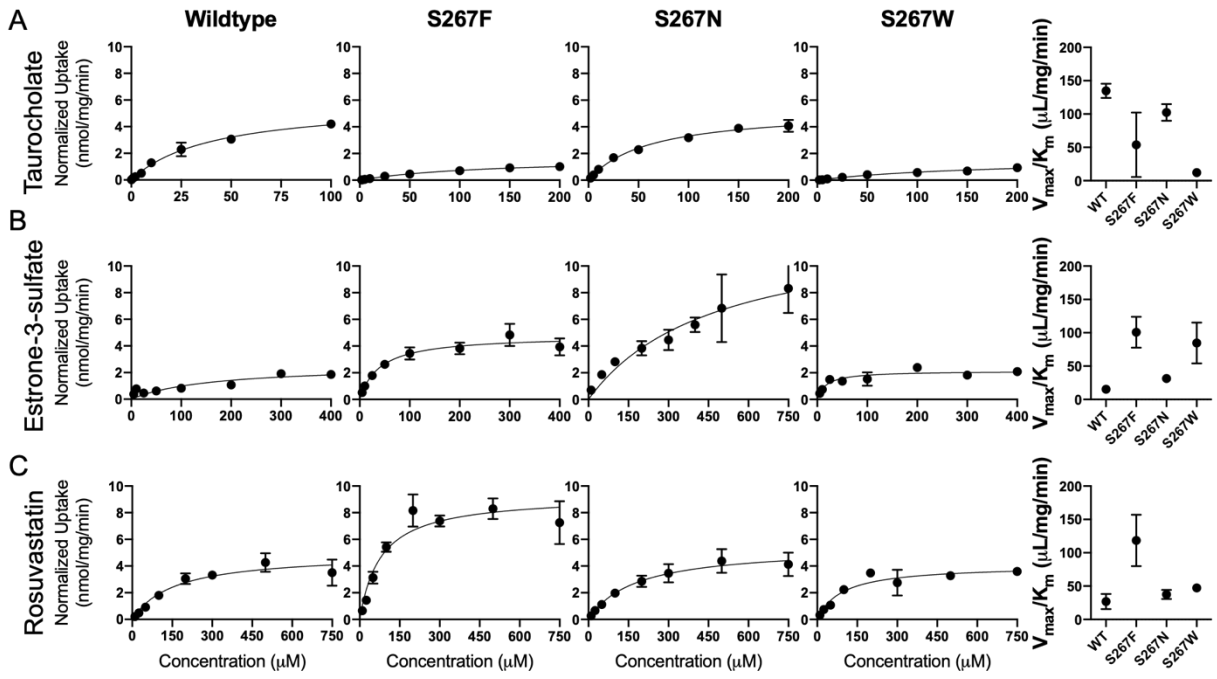


Figure 3-5. Kinetics of substrate transport mediated by WT Na^+/taurocholate (TCA) cotransporting polypeptide and select variants.

Kinetics of *A,* TCA, *B,* estrone-3-sulfate and *C,* rosuvastatin uptake by wildtype Na^+/TCA cotransporting polypeptide (first column), S267F (second column), S267N (third column), S267W (fourth column), and transport capacity (fifth column). Uptake of increasing concentrations of each substrate was measured under initial linear rate conditions in human embryonic kidney 293 cells 48 h after transfection. Results shown are the mean ± SD of a representative experiment completed with triplicate technical samples. The curves are best fits of the mean values using the Michaelis-Menten equation in GraphPad Prism 8. The results listed in Table 3-1 report the average and SD of at least three independent experiments, each comprising two to three technical replicates. WT. wildtype.

Table 3-1. Kinetic values for substrate uptake mediated by WT NTCP and selected variants

Substrate	NTCP	S267F	S267N	S267W
TCA				
K_m (μM)	42 ± 6.6	73 ± 56	59 ± 18	153 ± 50
V_{max} (nmol/mg/min)	5.7 ± 0.5	1.9 ± 0.6	5.9 ± 1.8	1.7 ± 0.1
V_{max}/K_m (μL/mg/min)	134 ± 24	26 ± 22	102 ± 44	11 ± 3.7
Estrone-3-sulfate				
K_m (μM)	155 ± 18	42 ± 5.7	420 ± 173	21 ± 6.7
V_{max} (nmol/mg/min)	2.3 ± 0.2	4.2 ± 0.6	13 ± 4.7	1.7 ± 0.5
V_{max}/K_m (μL/mg/min)	15 ± 2.3	99 ± 19	31 ± 17	81 ± 33
Rosuvastatin				
K_m (μM)	183 ± 51	73 ± 21	171 ± 26	88 ± 12
V_{max} (nmol/mg/min)	4.5 ± 1.0	8.2 ± 1.7	6.3 ± 0.9	4.1 ± 0.4
V_{max}/K_m (μL/mg/min)	25 ± 8.9	112 ± 39	37 ± 7.6	47 ± 7.4

NTCP, Na^+/taurocholate cotransporting polypeptide; TCA, taurocholate; WT, wildtype.

Uptake of increasing concentrations of TCA, estrone-3-sulfate, and rosuvastatin by human embryonic kidney 293 cells transiently transfected with either WT NTCP or variants S267F, S267N, and S267W was measured at 37°C under initial linear rate conditions. Net uptake was calculated by subtracting the transport from identical experiments using sodium-free buffer and was normalized for surface expression. Kinetic parameters, K_m and V_{max}, were determined by fitting the data to the Michaelis-Menten equation using GraphPad Prism 8. Transport capacity was determined by dividing the V_{max} by the K_m. The parameter averages and standard deviation presented were calculated from at least three independent experiments, each comprising three technical replicates.

Homology modeling of human NTCP structure

Both modeling and experimental data suggest that human NTCP has nine TM helices (Hu et al., 2011; Doring et al., 2012) with an extracellular glycosylated amino terminus (Appelman et al., 2017). These helices are arranged into "core" (TM 3, 4, 5, 8, 9 and 10) and "panel" domains (TM 2, 6 and 7) that flank the substrate-binding intracellular crevice (*e.g.,* Figure 3-6). Although no experimental structure is available for human NTCP, structural information could be derived from homology studies. Homologs in the SLC10A family exhibit 9.5% to 99.0% sequence identity and are present in most kingdoms of life (Supporting Information). Bacterial apical sodium-dependent bile acid transporters (ASBTs) are the best-structurally characterized; structures are available for *Yersinia frederiksenii* ASBT ($ASBT_{Yf}$, 26% sequence identity to NTCP) (Zhou et al., 2014) and *Neisseria meningitidis* ($ASBT_{Nm}$, 25% sequence identity to NTCP) (Hu et al., 2011). The pairwise sequence alignments of both relative to NTCP are presented as Supporting Figure 3-1.

In $ASBT_{Yf}$, transport of bile acids and other substrates appears to be accomplished by a transition between inward-open and outward-open conformations. More specifically, this is accomplished *via* a rigid body motion of the core domain (Zhou et al., 2014) that allows alternating exposure of ligand-binding sites to the intracellular or the extracellular space. Both ASBT and NTCP cotransport sodium ions along with bile salts. In the crystal structure of $ASBT_{Nm}$, two Na^+ binding sites have been identified at the junction of core and panel domains (Zhou et al., 2014).

We used the inward-open and outward-open crystal structures of two bacterial ASBTs as templates for comparative modeling of human NTCP. We then applied all-atom structural refinement to the homology models to generate the lowest energy inward-open and outward-

open homology models for WT human NTCP. In agreement with the bacterial structures, a well-defined pocket was present at the junction of the core and panel domains (Figures 3-6, *A*-*B*), in which TCA could bind before being transported. Comparison of the inward- and outward-open models suggest that conformational changes in TM domains 3, 4a, 4b, 7, 9a, and 9b led to closing of that pocket and an opening of a pocket where TCA binds before being released intracellularly (Figure 3-6*C*). Position S267 was located on TM9b, near the substrate-binding cavity in both the inward- and outward-open conformations, and substitutions therefore may directly influence substrate transport.

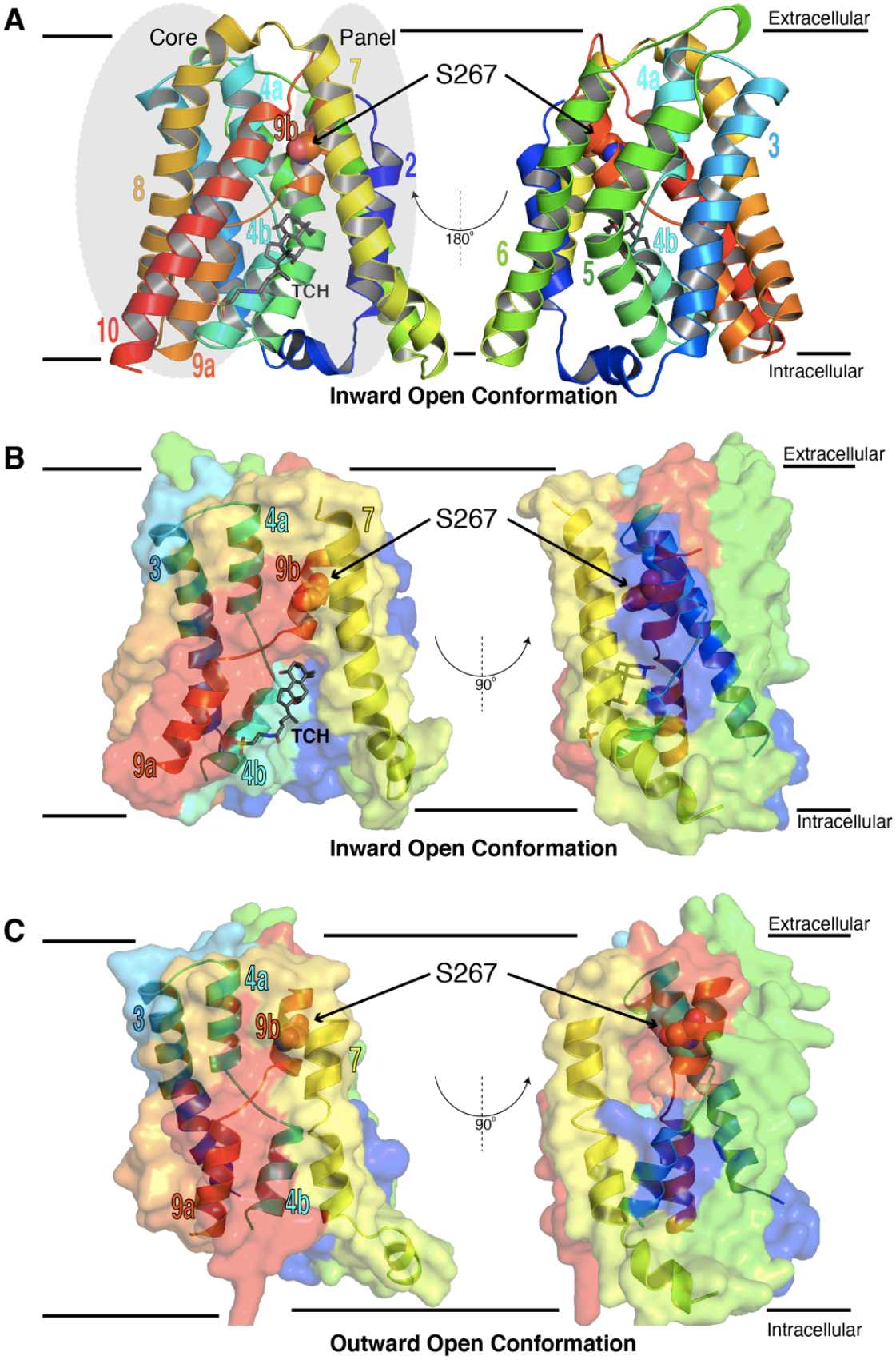

A
Core
Panel
S267
Extracellular
Intracellular
4a
7
9b
2
8
4b
TCH
10
9a
180°
6
5
3
Inward Open Conformation
B
Extracellular
S267
4a
7
3
9b
90°
TCH
9a
4b
Intracellular
Inward Open Conformation
C
Extracellular
S267
4a
3
9b
7
90°
9a
4b
Intracellular
Outward Open Conformation

Figure 3-6. Comparative models of human NTCP.

A, homology model of human NTCP in the inward-open conformation, built using apical sodium-dependent bile acid transporter from *Neisseria meningitidis* as a structural template. The structure comprises nine TM helices (denoted TMs 2, 3, 4a, 4b, 5-8, 9a, 9b, and 10). Position 267 is located on TM 9b. The inward-open conformation has a large crevice at the intracellular side of membrane that is formed between the core and panel domains. Taurocholate (labeled as TCH in this figure), in gray sticks, is included in this perspective by superimposition from the template to show the interaction of a substrate in the inward-open conformation. *B,* the inward-open model is shown in cross section, to highlight the arrangement of helices in this conformation. *C.* homology model of human NTCP in the outward-open conformation, built using apical sodium-dependent bile acid transporter form *Yersinia frederiksenii* as a structural template. Relative to the inward-open conformation, the substrate-binding pocket has closed in this conformation, because of a concerted movement of TM helices 3, 7, 4a, 9b, 4b, and 9a. NTCP, Na^+/taurocholate cotransporting polypeptide; TM, transmembrane. *This figure was completed by collaborators*

Evaluating stability changes arising from substitutions at position S267

Next, we modeled all 19 amino acid substitutions at position 267 of the WT homology models and assessed the predicted stability changes on the inward- and outward-open conformations. Starting with the inward-open conformation, we found that most substitutions were predicted to be stabilizing relative to the WT control (serine at position 267): 13 of the 19 potential substitutions yield energies more favorable (more negative) than serine (Figure 3-7*A*). This observation was striking because it is in stark contrast with typical results from such computation predictions, which have long found that the WT amino acid tends to score more favorably than any other substitution (Kuhlman and Baker, 2000).

To explore the structural basis for the observed energetics in this modeling experiment, we selected two substitutions that were stabilizing (S267I and S267W), one with little effect on stability (S267N), and one predicted to be slightly destabilizing (S267G). Inspection of the representative conformations for each of these variants revealed slight changes in the local environment; in particular, TM helices 2 and 7 – which face the side chain presented at position 267 – respond in a slightly different manner to each variant (Figure 3-7*B*).

In our model of the outward-open conformation, the WT S267 was more solvent exposed than in the inward-open conformation, with a solvent-accessible surface area of 17.76 $Å^2$ as compared to 4.26 $Å^2$. Remarkably though, the same unusual behavior emerged when probing stability differences in the outward-open conformation. Many substitutions were predicted to be more stable than the WT serine (Figure 3-7*C*), and again the notable tolerance for alternate amino acids at position 267 could be rationalized by malleability of the local structure, this time the nearby TM helix 10 (Figure 3-7*D*). In both cases, the small rearrangement of these helices

understates the dramatic differences in side chain conformations needed to accommodate these alternatively packed arrangements (Supporting Figure 3-2).

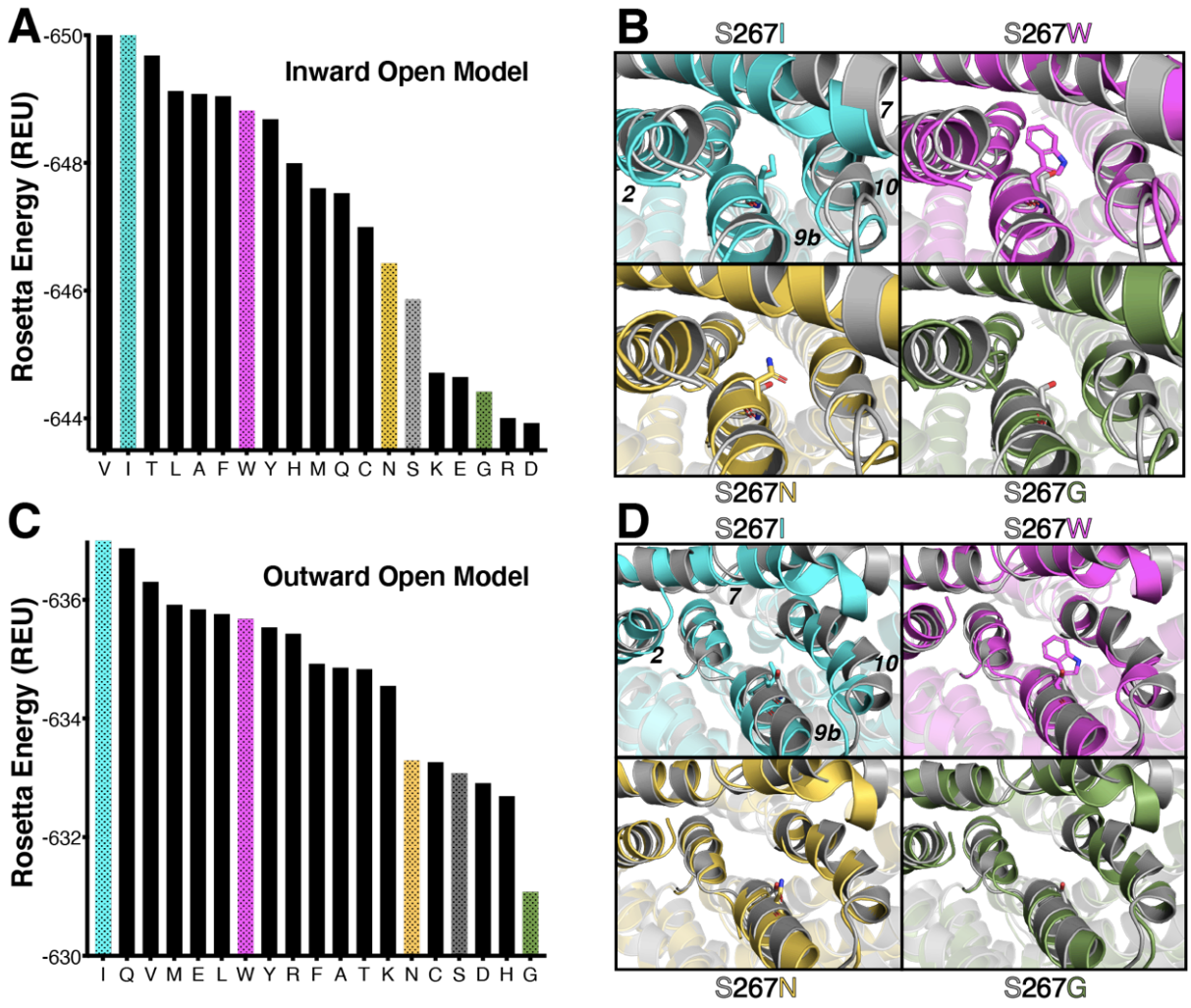


Figure 3-7. Predicted stability differences associated with sequence variation at the S267 position.

A, Rosetta energies for 19 sequence variants at the 267 position, using the inward-open model. Proline is not shown, because the energy associated with this residue is very unfavorable (off the scale); this is consistent with experimental results, in which proline showed the least amount of transport and somewhat diminished surface expression (Figures 3-1 through 3). *B,* structural details from the models underlying these energy differences. Four different sequence variants are compared with the wildtype S267; in each case, the conformation of TM helices 2 and 7 respond to changes in the amino acid at position 267, which is located on TM helix 9b. *C,* Rosetta energies using the outward-open model. Proline is again not shown, because the energy associated with this residue is very unfavorable (off the scale). *D,* structural details from the outward-open models. In this conformation, the position of TM helix 10 responds to changes in the amino acid at position 267. REU, Rosetta energy unit; TM, transmembrane. *This figure was completed by collaborators*

Correlation of structure models and experimental data

We next compared the computational stabilities of each S267 variant to the experimental data. The most direct comparison should be to the NTCP surface expression, which would be decreased or increased by altered protein stability. To that end, we examined the effect between cellular surface expression levels and the calculated energies using the inward-open NTCP model (Supporting Figure 3-3*A*), the outward-open model (Supporting Figure 3-3*B*), and the difference between the energies of the inward-open and the outward-open models. We anticipated that if the protein resides primarily in one conformation or the other, its surface expression may correlate with the calculated energies for that conformation; however, we found no statistically significant correlation between surface expression and the energies of the models in either conformation (Supporting Table 3-3).

Instead, we observed a correlation between surface expression and the difference in energy between the two conformations (Figure 3-8*A*) (Pearson and Spearman coefficients were -0.64 and -0.52, respectively; both correlation coefficients were nonzero with $p < 0.05$; Supporting Table 3-3). To rule out any possibility that the dramatic relationship observed between surface expression and the difference in energy between the two conformations was due to some quirk of the Rosetta energy function, we carried out the same calculation using the FoldX package (Schymkowitz et al., 2005). Unsurprisingly, the calculated energies were correlated with those from Rosetta (Supporting Figure 3-4, $p < 0.03$). We again observed a correlation to the observed surface expression, albeit it not to a statistically significant degree (Figure 3-8*B*; $p < 0.10$). We attribute the slightly stronger correlation from Rosetta's calculated energies to the fact that this protocol sought to capture slight backbone rearrangements in response to each mutation, whereas FoldX energies did not include backbone flexibility.

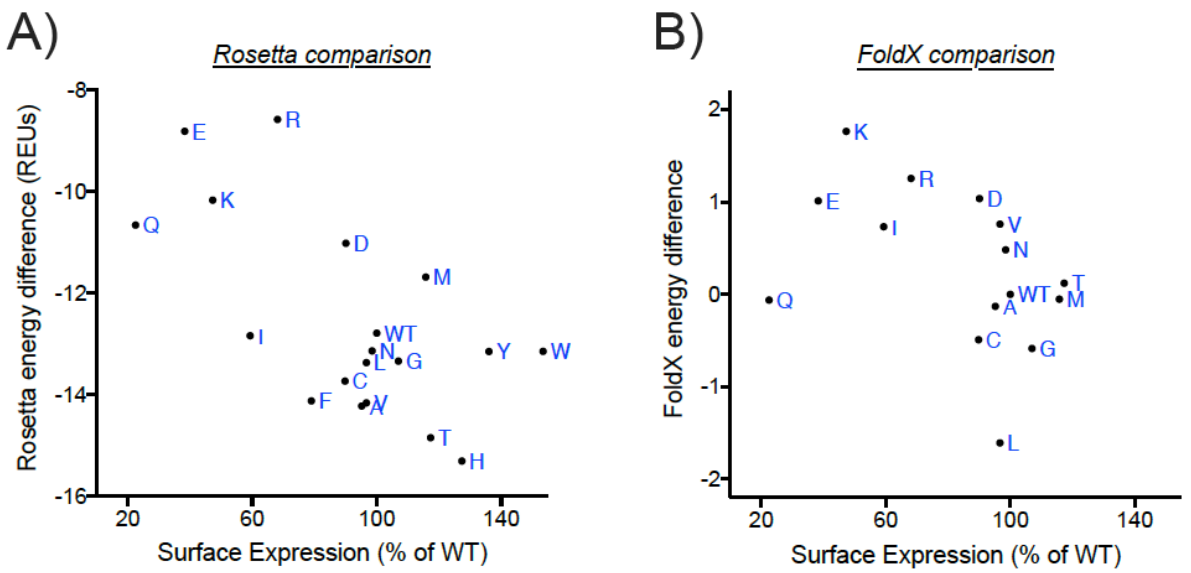


Figure 3-8. Correlation of surface expression levels with calculated stability differences for inward-open model minus outward-open model.

The percent surface expression (Figure 3-3) of each variant is plotted against results from structure-based modeling of each variant. Individual points are labeled with letters to indicate their amino acid replacements; wildtype is indicated as WT. Proline is excluded because the stability difference for this helix-breaking residue cannot be reliably estimated using these methods. *A,* energies calculated using Rosetta. *B,* energies calculated using FoldX. Aromatic residues (F/H/W/Y) were excluded because FoldX does not model backbone flexibility, making it unable to accommodate these large side chains. Note that these aromatic residues were the variants that dramatically altered substrate specificity (outliers in Figure 3-4). REUs, Rosetta energy units. *This figure was completed by collaborators*

There are several potential explanations for why sequence variants with more favorable energies in the inward-open model relative to the outward-open model (*i.e.,* the difference between the states) yielded higher surface expression. A direct physical interpretation may be that the inward-open conformation promotes surface expression, whereas a stable outward-open conformation serves as a trap that disfavors membrane insertion and/or correct folding required for presentation on the cell surface. In such a model, the outward-open conformation is critical for transport, but over-stabilization of this conformation would prove deleterious. As an alternate explanation, however, we note that the Rosetta energy function is parameterized primarily for study of soluble proteins. As a result, energy differences between sequence variants are computed relative to a reference state that may not be appropriate for membrane proteins (the unfolded protein in a soluble context). Because of this, the outward-open conformation (in which position 267 is more solvent exposed) may simply be serving as an alternate (more appropriate) reference state, implying that calculated stability of the inward-open conformation is the primary determinant of surface expression, once energies have been more appropriately normalized.

For completeness, we additionally explored whether there was a relationship between the Rosetta-calculated energies and substrate uptake; we did not expect the Rosetta energies to be predictive, given that the balance of multiple conformational states is likely critical for effective transport. This expectation proved to be the case, as we did not observe a statistically significant correlation between any of the energies with uptake of any substrate (Supporting Figure 3-5).

Overall, the strong correlation between surface expression and these computed energy differences supports an underlying structural rationale that diverse amino acids can be accommodated at position 267 in a manner that altered – but did not disrupt – protein activity in a rheostatic manner. Because stability did not appear to be greatly altered, we hypothesize that

the experimentally observed changes in transport for each substitution arose from altered dynamics or from altered energies of other conformations that surely exist in the NTCP ensemble.

III. **Discussion**

Rheostat positions were first described in soluble-globular proteins (Meinhardt et al., 2013; Hodges et al., 2018) and have been predicted in a wide variety of proteins (Miller et al., 2019). However, to our knowledge, no such positions have been biochemically/experimentally identified in TM proteins. In the present study, we confirmed that rheostat positions can occur in TM transport proteins such as NTCP. Indeed, the natural polymorphic position, S267, behaved as a rheostat position for all three substrates tested: TCA, estrone-3-sulfate, and rosuvastatin (Figure 3-1).

We were further able to parse contributions to NTCP cellular uptake into various biochemical processes, including transport, ligand specificity, membrane localization, and protein stability. Our biochemical approach provided an advantage over the deep mutational scanning studies (Roscoe et al., 2013; Fowler and Fields, 2014) that have become increasingly popular in recent years. Although the latter experiments substitute large protein regions with all amino acids at each position, they rely upon phenotypic competitions that are highly sensitive to environmental conditions/biological thresholds and are the composite of many functional parameters. Furthermore, data interpretation is based on the assumption (and/or requires extensive validation) that allele frequency within a library represents the degree of protein function. Biochemical assays, such as those reported here for NTCP position 267, directly report high-resolution

information about the protein's activity that can be resolved into multiple functional and structural components.

Here, the overall phenotype of each substitution at position 267 was primarily dominated by changes in altered transport kinetics. Full kinetic analyses with selected variants demonstrated that both K_m and V_{max} were differentially affected, along with substrate specificity. Nonetheless, both modeling results and experimental data indicated subtle changes in stability that were distinct from (not correlated with) changes in transport. Thus, changes at one rheostat position altered multiple functional and structural parameters, revealing a complex interplay that must be resolved to advance predictive pharmacogenomics. Because we also observed complex outcomes – affecting multiple functional parameters – arising from substitutions at rheostat positions in human liver pyruvate kinase, this complexity is likely common in a variety of proteins (Wu et al., 2019).

Of the affected functional parameters in NTCP, we were particularly intrigued by the altered substrate specificity that was striking for a subset of the amino acid substitutions at position 267 (changed rank order in Figure 3-3 and correlation outliers in Figure 3-4). Historically, altered substrate specificity has been defined from the perspective of the proteins, as changes in either the rank order of preferred substrate and/or changes in the fold change of transport (Creighton, 1993; Tungtur et al., 2019). The results shown here – from the perspective of the substrate – provide an orthogonal view of specificity.

The substrate-dependent effects became even more apparent when we compared the kinetics for selected variants (Figure 3-5, Table 3-1). The substitution-dependent effects on substrate specificity could arise if the translocation pathway or binding pocket for TCA differed from that

of the other two substrates. Distinct binding pockets within the translocation pathway were recently demonstrated for three substrates of the organic cation transporter 1 (Boxberger et al., 2018). Different binding pockets or translocation pathways may be a common feature of multi-specific drug transporters. Alternatively, substrate-dependent substitution outcomes might arise if the different substrate/substitution combinations had different effects on the NTCP conformational and equilibrium dynamics, similar to the types of changes that arise from amino acid substitutions in β-lactamase (Modi and Ozkan, 2018).

Indeed, substrate transport by solute carriers like NTCP is a dynamic process generally described by the alternating-access model. As such, NTCP is an intrinsically flexible protein that undergoes complex and hierarchical conformational changes while carrying out its biological function of transporting substrates into hepatocytes. In addition to the outward- and inward-open conformations considered here, NTCP must have multiple intermediate states. The stabilities and/or equilibrium dynamics for each of the conformations could be differentially affected by single amino acid substitutions, giving rise to the intermediate functional outcomes observed. Conformational changes are also modulated by interactions with substrates and sodium ions. Thus, an accurate evaluation of structural characteristics of intermediate conformations along the entire conformational transition pathway will be necessary to understand rheostatic substitution behavior. This remains a challenging task for both experimental and theoretical approaches.

When the rheostatic outcomes from the cellular uptake assays were further investigated, strong rheostatic effects were observed for transport, and the effects on stability of the inward-open conformation (in which position 267 was mostly buried) were slightly rheostatic. Furthermore, computational modeling and stability calculations were in agreement with the experimental measures of protein surface expression. This indicates that modeling merits further exploration

for identifying positions in functionally important regions that tolerate a wide range of substitutions, the hallmark characteristic of a rheostat position.

In conclusion, these combined results showed that NTCP position 267 is a rheostat position. Although individual substitutions had wide-ranging effects on the various aspects of function and stability that give rise to the overall phenotype of cellular transport, much of the change could be attributed to altered transport kinetics. Structurally, position 267 was strikingly tolerant to substitution, despite the fact that it was largely buried in the inward-open conformation. Based on these results, we propose that other polymorphic positions in NTCP and in other proteins might also be locations of rheostat positions (Fenton et al., 2020). Given the complex interplay between substitutions and substrate specificity observed at this NTCP rheostat position, it is imperative to expand recognition and understanding of rheostat positions to advance predictive pharmacogenomics.

Chapter 4

Examination and characterization of predicted rheostat, toggle and neutral positions within the Na^+/taurocholate cotransporting polypeptide (NTCP)

Contributions from John Karanicolas, Shipra Malhotra, Liskin Swint-Kruse, and Bruno Hagenbuch in this chapter include: text and figures. Detailed contributions were as follows: computational modeling of NTCP, structure-based Rosetta energy score calculations, and the corresponding figure (Figure 4-1. Stability scores predicted using the Rosetta Software Suite as a consequence of amino acid replacement at position 271.) were completed by Shipra Malhotra and John Karanicolas. Rosetta energy scores completed by these collaborators were then used for subsequent correlation studies. Figures completed by collaborators are denoted in the figure legends.

I. Introduction

Historically, the prediction of functional outcomes has been limited to specific protein mutations and only included two types of outcomes: 1) depleted function: deleterious or catastrophic mutations or 2) no change in function compared to wildtype: neutral mutations. If multiple amino acid substitutions result in similar outcomes then the amino acid residue is considered a toggle or neutral position, respectively. In addition, our group recently discovered a third potential outcome: intermediate mutations. These amino acid replacements result in an alteration that is neither diminished function nor similar to wildtype but rather somewhere in between (Meinhardt et al., 2013). Further, if numerous substitutions result in a continuum of functional outcomes including function that is more efficient or similar to wildtype, depleted function, and intermediate function then the position is considered a rheostat position.

These rheostat positions often do not conform to the existing algorithm parameters. Most importantly, rheostat positions often occur at nonconserved positions whereas algorithms were

biased towards the prediction of mutations at conserved amino acid locations. Mutations at rheostat positions also do not result in only detrimental or neutral outcomes. And finally, there does not seem to be a correlation to the biochemistry of amino acid substitutions and the functional outcomes, and the same amino acid substitution behaved differently in different homologs (Meinhardt et al., 2013; Fenton et al., 2020). The combination of a rheostat's unique functional outcome and location makes these positions unpredictable by existing algorithms (Miller et al., 2017).

While intermediate functional outcomes may not be deleterious under "normal" conditions, the effects on structure or function could still impact a patient's therapeutic response to a drug. In addition, it is likely that these intermediate outcomes are a result of substitutions at a rheostat position. Rheostat locations are found in numerous types of proteins, indicating that they are not only relevant but also a prevalent mutation (Meinhardt et al., 2013; Ohnishi et al., 2014; Adamski and Palzkill, 2017; Hodges et al., 2018; Wu et al., 2019; Procko, 2020). These factors further emphasize the importance of being able characterize, understand and predict these positions. If predictive computer algorithms were improved to include rheostat positions, this would have a significant positive impact on predictive pharmacogenomics.

Following the confirmation of rheostats within NTCP in Chapter 3, we next wanted to determine if it was possible to predict alterations in protein expression as well as mutational tolerance based on the calculated energy scores which could indicate potentially rheostatic functional outcomes. We hypothesized that we could use the calculated energy scores to predict which amino acid positions could tolerate a wide range of substitutions without greatly impacting protein stability. Thus, we selected position 271 for this study for the following reasons: First, N271 is close to S267 but located away from the substrate binding site and should not impact the translocation of

substrates. Therefore, we anticipated that when comparing transport by the variants, there would be strong substrate-to-substrate correlations. Second, the calculated energy scores predicted that several variants would be energetically similar to wildtype, indicating that amino acid replacements at this position would be tolerated and the protein would remain stable. Finally, N271 has a high conservation score of 2.2 (in an overall range of 0.0-2.8) compared to S267 which had a score of 1.54 (Chapter 3), indicating that position 271 is less evolutionarily conserved. In addition, N271 has a ConSurf (Ashkenazy et al., 2016) score of 6 as compared to S267 whose score was 8. We hypothesize that this decreased conservation of position 271 indicates that it may be an amino acid location that allowed for functional variations to evolve between homologs.

In addition, NTCP's ability to adapt to mutations thus far led us to question if we could detect non-rheostatic positions within NTCP. To investigate this question, two additional positions were selected to determine if mutations at all positions in NTCP would result in rheostat-like functional outcomes. We utilized the multiple sequence alignment and entropy calculations from Chapter 3 as well as the ConSurf analysis and selected glycine 102 and tyrosine 146 to target for studies.

Glycine 102 is evolutionarily highly conserved with a calculated sequence entropy of 0.0054. In addition, position 102's ConSurf score of 9 was in agreement with the sequence entropy. Therefore, glycine at position 102 would historically be considered a structurally and functionally important amino acid. Thus, we hypothesized that mutations at this location would be detrimental and the result would functionally be similar to a toggle (Landau et al., 2005). In stark contrast, tyrosine 146 is much less conserved with an entropy of 1.91. In addition, homology modeling predicted position 146 to be located in an extracellular loop, which is often

expected to be neutral (Martin et al., 2020). These characteristics suggest that mutations at Y146 would be well tolerated with little to no impact on the function of NTCP and therefore resemble a neutral position.

We first experimentally characterized the initial substrate transport and expression of all 19 NTCP variants with substitutions at position 271. We also determined if transport kinetics of select variants would be altered. To test the hypothesis that the homology model simulations and energetic scoring could accurately predict the function or, more importantly, the expression of mutations, we compared the experimental results to simulation energy scores for the inward-open, outward-open, and the difference of the two models. In addition, we assessed the function and expression of NTCP containing substitutions at positions G102 and Y146. Finally, we used the RheoScale calculator to systematically assess the substitution outcomes observed for all characterized positions, including the previously published 267 position. These scores allowed us to classify each experimental result as rheostat, toggle, neutral, or somewhere in between (Hodges et al., 2018).

II. **Results**

Structure model simulated energy scores

Several locations within the NTCP homology model were computationally mutated to all other nineteen amino acids and simulations with the variant proteins were completed using the Rosetta macromolecular modeling suite. These simulations resulted in a variety of modest predicted energy changes. In particular, as shown in Figure 4-1, the simulation energy scores predicted that amino acid substitutions at position 271 would be tolerated in both the inward-open facing model

(left side) and the outward-open facing model (middle), despite 271's buried location. Mutations that occur in buried location are normally expected to be highly destabilizing, however this was not the case for position 271. This is, the energy was altered but not substantially increased to indicate protein instability. In addition, the energy scores for the difference of the inward and outward-open facing model (right-side) does not seem to differ dramatically from wildtype, indicating that the individual variant energy scores differ depending on the protein's conformation. For example, N271H shows increased energy in the inward-open model but decreased energy in the outward-open model when compared to wildtype. These predictions led to the selection of position 271 to be examined for potential mutational tolerance and thus rheostat-like functional behavior.

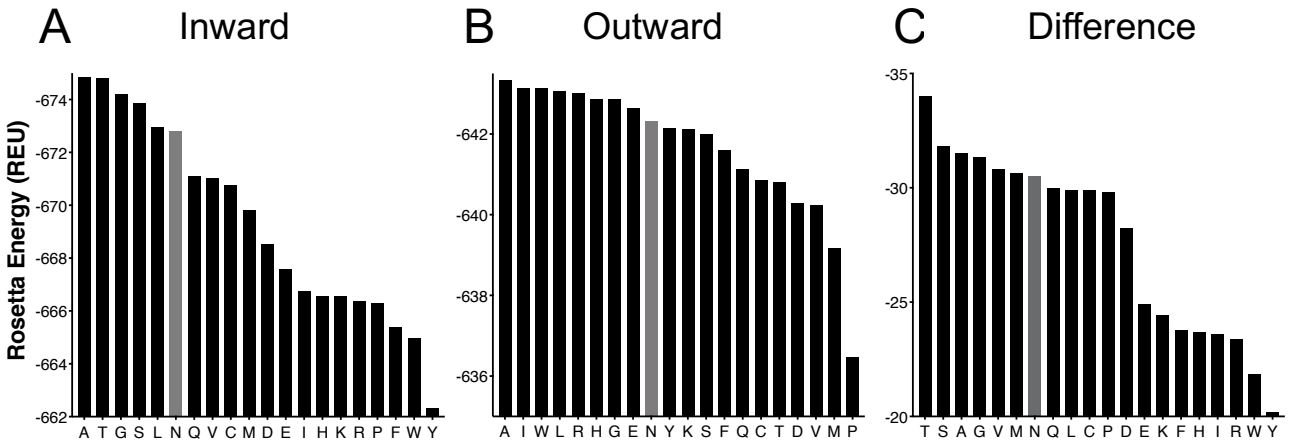


Figure 4-1. Stability scores predicted using the Rosetta Software Suite as a consequence of amino acid replacement at position N271.

A. Rosetta energy scores for 19 variants using the inward-open model. *B.* Energy scores calculated using the outward-open model. *C.* Energy scores calculated by taking the inward minus the outward-open model scores. Native amino acid, asparagine, is designated by the lighter gray color. *This figure was completed in part by collaborators*

Substrate transport by N271 variants

We first wanted to determine the effects that amino acid substitutions at position 271 have on the function of NTCP. Thus, we began by mutating position 271 to all other nineteen amino acids. Confirmed variants were transfected into HEK293 cells and substrate uptake was determined (Figure 4-2). We used three model substrates to determine if variants demonstrated substrate dependent alterations in function: taurocholate (top row), estrone-3-sulfate (middle row) and rosuvastatin (bottom row). The left plot displays the variants in alphabetical order with the exception of wildtype which is listed first. In contrast, the right plot shows the changes ordered by transport rate. Overall, the ordered side of Figure 4-2 shows that position 271 behaves like a rheostat for all three substrates. The variants present a continuum of functional outcomes from approximately 135% of wildtype transporter function down to 2%, with the wildtype at (100%). Subjectively, changes in taurocholate uptake are the least rheostatic while estrone-3-sulfate and rosuvastatin show a wider range in transport (ordered, right side). However, there are similarities between the three substrates. For instance, proline, glutamic acid, and aspartic acid are the three amino acid variants with the lowest transport. In contrast, variants like alanine and tyrosine seem to show elevated transport for all substrates. Furthermore, there are no drastic changes in rank order for the variants in the right side of Figure 4-2 from substrate to substrate, indicating that changes in transport are not substrate dependent. This indicates that amino acid substitutions at position 271 neither disrupt the substrate binding site nor the translocation pathway for any of the three substrates examined. This is in contrast to the clear substrate dependent changes that were seen in Chapter 3 with S267 variants.

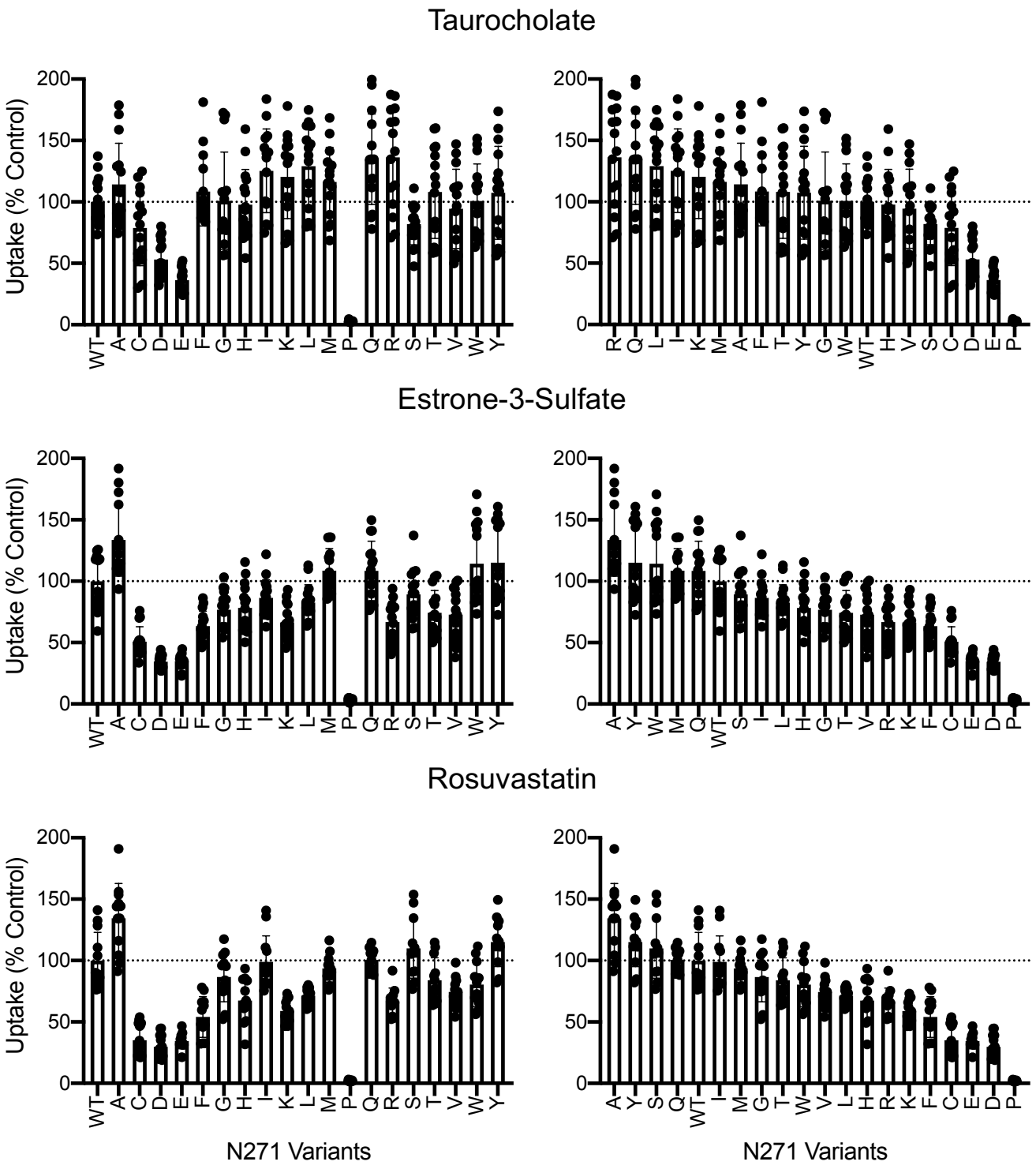


Figure 4-2. Transport of select substrates by wildtype NTCP and its' N271 variants.

HEK293 cells were transiently transfected with either wildtype NTCP (WT), N271 variants, or empty vector. Transport of ^{3}[H]taurocholate (30 nM), ^{3}[H]estrone-3-sulfate (5.8 nM) and ^{3}[H]rosuvastatin (50 nM) was measured at 37°C for 5 minutes, 48 hours post transfection. Uptake by the empty vector transfected cells was considered background and subtracted from the uptake by WT and N271 variant-expressing cells to yield net

uptake. Results for variants were normalized relative to wildtype (100%). The left-hand plots show transport by the variants alphabetically by their amino acid substitutions, apart from wildtype which is listed first. The right-hand plots show the results ordered from highest to lowest transport activity. Individual data points from n=3 biological replicates with at least 2-3 technical replicates are reported with the bar indicating the mean of all replicates ± SD. Horizontal line indicated wildtype at 100% to aid in visual comparisons.

Examining the reasons for variation in substrate transport by N271 variants

To elucidate if the changes seen in Figure 4-2 were due to alterations in transport or variant protein expression at the plasma membrane, we performed surface biotinylation experiments. The results shown in Figure 4-3 revealed over a three-fold difference between the lowest to the highest expressed variants (48% to 160% compared to wildtype set at 100%) (Figure 4-3B). For example: serine, glutamine, and alanine were all overexpressed compared to wildtype while expression of other variants like threonine and isoleucine was significantly decreased. Many additional variants were expressed at similar levels to wildtype, such as: proline, tyrosine, and cysteine. These variations, or lack thereof, in surface expression could be attributed to a number of factors including: altered post-translational modifications, modified protein stability, and/or mutational tolerance (such as protein repacking).

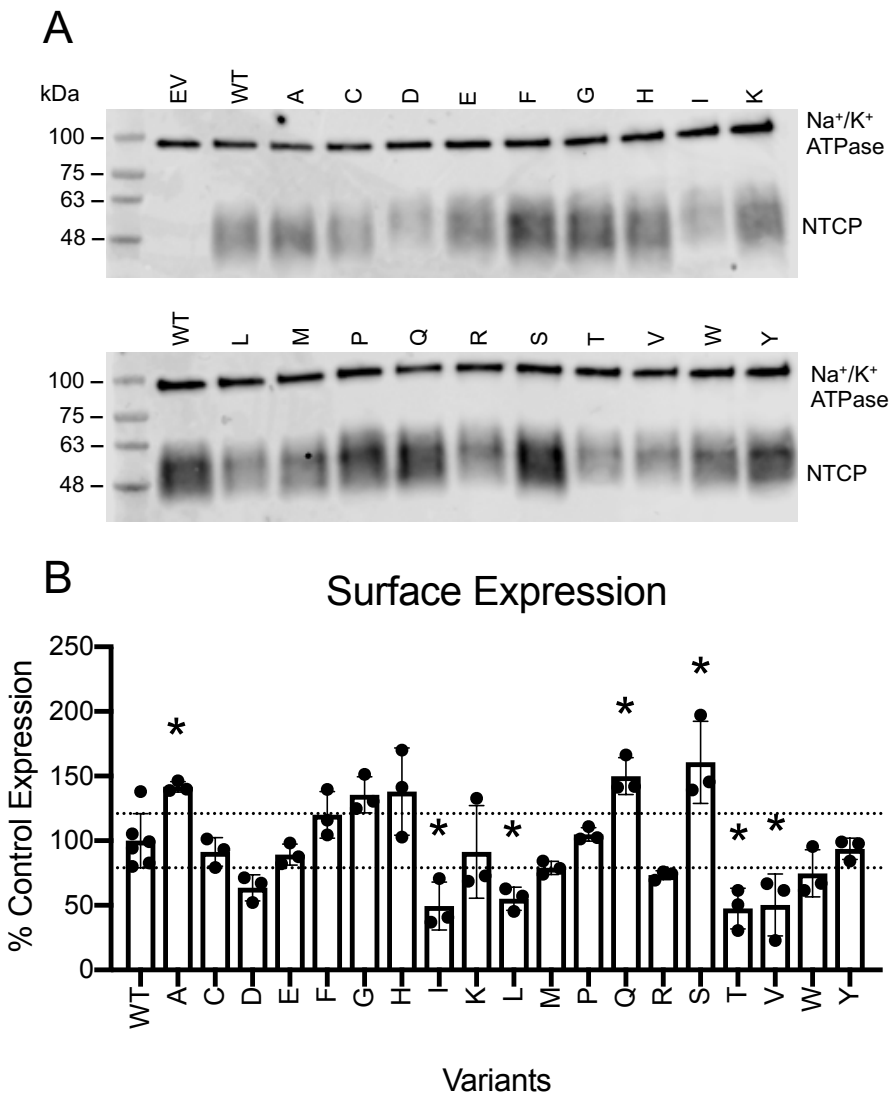


Figure 4-3. Surface expression and quantification of wildtype NTCP and N271 variants.

A. Surface expression of wildtype NTCP and N271 variants on a representative western blot. Proteins from HEK293 cells transiently transfected with empty vector (EV), wildtype NTCP (WT) and N271 variants were separated using a 4-20% polyacrylamide gradient gel and were then transferred to nitrocellulose membranes. Blots were probed simultaneously with Na^+/K^+-ATPase as a loading control (100kDa) and Tetra-His antibodies which detects His-tagged proteins. *B*. Quantification of western blots with N271 variants compared to wildtype NTCP. Three independent surface expression experiments were quantified using Image Studio Lite. Individual data points are shown with the bar representing the mean ± SD. Horizontal lines show the upper and lower limit of the wildtype average plus or minus its standard deviation, respectively. Asterisks denote significant difference from wildtype NTCP with a $p<0.05$ level.

Results of the initial uptake experiments (Figure 4-1) were then corrected for surface expression. Figure 4-4 demonstrates that, subjectively, this enhanced the rheostat-like behavior position 271 for all three substrates, most noticeably for taurocholate. In addition, many of the variants changed location within the hierarchal ordering for each substrate. For example, in Figure 4-2 alanine showed elevated transport. However, after correction for protein overexpression, alanine transport was more similar to wildtype than initially thought. In contrast, isoleucine showed transport similar to wildtype but after surface correction isoleucine is the best transporting variant for all three substrates.

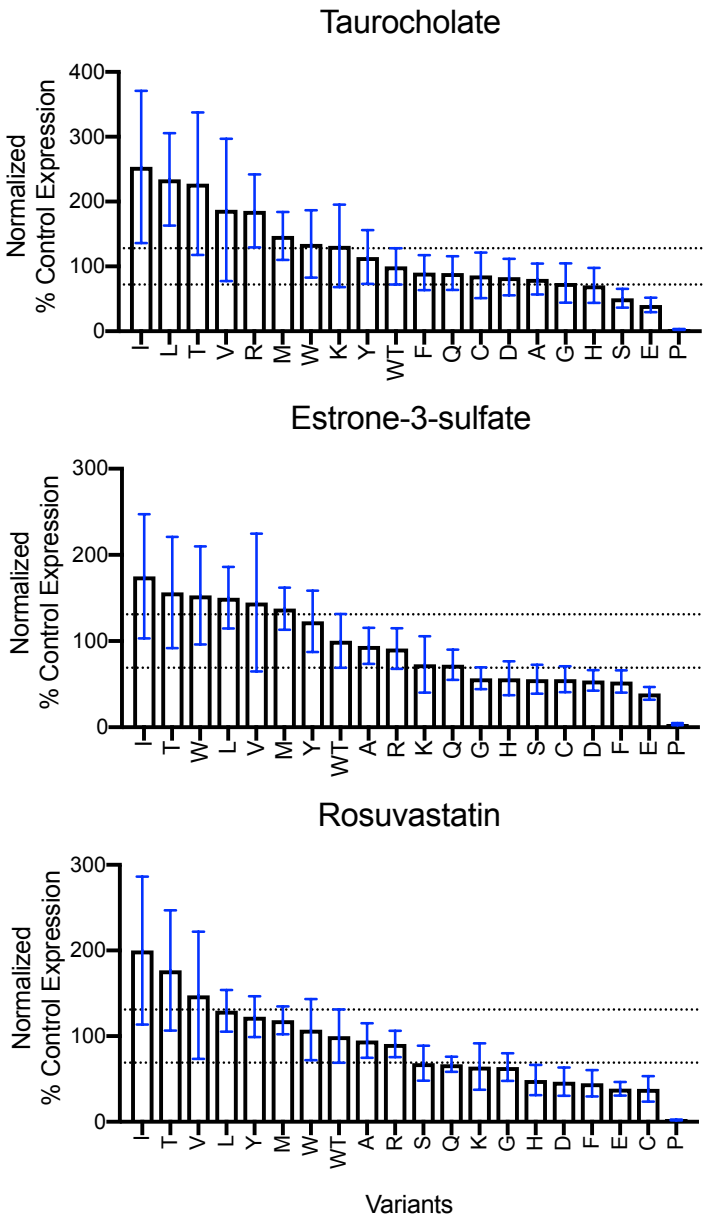


Figure 4-4. Substrate transport by N271 variants corrected for surface expression.

Initial uptake results from Figure 4-2 were normalized for the surface quantification in Figure 4-3. Corrected transport is shown with variants and wildtype NTCP ordered from greatest to least uptake for each substrate. Bar graphs indicate the average of the corrected values ± propagated SD. Horizontal lines denote the upper and lower limits of the wildtype standard deviation. Two to three technical replicates from at least three independent experiments are shown.

Comparisons of experimental results with structure modeling and energy calculations

To determine if homology modeling along with Rosetta simulations could accurately predict mutational tolerance and therefore the functional outcomes of variants at rheostat locations, we compared the calculated energy scores with both, the expression corrected initial transport rates (Figure 4-5) and surface expression (Figure 4-6). Both initial transport and surface expression were compared to the inward- and outward-open model energy scores (left side and middle, respectively) as well as to the difference between the two model scores (right side). Pearson and Spearman correlation (Supporting Table 4-1) calculations showed no significant correlation between any of the Rosetta energy score sets and initial transport or the surface expression.

The lack of correlation for position 271 was in contrast to the correlations shown for position 267. In Chapter 3, there was a modest correlation between the inward- minus outward-open energy scores and the surface expression for position 267. In addition, it was hypothesized that the inward-open model, and therefore the energy scores, were indicative of the surface expression, while the outward-open scores were an indicator of the transport. In agreement with this hypothesis, the strongest correlation (lowest p-value) shown for position 271 was between the transport of estrone-3-sulfate and the outward-open energy scores. In addition, the strongest correlation between the surface expression and simulated energy scores was shown for the inward-open model for position 271. However, neither correlation reached significance. In summary, while there are trends indicating our previous hypotheses are correct, the absence of correlation between the simulated energy scores and the surface expression for position 271 indicates that our interpretation of the structure modeling needs to be reevaluated.

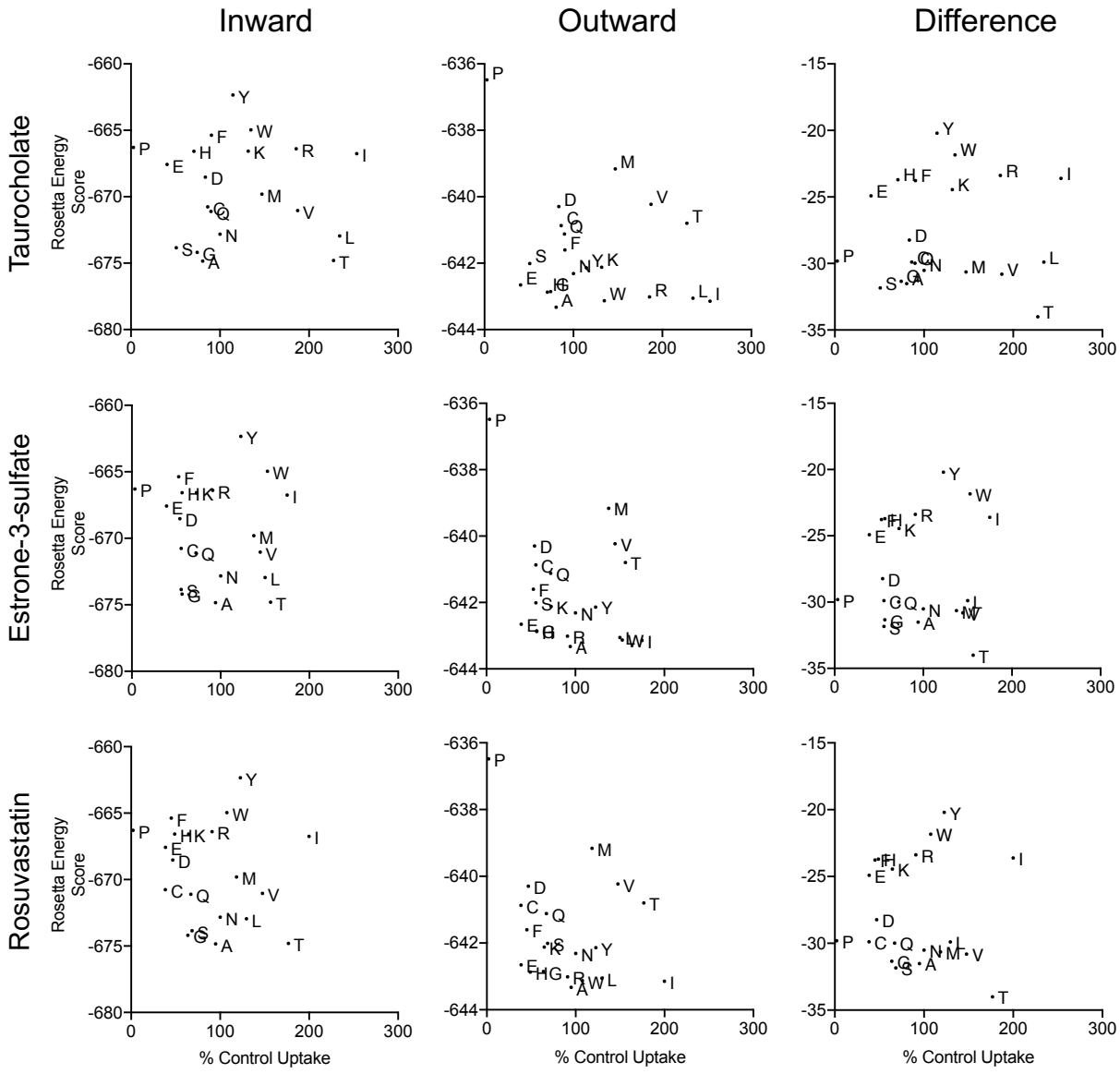


Figure 4-5. Comparison of N271 normalized transport to Rosetta energy scores.

Corrected values from Figure 4-4 are plotted against Rosetta energy scores from Figure 4-1. Plots in the left-most column shows a comparison of the normalized transport compared to the inward-open model Rosetta energy scores. Plots in the middle column show normalized transport versus outward-open model energy scores and plots in the right-most column shows the normalized transport versus the difference between the inward and outward-energy scores. Letters indicate the amino acid replacement for the plotted variant. The "N" variant is wildtype NTCP. Correlation calculations using the Pearson (linear correlation) and Spearman (rank-order) coefficients were calculated and their values are shown in Supporting Table 4-1.

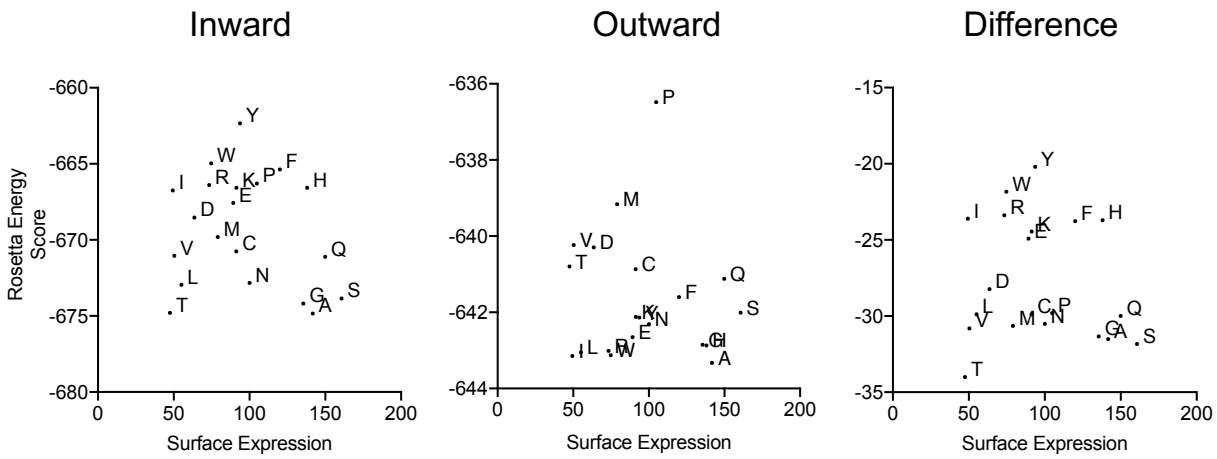


Figure 4-6. Comparison of Rosetta energy scores to NTCP N271 variant surface expression.

Surface expression values from Figure 4-3 are plotted against the energy scores calculated from the structure-based modeling in Figure 4-1. The energy scores from inward-open model is pictured on the left, outward-open model in the middle and the difference between the two models on the right. Correlation calculations using the Pearson and Spearman coefficients were completed and values are reported in Supporting Table 4-1. Wildtype is represented as "N" and all other points are labeled by their amino acid replacement.

In Chapter 3, we demonstrated a strong correlation between the transport of estrone-3-sulfate and rosuvastatin, and a lack of a correlation between these substrates and taurocholate, with respect to S267 variant transport. Thus, we concluded that estrone-3-sulfate and rosuvastatin likely share the same translocation pathway and as a result transport of these two substrates were similarly affected when position 267 was mutated. Given that position 271 is thought to be located outside of the translocation pore, we hypothesized that there would be fewer substrate dependent alterations in function. As a result, there would be strong correlations not only between the transport of estrone-3-sulfate and rosuvastatin but also between these two substrates and taurocholate when N271 was mutated. To test this hypothesis, we compared the transport rates of the different substrates by individual variants to determine whether alterations in transport were consistent from substrate to substrate (Figure 4-7) and calculated Pearson and Spearman correlation coefficients. These comparisons revealed strong correlations between the transport of all three substrates. All correlations had p-values of <0.0001 and coefficients varied between 0.8150 to 0.9564 with the greatest values for estrone-3-sulfate and rosuvastatin comparison (Pearson coefficient: 0.9527; Spearman: 0.9564) (Supporting Table 4-1). These strong correlations indicate that mutations at position 271 do not alter the specificity of the transported substrates.

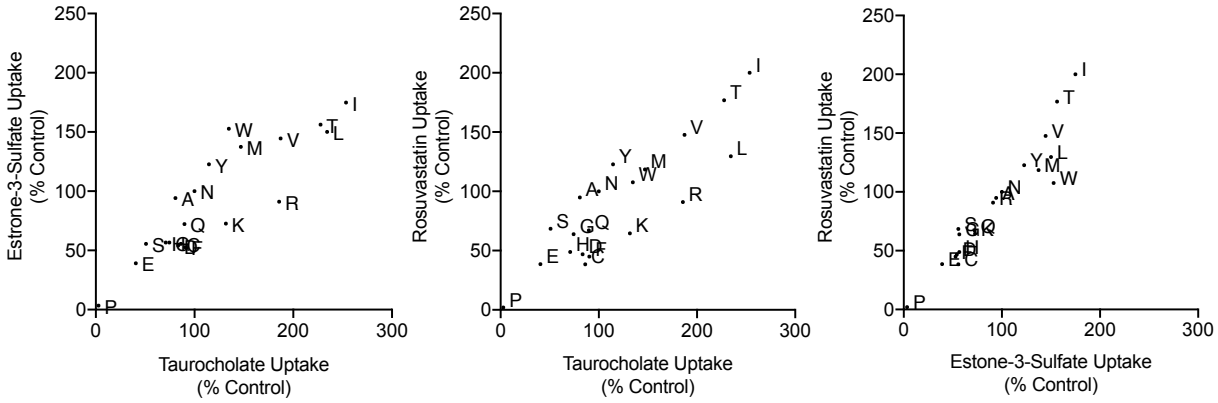


Figure 4-7. Correlation of normalized N271 variant transport.

Surface corrected uptake values from Figure 4-4 are plotted against each other. Variant amino acid replacements are indicated by their respective one-letter abbreviations; "N" is wildtype. The left-hand graph compares ^{3}[H]taurocholate to ^{3}[H]estrone-3-sulfate transport. Middle graph is showing the comparison of ^{3}[H]taurocholate transport to ^{3}[H]rosuvastatin uptake. Right-hand graph is demonstrating correlation of ^{3}[H]estrone-3-sulfate to ^{3}[H]rosuvastatin uptake. Linear (Pearson) and rank order (Spearman) correlation calculations were completed, and results are reported in Supporting Table 4-1.

Kinetic characterization of select N271 variants

The changes in variant transport demonstrated in Figure 4-4 could be due to changes in substrate affinity or turnover rate. To elucidate which of these parameters where altered, we selected three variants based on similarities and differences to wildtype in the initial surface corrected transport studies (Figure 4-4) and determined transport kinetics (Figure 4-8). Michaelis-Menten kinetics were calculated for each substrate and selected variants and are presented in Table 4-1. Visual presentation of kinetic values compared to surface corrected initial uptake results can be seen in Figure 4-9.

For these studies, we selected three variants. First, we chose cysteine because it demonstrated unique changes for all three substrates. Taurocholate transport was the same as wildtype while transport of estrone-3-sulfate and rosuvastatin was decreased. Kinetic experiments showed no real change in the transport capacity (V_{max}/K_m) of taurocholate or rosuvastatin by the cysteine variant compared to wildtype. However, the capacity of N271C to transport estrone-3-sulfate increased from 15 ± 2.3 μL/mg/min for wildtype to 30 ± 8.6 μL/mg/min (Table 4-1).

Next, the histidine variant was chosen because it showed a trend towards decreased transport compared to wildtype for all three substrates. However, these differences were not statistically significant. Unsurprisingly, there was no significant difference in the transport capacity of N271H for any of the substrates compared to wildtype.

Finally, we chose the leucine variant because it demonstrated elevated levels of initial transport for all three substrates compared to wildtype. Michaelis-Menten kinetics showed that the capacity of N271L was elevated for all three substrates from approximately 2-fold for taurocholate to almost 5-fold for estrone-3-sulfate and rosuvastatin. The lack of correlation between N271L and the other variants in Supporting Figure 4-1 further confirms that N271L shows the greatest differences in transport velocity (V_{max}) for all three substrates.

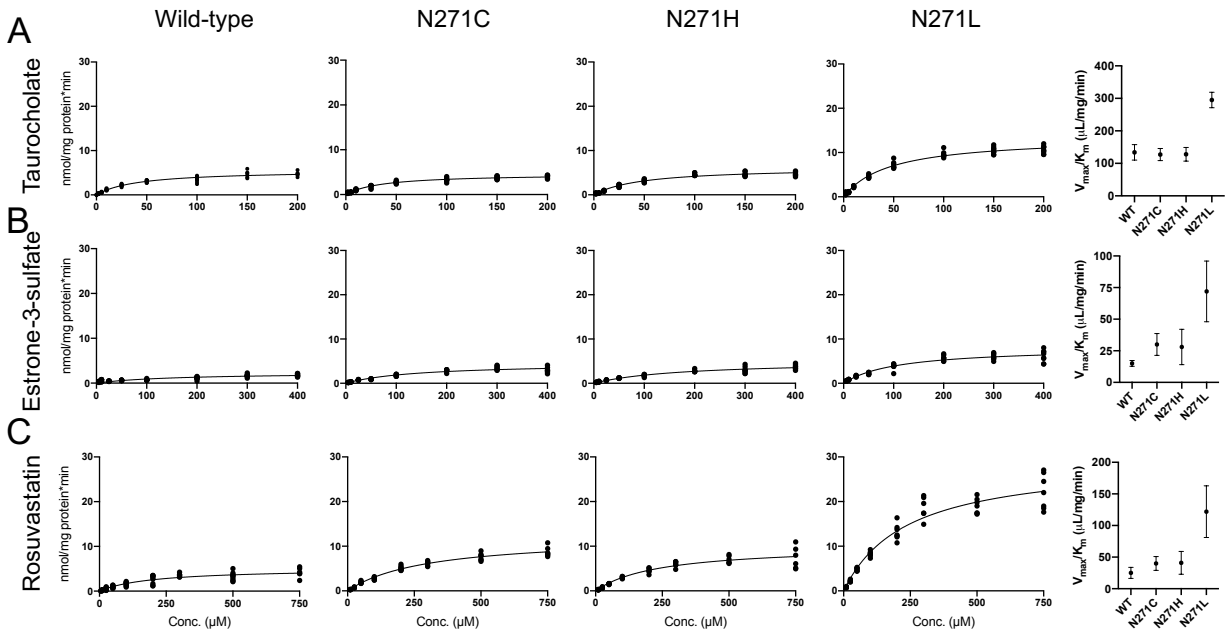


Figure 4-8. Substrate transport kinetics by wildtype and select NTCP variants.

Kinetic experiments were measured 48 hours post transfection in HEK293 cells under initial linear rate conditions using increasing concentration of the respective substrates. Kinetics for wildtype (first column) were previously reported in Ruggiero et al. 2021 and are shown here for visual comparison. NTCP variants N271C (second column), N271H (third column), and N271L (fourth column) were determined using *A.* taurocholate, *B.* estrone-3-sulfate, and *C.* rosuvastatin. The transport capacity (V_{max}/K_m) of all variants and substrates are plotted in the fifth column. The Michaelis-Menten equation in GraphPad Prism 8 was used to plot the curves of best fit and kinetic parameter results are reported in Table 4-1. Results were calculated and plotted from at least three independent experiments completed with 2-3 technical replicates and are shown as the mean ± SD.

Table 4-1. Kinetic values for substrate transport by wildtype and select N271 variants.

		NTCP	N271C	N271H	N271L
Taurocholate	K_m (μM)	42 ± 6.6	36 ± 5.1	49 ± 7.3	46 ± 3.7
	V_{max} (nmol/mg/min)	5.7 ± 0.5	4.6 ± 0.2	6.2 ± 0.4	13.5 ± 0.2
	V_{max}/K_m (μL/mg/min)	134 ± 24	127 ± 19	128 ± 21	295 ± 24
Estrone-3- Sulfate	K_m (μM)	155 ± 18	155 ± 39	190 ± 76	117 ± 30
	V_{max} (nmol/mg/min)	2.3 ± 0.2	4.6 ± 0.3	5.3 ± 1.7	8.5 ± 1.8
	V_{max}/K_m (μL/mg/min)	15 ± 2.3	30 ± 8.6	28 ± 14	72 ± 24
Rosuvastatin	K_m (μM)	183 ± 51	308 ± 74	253 ± 87	235 ± 60
	V_{max} (nmol/mg/min)	4.5 ± 1.0	12 ± 1.6	10 ± 2.7	29 ± 5.9
	V_{max}/K_m (μL/mg/min)	25 ± 8.9	40 ± 11	41 ± 18	122 ± 41

Kinetic values, K_m and V_{max}, calculated using the Michaelis-Menten equation in GraphPad Prism 8 from Figure 4-8 are reported. Transport of increasing concentrations of taurocholate, estrone-3-sulfate, and rosuvastatin by HEK293 cells transiently transfected with either empty vector or NTCP variants N271C, N271H, or N271L was measured under initial linear rate conditions, 48 hours post-transfection. Net uptake was determined by

subtracting transport by empty vector transfected cells from the transport by the N271 variants. Wildtype NTCP results were previously published in Ruggiero et. al 2021 and are shown here for comparison. Transport capacity (V_{max}/K_m) was calculated by dividing the V_{max} by the K_m. Kinetic experiments were repeated at least three individual times with 2-3 technical replicates completed in each experiment. Results in this table are reported as the mean of those repeated experiments ± SD.

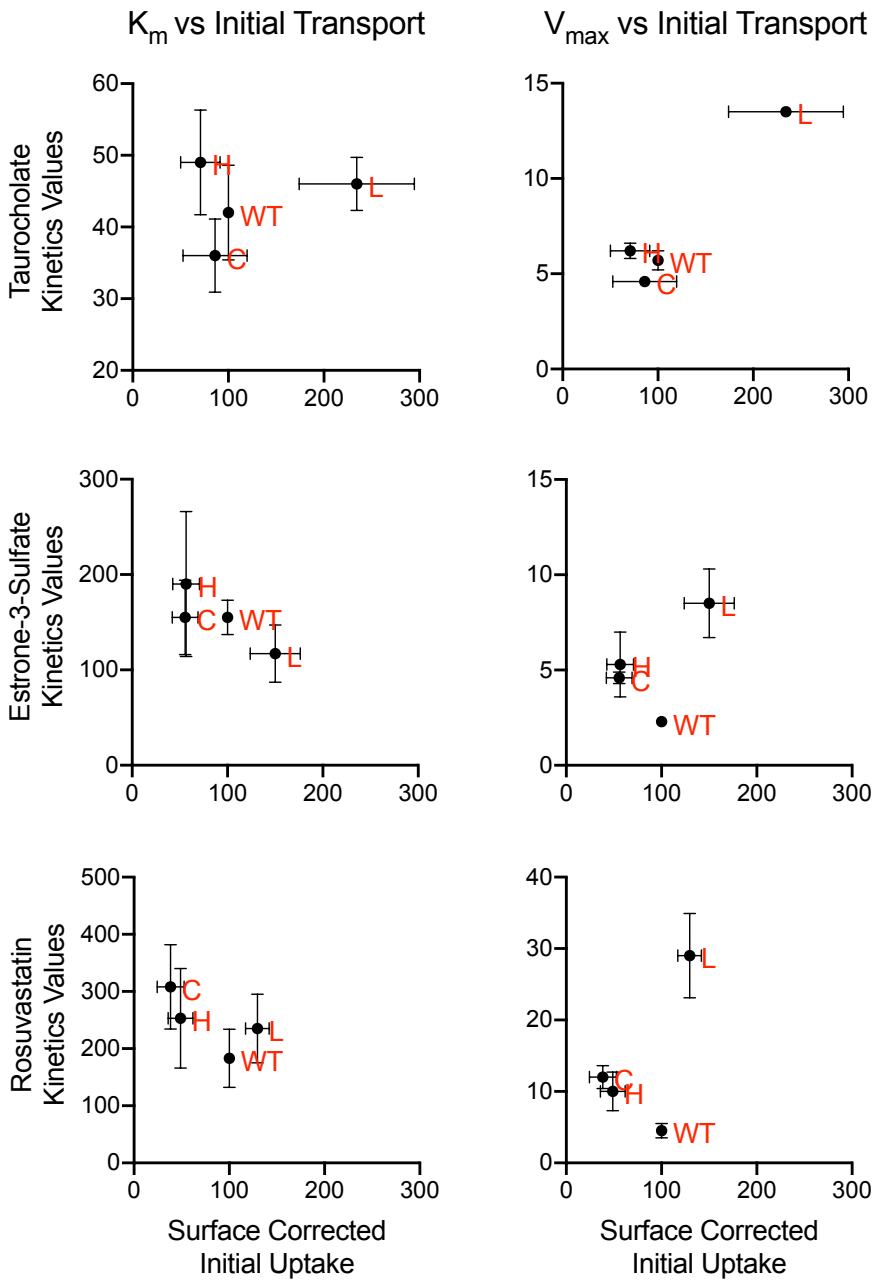


Figure 4-9. Visual representation of correlation between kinetic values versus surface corrected initial transport for select N271 variants.

Kinetic values, K_m (left column) and V_{max} (right column), for select N271 variants from Table 4-1were plotted against their surface corrected initial uptake values (Figure 4-4) for each substrate (taurocholate, first row; estrone-3-sulfate, second row; and rosuvastatin, third row). Comparisons are meant to aid in visual observations of changes in kinetic values. Red letters correspond to wildtype (WT) and the amino acid substitutions in position 271.

Selection of additional positions and their function and expression outcomes

At this point, we have demonstrated that two positions, 267 (see Chapter 3) and the current position of interest 271, where both subjectively rheostatic for functional and expression outcomes. This led to the question of whether mutations at all amino acid positions in NTCP would result in rheostatic results. Thus, we used the multiple sequence alignment from Chapter 3 along with entropy and ConSurf analyses to select two additional positions. First, we searched for a highly conserved and thus "evolutionarily important" position. Glycine at position 102 has a ConSurf score of 9, the highest possible value, and an entropy score of 0.0054, one of the lowest values found in the sequence alignment. Both values, in agreement with each other, indicate that the glycine at position 102 is important for protein stability and function. Therefore, any deviation from the natural amino acid at position 102 will likely results in detrimental if not catastrophic outcomes. Furthermore, if we were to replace the native glycine with all other amino acids, we would expect to see an overall toggle-like outcome.

To test this, we substituted glycine for all other 19 amino acids at position 102 and examined changes in the function using taurocholate, estrone-3-sulfate, and rosuvastatin (Figure 4-10). Figure 4-10 shows that position 102 seems to be most rheostatic for taurocholate and less so for estrone-3-sulfate and rosuvastatin. As is generally expected for a highly conserved position, with the exception of taurocholate, the wildtype was by far the best functionally for each substrate. The only variant similar to wildtype for taurocholate transport is alanine. In addition, alanine is the next best variant compared to wildtype for estrone-3-sulfate and rosuvastatin transport. However, its transport is only 24 and 44 percent of wildtype, respectively. Most variants show diminished transport for all three substrates, again something expected for an important and conserved position.

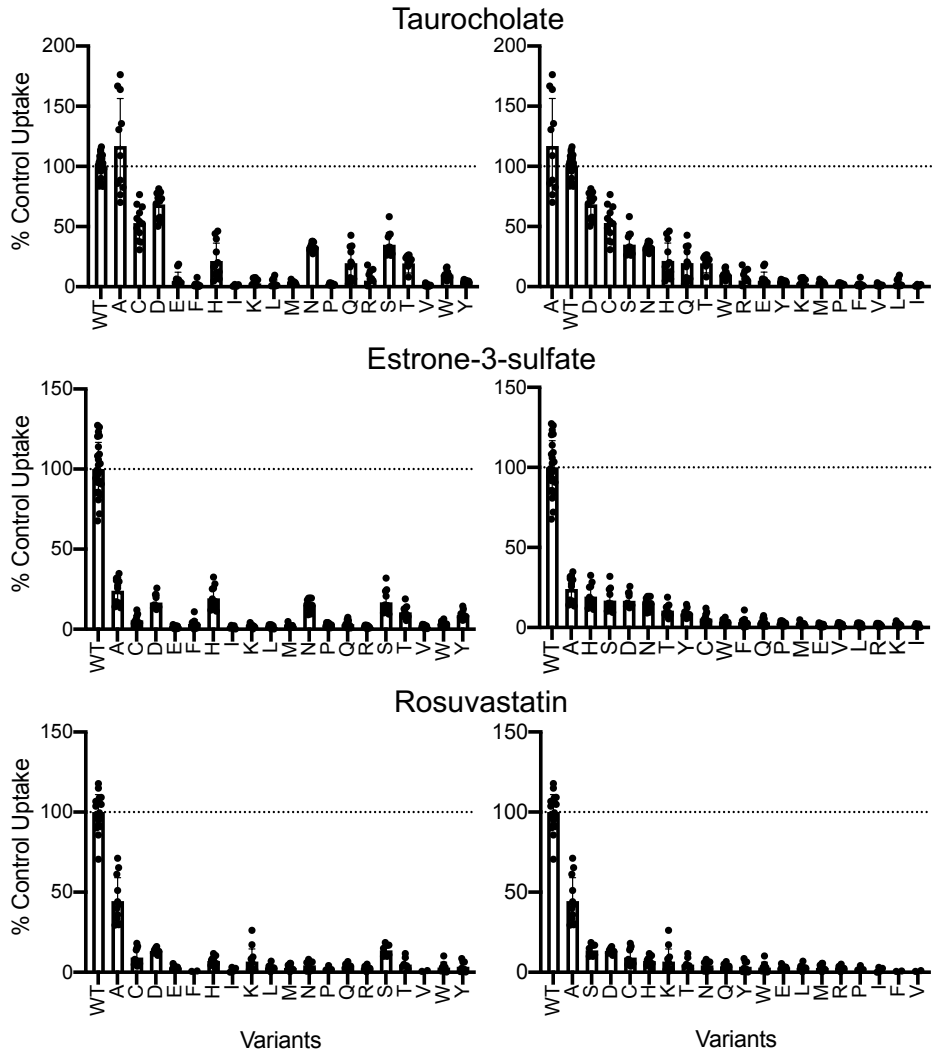


Figure 4-10. Transport of model substrates by wildtype NTCP and G102 variants.

Similar experiments to Figure 4-2 were completed with the NTCP G102 variants. In brief, HEK293 cells were transiently transfected with either empty vector, wildtype NTCP (WT), or G102 variants and then transport of tritiated taurocholate, estrone-3-sulfate, and rosuvastatin was measured. Results are net uptake and variants were normalized to wildtype, which was set to 100%. The left side shows transport of wildtype followed by the G102 variants sorted alphabetically by their amino acid substitutions and the right-hand side is organized using rank order of transport. Individual data points from three biological replicates with at least 2-3 technical replicates are reported with the bar indicating the mean of all replicates ± SD. Horizontal line indicates 100%. For more detailed description see Figure 4-2 legend or text.

So many G102 variants showing diminished transport raised the question of whether these variants are expressed at the surface of the cell and are not functional or whether they are not expressed. Figure 4-11A shows the result of the surface biotinylation assay. The quantification in Figure 4-11B demonstrates that a lack of surface expression for isoleucine, leucine, lysine, arginine, and valine is the explanation for the non-functional variants seen in Figure 4-10. However, most variants are still being made but are not making it to the membrane as seen in Figure 4-12. When corrected for surface expression in Figure 4-13, overall, the order and message does not change from Figure 4-10 to Figure 4-13. However, there are some variants with lower expression that when normalized for surface expression show increased transport activity. One example is aspartic acid, which is closer to wildtype after surface correction.

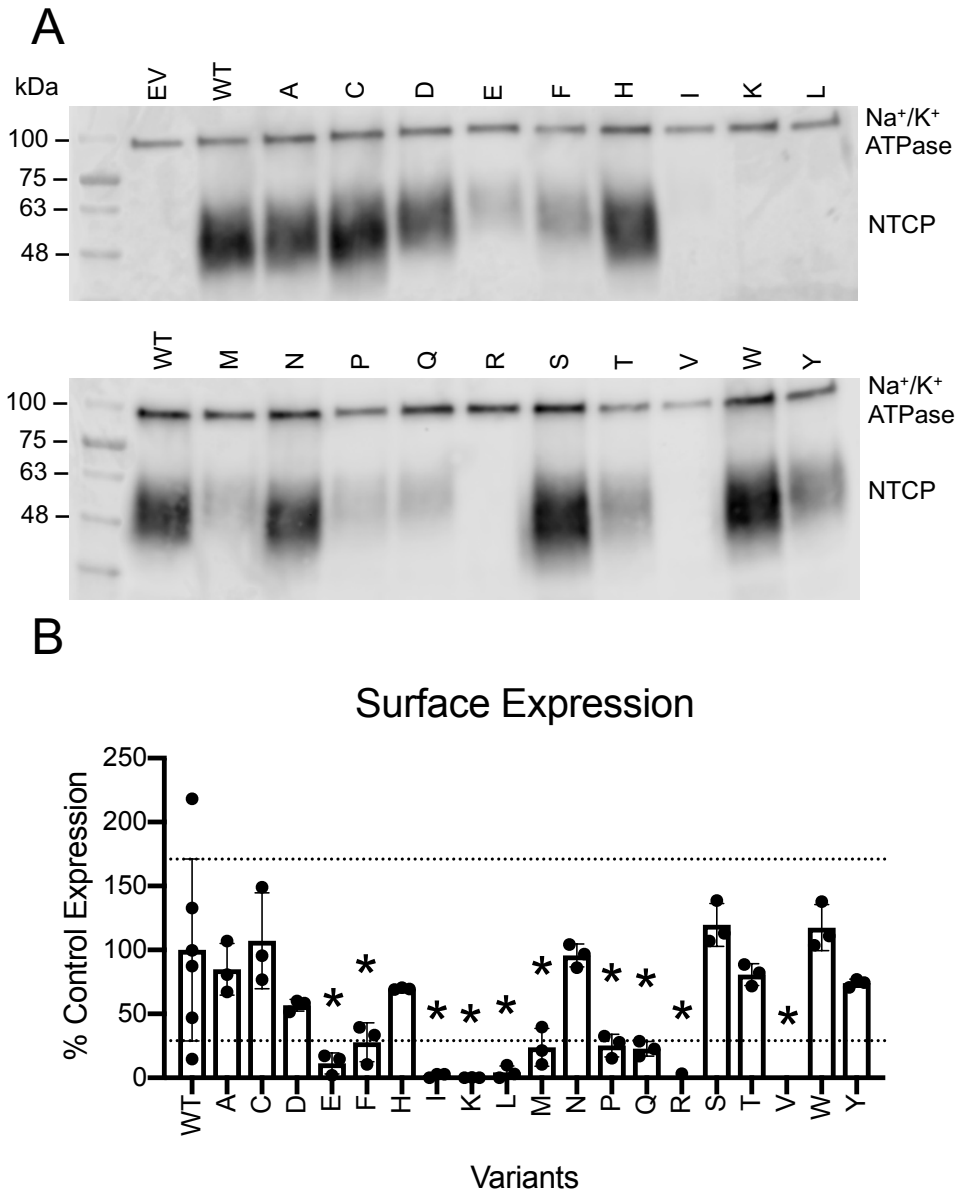


Figure 4-11. Surface expression and quantification of wildtype NTCP and G102 variants.

Experiments similar to those in Figure 4-3 were completed with the G102 variants. In short, *A.* a representative western blot of the surface expression of wildtype (WT) NTCP and G102 variants. Blots were probed simultaneously with Na^+/K^+-ATPase (100kDa) and Tetra-His antibodies. *B.* Quantification of wildtype (WT) and G102 variant surface expression. Three independent surface expression experiments were quantified, and individual data points are shown with the bar graphs representing the mean ± SD. Horizontal lines are shown at the upper and lower limit of the wildtype standard deviation. Asterisks denote statistically significant differences ($p<0.05$) from wildtype NTCP. For more detailed description see Figure 4-3 legend or text.

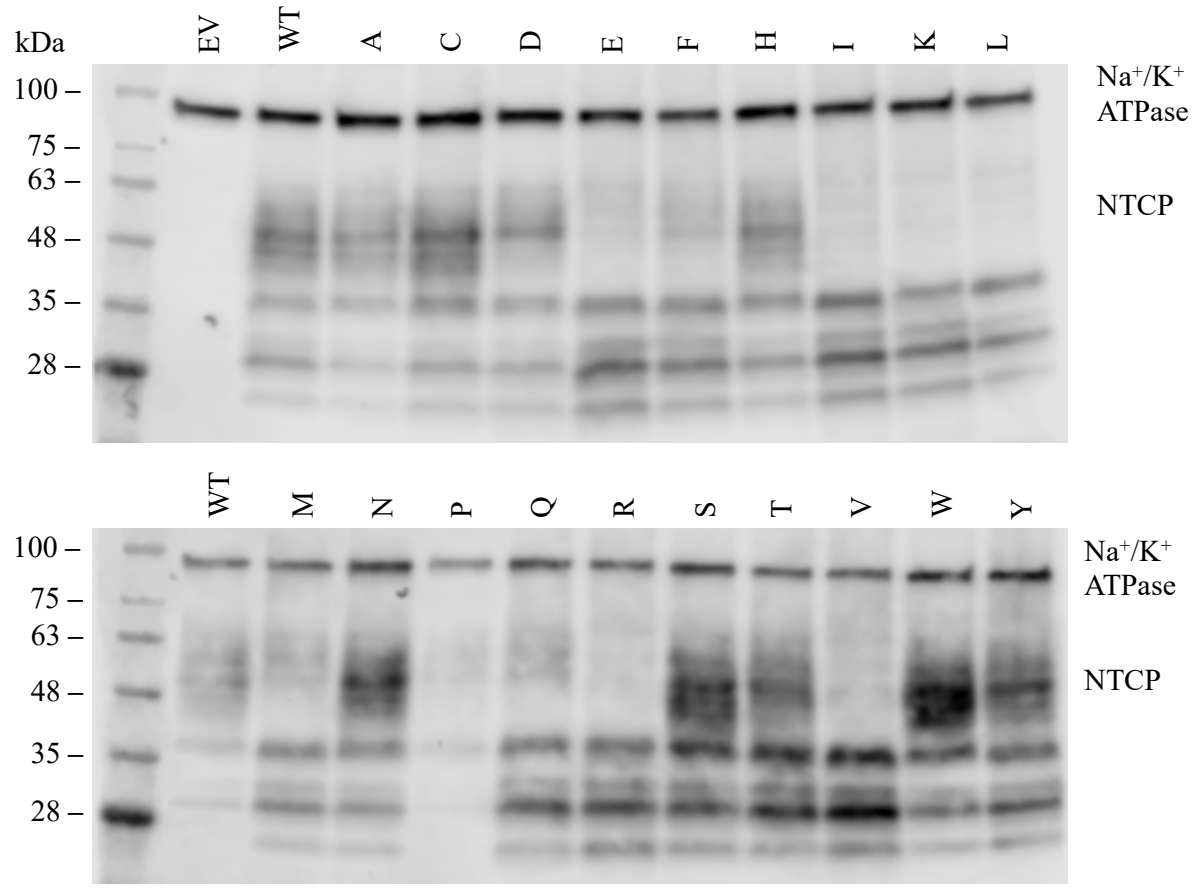


Figure 4-12. Representative western blots of total protein expression of wildtype and G102 variants in transiently transfected HEK293 cells.

Total protein samples from surface biotinylation experiments were electrophoresed on 4-20% gradient polyacrylamide gels for one hour at 150 volts. Separated protein was then transferred to nitrocellulose membranes and then probed with Na^+/K^+-ATPase as a loading control (100kDa) and Tetra-His antibodies to detect His-tagged protein.

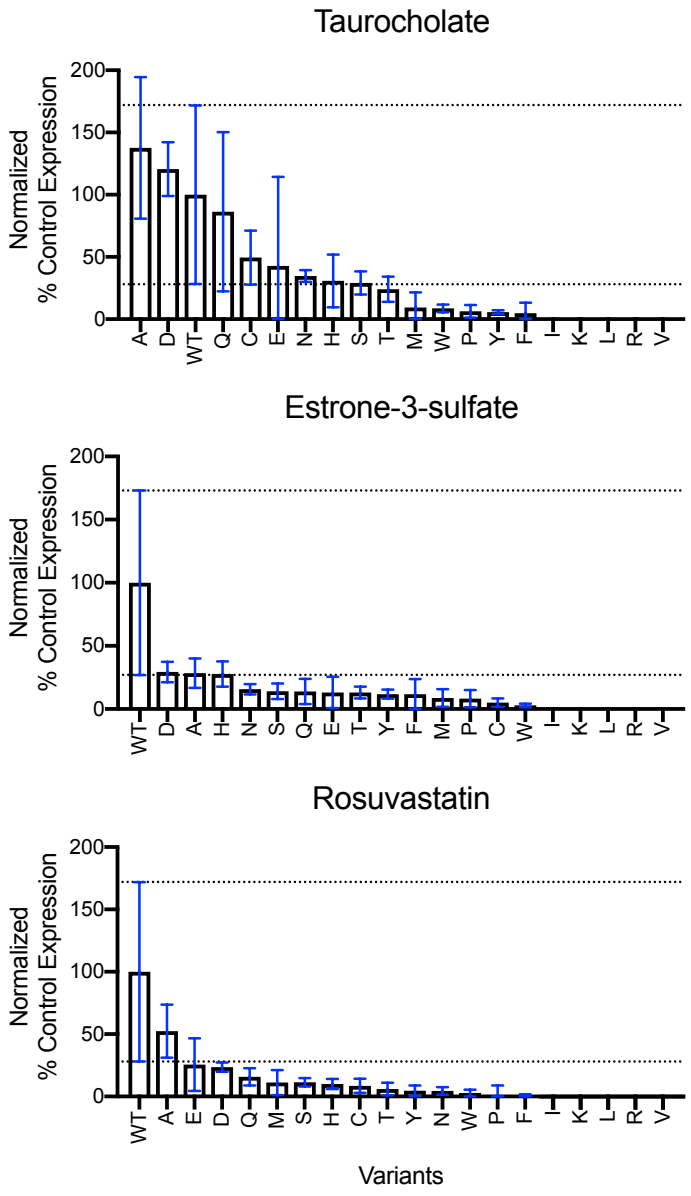


Figure 4-13. Substrate transport of wildtype NTCP and G102 variants corrected for surface expression.

Initial uptake results from Figure 4-10 were corrected for the surface quantification in Figure 4-11*B*. Normalized transport is shown with variants and wildtype NTCP (WT) ordered from greatest to least uptake for each substrate. Averaged corrected values are indicated by bars ± the propagated SD. Horizontal lines denote the upper and lower standard deviation limit for wildtype. At least two technical replicates from three independent experiments are shown. Variants G102I, K, L, R, and V were excluded due to no surface expression and low transport indicating that these variants are not expressed at the cell surface and therefore transport cannot be corrected for expression.

In addition to position 102, position 146 was also selected to be examined. It was chosen because it has an entropy and ConSurf score similar to position 271, 1.91 and 7 respectively. However, it differs from position 271 in that it is predicted to be located in an extracellular loop while N271 is predicted to be buried in a transmembrane domain. Again, we substituted the native amino acid for all other 19 amino acids and measured uptake of taurocholate, estrone-3-sulfate, and rosuvastatin, as well as the surface expression of the resulting variants. With the exception of serine and asparagine, all variants at Y146 showed very similar transport compared to wildtype, especially for estrone-3-sulfate (Figure 4-14). Surface quantification of the Y146 variants revealed that neither serine nor asparagine were expressed at the plasma membrane (Figure 4-15). A western blot of total protein samples confirmed that a non-glycosylated Y146S (at approximately 35 kDa) is made but this variant does not make it to the membrane (Figure 4-16). In contrast, Y146N is not expressed at all. Given the only minor variations in surface expression compared to wildtype, there is little change after normalizing transport for surface expression (Figure 4-17). Further, most difference that were observed were not statistically significant.

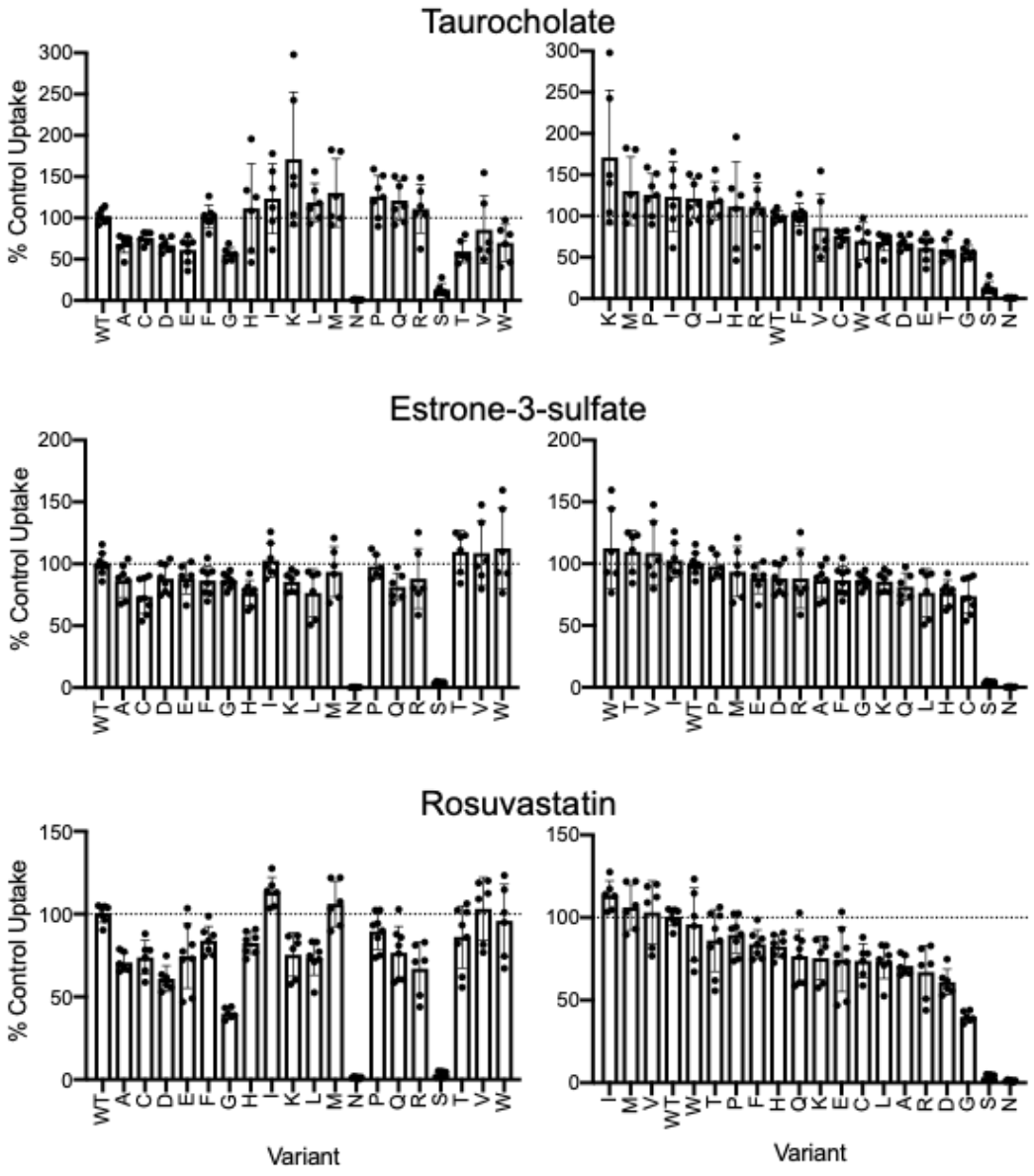


Figure 4-14. Uptake of model substrates by wildtype NTCP and Y146 variants.

Experiments comparable to those in Figure 4-2 were completed with NTCP Y146 variants. Briefly, transport of radiolabeled taurocholate, estrone-3-sulfate, and rosuvastatin by empty vector, wildtype NTCP (WT) and Y146 variants was measured. Net uptake was determined by subtracting the transport by empty vector transfected cells from the transport by NTCP and NTCP Y146 variant expressing cells. Variants are shown in order from highest to lowest transport. Individual data points are biological replicates from three individually measured experiments with at least 2-3 technical replicates each. Mean results are indicated by the bar graph ± SD. Variant transport is shown as a percent of the wildtype control, which is set to 100% and indicated by the horizontal line.

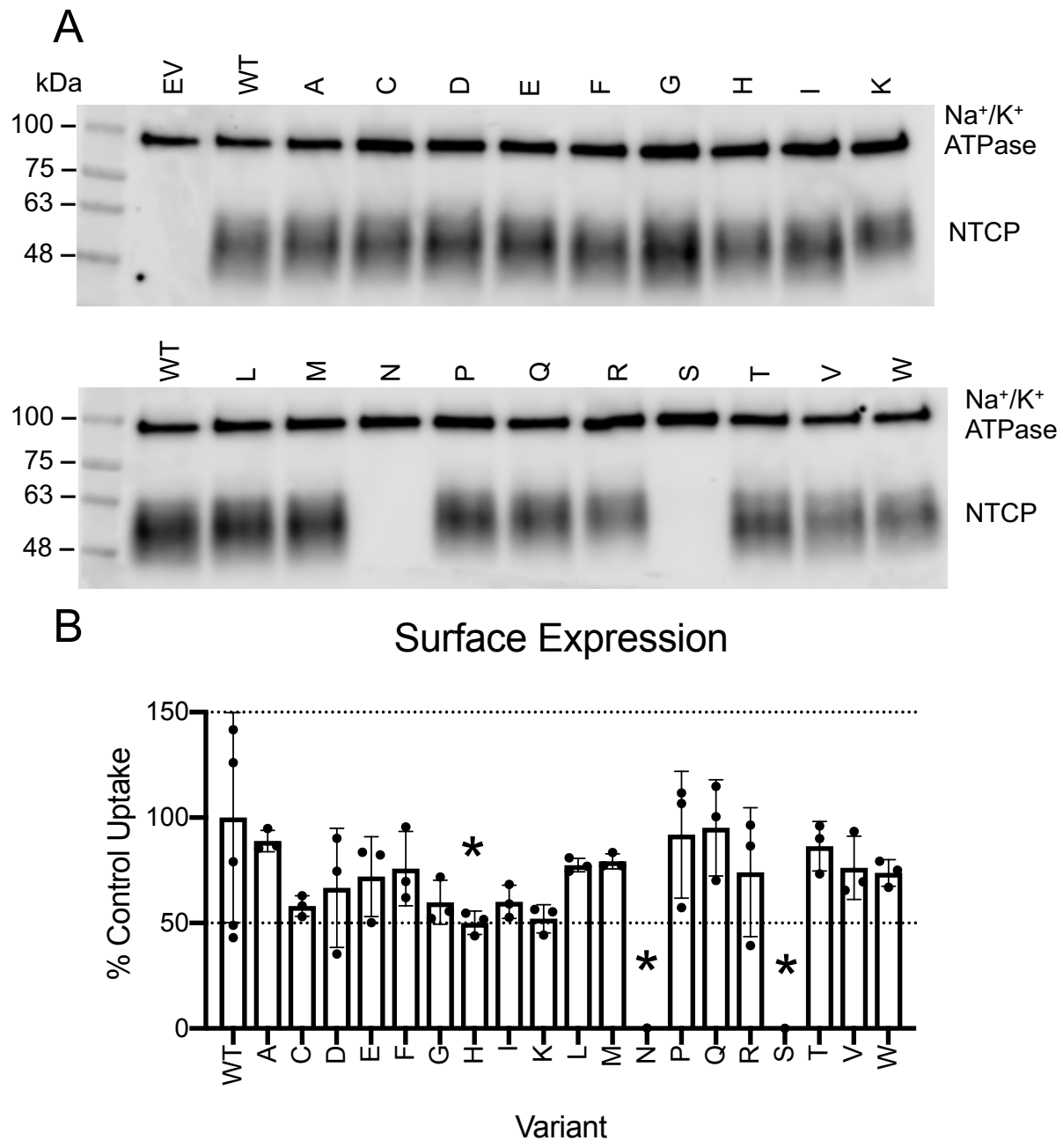


Figure 4-15. Surface expression and quantification of wildtype NTCP and Y146 variants.

Experiments similar to those in Figure 4-3 were completed with the Y146 variants. In short, *A.* surface expression of wildtype (WT) NTCP and Y146 variants on a representative western blot. *B.* Quantification of wildtype (WT) and Y146 variant surface expression. Three independent surface expression experiments were quantified, and individual data points are shown with the bar graphs representing the mean ± SD. Horizontal lines shown at the upper and lower limits of wildtype NTCP standard deviation. Significant differences from wildtype NTCP are denoted by asterisks ($p<0.05$). For more detailed description see Figure 4-3 legend or text.

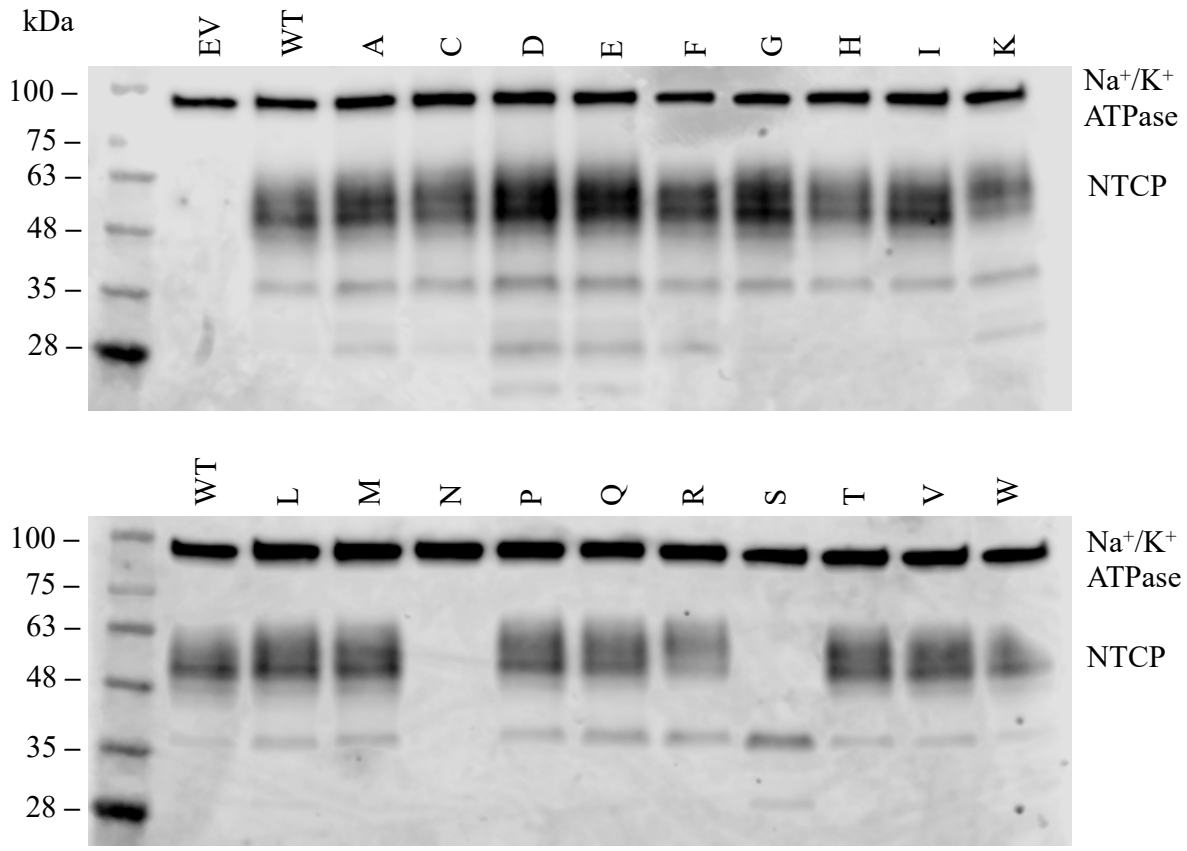


Figure 4-16. Representative western blots of total protein expression of wildtype and Y146 variants in transiently transfected HEK293 cells.

Samples were electrophoresed and probed using the method outlined in Figure 4-12.

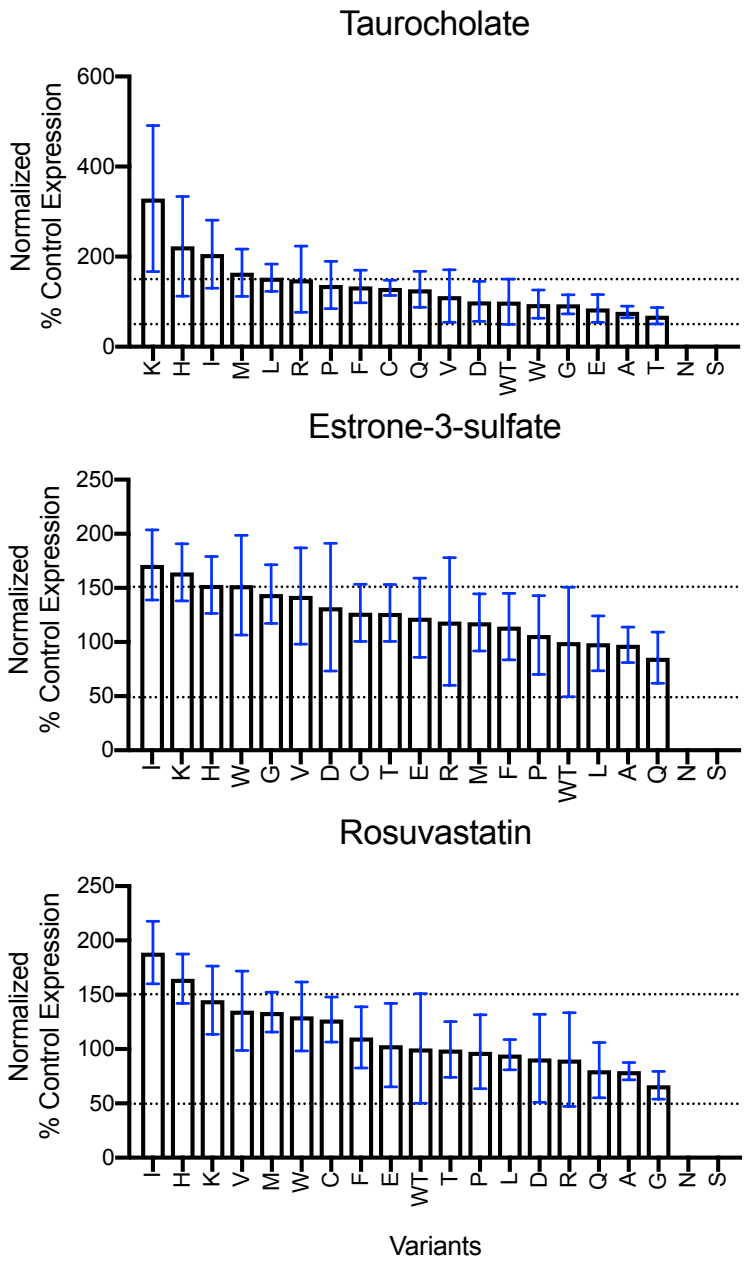


Figure 4-17. Y146 variant substrate transport corrected for surface expression.

Initial uptake results from Figure 4-14 were corrected for the surface quantification in Figure 4-15*B*. Normalized transport is shown with variants and wildtype NTCP (WT) ordered from greatest to least uptake for each substrate. Variants Y146N and S were excluded due to no measurable surface expression (Figure 4-15) and low transport (Figure 4-14) indicating that these variants are not expressed at the cell surface and therefore transport cannot be corrected for expression. Bars indicate the mean ± the propagated SD of the corrected values from at least two technical replicates from three independent experiments. Horizontal lines denote the upper and lower limits of wildtype propagated error.

RheoScale calculations and outcomes

Up to this point, our observations about the strength or weakness of the observed outcomes were strictly subjective. However, in order to make conclusions about the observed outcomes we needed a method to objectively quantify the results. After characterizing the three positions N271, G102, and Y146, we systematically assessed the overall substitution outcomes of these positions using the RheoScale calculator created by Hodges et al. 2018. For these calculations, we compared the surface expression and the surface-normalized transport from all four characterized positions, including the previously described 267 position in Chapter 3. This calculator takes all the data and distributes the variants into bins in a histogram analysis. Histogram bin analyses for the 10 bin analysis is shown in Figure 4-18. If the majority of the variants (at least 70%) fall into the same bin as wildtype, the position is trending towards "neutral" (score of 0.7 or more) (Martin et al., 2020). An example histogram for a neutral position is demonstrated by estrone-3-sulfate for Y146 in Figure 4-18 (bottom row, second column). In contrast, if at least two-thirds of the variants fall into the bin designated as "dead", meaning the variant is not functional or not expressed, then it is classified as a toggle (score of 0.67) (Wu et al., 2019). Lastly, if the variants sample at least half of the possible functional range, meaning that variants fall into at least 50% of the bins, then the position is considered a rheostat (score of 0.5 or greater) (Hodges et al., 2018). Strong rheostatic histograms are demonstrated by position 102 for both rosuvastatin transport and surface expression in Figure 4-18 (third row, third and fourth columns). In addition, to ensure we appropriately captured the neutral variants, we manually calculated the number of variants whose average surface expression or normalized transport fell within the wildtype standard deviation (dashed lines in Figures 4-3, 4-4, 4-11, 4-13, 4-15, 4-17) and summarized them in Table 4-2.

Subjectively, position 146 trends the most towards neutral for all three substrates for both the manually calculated and the RheoScale scores. However, it does not reach the neutral threshold of 0.7 for surface expression when calculated using RheoScale. Position 102 is a rheostat for surface expression and all three substrates, while position 267 is a stronger rheostat for all substrates but a neutral for surface expression. Position 271 is rheostatic, with a score of 0.52, for both taurocholate and rosuvastatin transport but does not reach the threshold for estrone-3-sulfate. Nevertheless, since it is a rheostat position for any of the measured functional parameters, we classify this as a rheostat position (Wu et al., 2019).

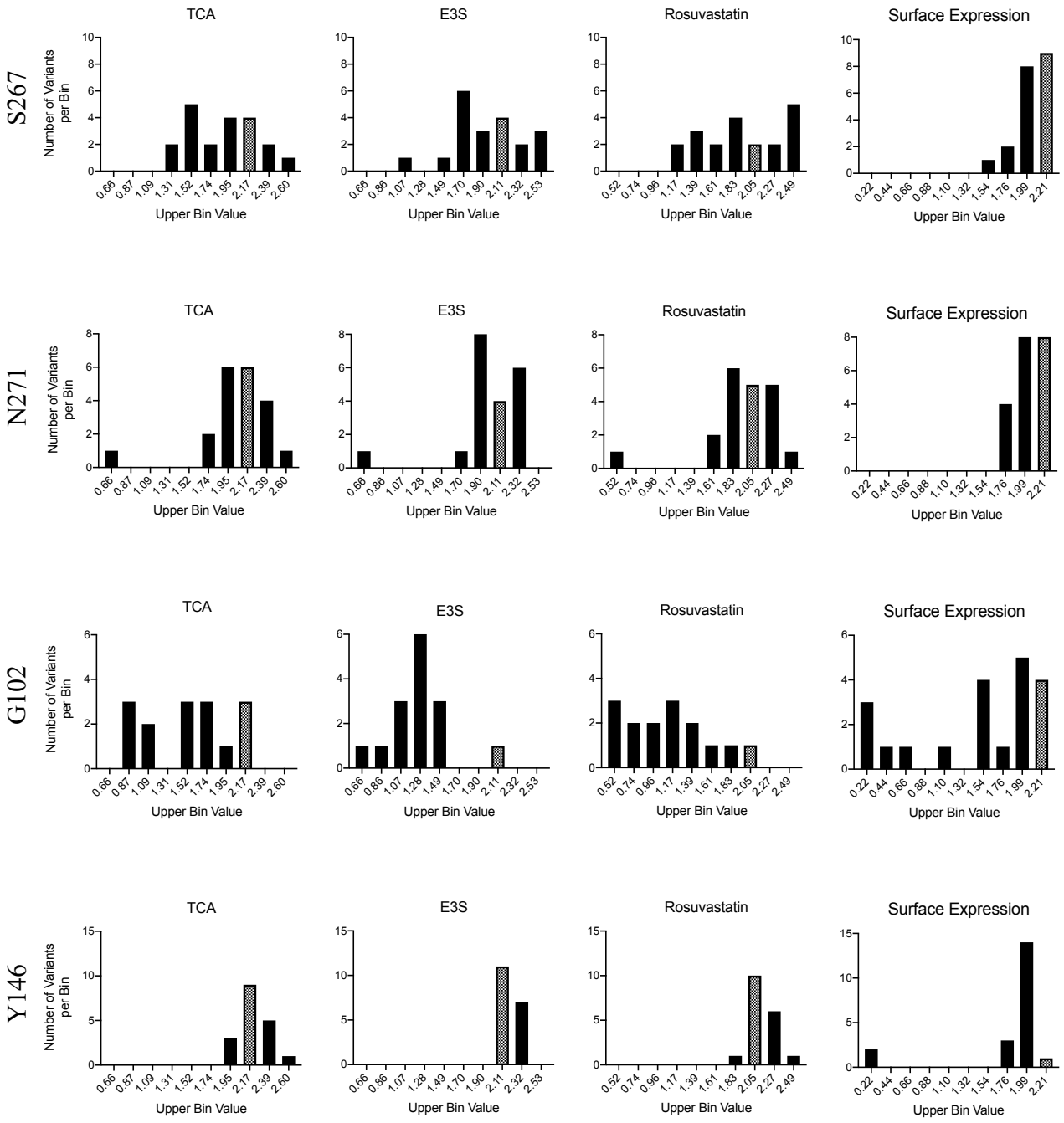


Figure 4-18. RheoScale bin analysis histograms.

Graphical depiction of the 10 bin analysis used by RheoScale for the determination of rheostat and neutral scoring for each position: S267 (top row), N271 (second row), G102 (third row), and Y146 (bottom row) and their data sets: taurocholate (first column), estrone-3-sulfate (second column), and rosuvastatin (third column) transport as well as surface expression (fourth column). Scores for analysis are presented in Table 4-2 (column 6). Bin that includes wildtype is indicated by the patterned bar.

Table 4-2. RheoScale values of mutational outcomes in NTCP

Experiment	Position	# of variants	Neutral	Rheostat	Toggle
Surface Expression	S267	20	**0.79**	0.38	0.00
	N271	20	0.32	0.25	0.00
	G102	20	0.47	**0.75**	0.16
	Y146	20	**0.84**	0.29	0.11
TCA Initial Uptake	S267	20	0.63	**0.74**	0.00
	N271	20	0.32	**0.52**	0.05
	G102	15	0.50	**0.61**	0.00
	Y146	18	**0.76**	0.35	0.00
E3S Initial Uptake	S267	20	0.37	**0.74**	0.00
	N271	20	0.26	0.39	0.05
	G102	15	0.14	**0.57**	0.07
	Y146	18	**0.88**	0.13	0.00
Rosuvastatin Initial Uptake	S267	20	0.42	**0.74**	0.00
	N271	20	0.32	**0.52**	0.05
	G102	15	0.07	**0.78**	0.21
	Y146	18	**0.88**	0.35	0.00

Experimental data for the four characterized positions in NTCP (S267, N271, G102 and Y146) were analyzed with the RheoScale calculator. Surface expression data (top section) along with the data from surface corrected initial uptakes for each substrate, taurocholate (TCA), estrone-3-sulfate (E3S) and rosuvastatin, were compared. Variants with no surface expression (less than 1%) were excluded from the uptake calculations and denoted by a change in the # of variants column (G102 variants I, K, L, R and V; Y146 variants N and S). Rheostat and toggle scores were calculated using 10 bins while neutral scores were manually calculated. Neutral scores (column 4) were determined by visually counting the variants whose averages fell within the wildtype's SD for each data set and then dividing by the total number of variants in the data set. Scores are calculated on a scale of 0 (low) to 1 (high). Empirical thresholds determined for significant scores in previous works were: 0.5 rheostat, 0.7 neutral, and 0.67 toggle (Hodges et al., 2018; Wu et al., 2019; Martin et al., 2020). Values that exceed their significance thresholds are in bold.

III. **Discussion**

At the onset of this study, our fundamental question was whether we could use structure-based simulated Rosetta energy scores to predict mutational tolerance indicating a rheostat location within NTCP and whether these scores would correlate to variant protein expression. Our results in Chapter 3 showed a slight correlation between the surface expression of S267 variants and the variants' simulated energy scores. Based on this we hypothesized that there was a correlation between the energy scores of a mutated protein and the protein's overall stability. For example, if the energy scores were increased for a particular variant, the protein would be less stable and, as a result, more likely to be degraded or inefficiently trafficked to the cell surface. In other words, the calculated energy scores would give insight into NTCP's stability which would directly impact its surface expression. To test this hypothesis, we selected N271, an amino acid position whose energy scores resembled a rheostat, indicating that mutations at this location would be tolerated. We then investigated whether amino acid substitutions at this location would result in rheostatic outcomes for surface expression.

In order to quantify the outcomes for surface expression and transport, we applied the RheoScale calculator to each data set (Hodges et al., 2018). The results from the RheoScale showed that the surface expression for position 271 was close to neutral, demonstrating that mutations were well tolerated but not rheostatic as originally predicted (Table 4-2). Moreover, the comparison between the simulated energy scores and surface expression in Figure 4-6 further confirmed that there is no correlation between these two measurements. With respect to function, position 271 reached the rheostat threshold for both taurocholate and rosuvastatin transport, but did not show any strong scores for estrone-3-sulfate uptake. Thus, based on the functional outcomes we can classify position 271 as a rheostat, but there was no basis for the claim that the energy scores

predicted this outcome. Therefore, we conclude that the methods we are currently using to calculate the simulated energy scores are able to accurately predict mutational tolerance at position 271 and thus leading to rheostatic outcomes. However, the energy scores did not correlate with the surface expression thus they cannot currently be used to predict protein expression. Consequently, our hypothesis from Chapter 3 that simulated energy scores could predict protein expression, was false. However, the energy scores did correctly predict that mutations at position 271 would be tolerated and position 271 was confirmed as rheostatic for certain substrates.

One explanation for the disagreement between the outcomes observed and the rheostatic prediction could be discrepancies in our structure model. To our knowledge, human NTCP has not been crystalized yet. Thus, the model we used is based on crystalized bacterial homologs that only show low amino acid similarities to NTCP (Claro da Silva et al., 2013; Zhou et al., 2014). It is possible that the folding and structure of NTCP may vary significantly from the homologs that were used for our modeling. This, in turn, could impact the simulations and therefore the energy scores. In addition, other factors can impact protein stability and therefore the surface expression. For example, mutations may slightly alter the structure of the protein. While the protein is able to re-structure itself or "re-pack" there may be additional downstream effects, including disruption of post-translational modifications that as a result alter localization in the plasma membrane. However, at this point in time, our structure-based modeling is unable to consider these factors. Nevertheless, the models will need to be improved to determine if this method can predict rheostatic outcomes in the future.

As indicated above, with the progression of this study it became increasing clear that mutations in a protein like NTCP could be classified as rheostatic in terms of function or expression. This

observation led us to reevaluate how we classify rheostatic outcomes and begin differentiating between functional and expression rheostats. As mentioned, the stability and energy of a protein is more likely to correlate with the expression rather than the transport function. Furthermore, if a position is rheostatic it can be either rheostatic for function or expression or both. To complicate this dynamic further, multi-specific transporters known to transport a variety of substrates could be rheostatic for certain substrates and not others. As explained above, this was the case for position 271 where we observed rheostatic behavior for the uptake of taurocholate and rosuvastatin but not for estrone-3-sulfate.

Up to this point we had identified two rheostatic amino acid locations in NTCP, S267 and N271. Our second question was if we could accurately identify locations within NTCP that, when mutated to other amino acids, would not result in rheostatic outcomes. We selected a nonconserved position, tyrosine 146 because numerous amino acids were naturally allowed at this position in a variety of different species. Our hypothesis was that amino acid substitutions at Y146 would be widely accepted and as a result exhibit more neutral-like outcomes. In support of our hypothesis, minor changes in the expression and transport for most variants at position 146 confirmed that the majority of amino acid substitutions were accepted. RheoScale calculations (Table 4-2) identified that position 146 is neutral, ultimately demonstrating that not all amino acid substitutions within NTCP will result in rheostatic outcomes.

Interestingly however, two amino acids at position 146, serine and asparagine, were detrimental to surface expression and therefore function as well. Surprisingly, there is nothing biochemically remarkable about these two amino acids that could explain this phenomenon. In addition, examination of total protein expression (Figure 4-18) demonstrated that Y146S was produced as a non-glycosylated protein and unable to reach the membrane, while Y146N was not detectable

in the total membrane fraction. This indicates that the non-glycosylated Y146S is not properly trafficked to the membrane while Y146N likely gets degraded during translation or shortly thereafter. One explanation is that NTCP is unable to properly re-pack to accommodate these two amino acids thus causing instability, protein degradation, and disruption of post-translational modifications. Further experiments would be necessary to determine at which stage these two variants are degraded or sequestered.

In addition to Y146, we selected an evolutionarily highly conserved position, G102, where mutations should result in detrimental or toggle-like outcomes. Unsurprising, numerous G102 variants did not transport well and had low surface expression (Figure 4-10, 4-11, and 4-13). These results indicate that most mutations at G102 are indeed detrimental, consistent with the interpretation that nature naturally eliminates variations at position 102 due to these deleterious effects and explaining why this position is evolutionarily so highly conserved. In addition, examination of total protein samples from surface experiments showed that variant proteins were in various states of glycosylation and degradation, further indicating instability and explaining the decreased surface expression (Figure 4-13).

However, when the RheoScale calculator was used to examine the outcomes of amino acid substitutions at position 102, different conclusions could be drawn. Position 102 is the only location examined in these studies that reached the threshold for rheostat for all four data sets: surface expression as well as transport of all three substrates. This indicates that, while substitutions at G102 cause significant decreases in function and expression, NTCP is able to re-structure itself to accommodate the altered amino acid. Thus, NTCP allows for some function and expression depending on the amino acid replacement. Although these results contradict the prediction that G102 would be a toggle position, it is still clear that wildtype demonstrates the

greatest fitness. However, this does lead to the question of whether or not a true toggle position exists in NTCP.

In conclusion, we have identified two additional rheostat positions as well as a neutral position within NTCP. While these positions are not clinically relevant, they have allowed us to gain greater insight into the complexity of rheostat positions as well as further confirm that current prediction methods are insufficient. We also determined that stability and surface expression do not always influence function. This may be due to conformational changes, interactions with cellular machinery, or other factors that we are unable to consider or measure at this time. However, another concept is becoming more clear: the more locations we examine the more complex and dynamic function, expression, and prediction of rheostat positions become. In order for patterns to begin to emerge to aid in the prediction of mutational outcomes at "unpredictable" locations, even more positions and other transport proteins may need to be examined.

Chapter 5

Overall Summary and Conclusions

Predictive pharmacogenomics can shift the medical field from the use of standard care into an era of personalized or precision medicine characterized by effective treatments, decreased adverse events, and lower medical costs (Sankar and Parker, 2017; Phillips et al., 2018). Ideally, every mutation in each patient would be experimentally characterized to understand the effects of mutations on drug response. However, this approach is impractical in terms of both time and money. As a result, we must rely on bioinformatics and computational models to predict protein mutation outcomes and their impact on drug metabolism, disposition, and efficacy. While prediction algorithms are continually modified, more improvement is needed (Radivojac et al., 2013). For example, current algorithms are primarily based on outcomes at conserved amino acid locations. Further, they predict binary outcomes, detrimental (deleterious) and non-detrimental (allowed), regardless of the mutated amino acids conservation. Consequently, mutations that do not fit these criteria are inaccurately predicted. For instance, studies conducted on variants of LacI/GalR homologs demonstrated that mutations at nonconserved locations did not follow the same outcomes at conserved locations (Miller et al., 2017). Rather, when a nonconserved location was mutated to all other nineteen amino acids, the results demonstrated a continuum of functional outcomes akin to a dimmer switch. This outcome is now known as a rheostat position (Meinhardt et al., 2013; Hodges et al., 2018; Wu et al., 2019).

In this dissertation, we aimed to improve bioinformatic and computational predictions of mutational outcomes by locating and characterizing rheostat positions as well as attempting to

establish a method to predict them. We selected a transmembrane protein, Na^+/taurocholate cotransporting polypeptide (NTCP), as a model protein. We chose NTCP for a variety of reasons. First, two bacterial homologs of ASBT, a fellow SLC10A family member of NTCP, had been crystalized in two conformations. Based on these crystal structures, models of NTCP in both an inward- and outward-open facing conformation were created (Hu et al., 2011; Claro da Silva et al., 2013; Zhou et al., 2014). These models allowed for a comparison of structural changes between the two major conformations. Next, NTCP is a transmembrane protein that is important for the enterohepatic circulation of conjugated bile acids (Hagenbuch and Dawson, 2004; Eloranta et al., 2006; Claro da Silva et al., 2013). In addition to bile acids, NTCP transports other endogenous and exogenous compounds like sulfated hormones and the cholesterol-lowering drug rosuvastatin (Claro da Silva et al., 2013). The use of different substrates made it possible to scrutinize substrate-dependent changes in the presence of mutations. Finally, NTCP is naturally polymorphic indicating that mutations in this protein are tolerated and therefore potentially rheostatic (Ho et al., 2004; Choi et al., 2011; Qiu et al., 2017).

In Chapter 3, we addressed the first specific aim by confirming the existence of a rheostat position within NTCP. We examined the site of a polymorphism, position 267, that showed substrate-dependent changes in function (Ho et al., 2004; Choi et al., 2011). There were several outcomes seen with the polymorphism including decreased taurocholate transport, unchanged estrone-3-sulfate uptake, and increased rosuvastatin transport, which led us to hypothesize that position 267 would be a rheostat position. As expected, position 267 demonstrated rheostatic outcomes for all three substrates (Figure 3-1 and 3-3). Interestingly, the rank order of S267 variants changed depending on the substrate. Transport of each substrate by the variants was compared and showed clear correlations between estrone-3-sulfate and rosuvastatin. Further,

there was a weaker correlation between either substrate and taurocholate (Figure 2-4). These results indicate that estrone-3-sulfate and rosuvastatin share similar translocation pathways and likely bind to similar locations within the translocation pore. Thus, we concluded that mutations at position 267 disrupt the translocation pathway and substrate specificity for taurocholate differently than for the other substrates. Further investigation of the transport kinetics of select variants confirmed that both the maximal transport rate (V_{max}) and the variant's substrate affinity (K_m) were altered (Figure 3-5 and Table 3-1). These finding provided insight into the substrate translocation pathways and the effect mutations can have on their specificity. In addition, we demonstrated that there was a slight correlation to changes in the variant surface expression and the simulated structure model energy scores (Figure 3-8). This comparison implied that the energy scores may give insight into the overall stability of NTCP in the presence of mutations and therefore may be a valuable tool for the prediction of protein expression outcomes.

In Chapter 4, we investigated our second specific aim, to determine if protein expression could be correlated with calculate energy scores and whether mutational tolerance could indicate a rheostat position. First, we selected position 271 because its simulated energy scores suggested it would be a rheostat (Figure 4-1). One crucial distinction we had not yet considered when selecting N271 was that a position could be rheostatic for function, expression, neither, or both. Furthermore, when selecting N271 based on the predicted energy scores, those scores should indicate protein stability and thus correlate with the surface expression. Therefore, if the energy scores were able to predict a rheostat then N271 should have been rheostatic for variant surface expression. Unfortunately, there was neither correlation between the simulated energy scores and the surface expression (Figure 4-6), nor was the surface expression rheostatic (Figure 4-3 and

Table 4-2). As a result, we concluded that we cannot use the current simulated energy scores to predict protein expression.

However, the energy scores did accurately predict mutational tolerance. Thus, we did confirm that position 271 was a functional rheostat for taurocholate and rosuvastatin; it was not rheostatic for estrone-3-sulfate transport, however (Figure 4-4 and Table 4-2). In addition, one interesting observation was the significantly higher correlation between the transport of the individual substrates. Further, investigation of select variants yielded few changes in substrate affinity (K_m) compared to wildtype, but rather changes were seen in the maximal transport rate (V_{max}) (Figure 4-8 and Table 4-1). When taken together, these results demonstrate that mutations at position 271 do not alter the substrate specificity confirming that N271 is located away from the substrate binding pocket.

In Chapter 4 we also investigated two locations that we predicted would not result in rheostatic functional outcomes. Up to this point, we had established that two locations in NTCP were both rheostatic. Therefore, we began to question if all amino acid positions in NTCP would be rheostatic. Thus, we selected one evolutionarily conserved position, G102, and one nonconserved position, Y146. Unsurprisingly, numerous G102 variants showed tremendous decreases in function and lack of surface expression (Figure 4-9, 4-10, and 4-11). Conversely, Y146 variants showed little to no changes in expression or function except for the asparagine and serine variants (Figure 4-12, 4-13, and 4-14). These results indicate that most mutations at G102 are deleterious, while mutations at Y146 are well-tolerated. When taken together, these observations confirmed the conservation, or lack thereof, of these two positions. Further investigation of total protein samples from G102 and Y146 surface biotinylation experiments gave insight into whether the observed decreased or diminished surface expression was due to

degradation, improper post-translational modifications, or if the protein was not produced at all. We found that G102 variants were in various states of glycosylation and degradation, Y146S was made but not glycosylated, and Y146N was not detectable at all. These results explain the lack of surface expression and demonstrate that the G102 variants were unstable. More experimentation is needed to understand the underlying mechanism of the Y146S and N variants' results.

There were a few unexpected observations encountered in Chapter 4. First, there was a discrepancy seen between the surface expression and the simulated energy scores, as discussed. We hypothesize that this disagreement may be due to the homology modeling of NTCP based on the bacterial structures. As mentioned, these models were calculated using ASBT bacterial homologs. These homologs have a sequence identity of approximately 25% with NTCP (Hu et al., 2011; Zhou et al., 2014). While the structures were only used as a template and other factors, like the known number of transmembrane domains, influenced the models, there is room for error and improvement. Ideally, NTCP would be crystalized in numerous confirmations, including intermediate structures, to fully understand the changes during substrate translocation.

Next there were unexplained changes in the surface expression of some variants, such as Y146N and Y146S. We predict that there are many factors that influence the stability of NTCP that were not and could not be considered in the simulated energy scores. For instance, post-translational modifications such as N-glycosylation, phosphorylation, acetylation, and more can greatly impact a protein's stability and expression. For example, NTCP needs to be both glycosylated and dephosphorylated to be properly trafficked to the plasma membrane (Anwer and Stieger, 2014; Appelman et al., 2017). Disruption of NTCP glycosylation can lead to endocytosis and lysosomal degradation (Appelman et al., 2017). While the glycosylation sites are located on the extracellular N-terminal end of NTCP, away from the positions we mutated, there could be long

range effects that impact glycosylation. There may also be additional post-translational modifications, such as palmitoylation, that could be necessary for NTCP regulation. A recent study found that S-acylation was crucial for human ASBT stability, phosphorylation, function, and expression. This, along with other post-translational modifications that have not yet been discovered for NTCP, may be playing a role in the results we have observed in this dissertation.

Another unexpected observation from these overall findings was that numerous variants were more efficiently expressed and transported substrates better than the wildtype protein. While this was not always the case, as seen in Chapter 4 with G102, it was very prevalent for all other positions examined (S267, N271, and Y146). We initially thought that nature would have selected amino acids that produced the greatest transport and expression in each location. However, as mentioned, this was not what we observed. One possible explanation is that nature selects amino acids that impact the structure and function of a protein to accomplish a task. Furthermore, disruption of the natural homeostasis of a system by replacing a normal protein with a highly efficient transporter may not be necessary and may even cause toxicity. Thus, it is likely that the most active or efficient protein may not be the most optimal for the organism. Further, having suboptimal proteins allows room for improvement in the future as well as functional differentiation between species.

One of the most valuable tools we utilized in this dissertation was the RheoScale calculator. This tool allowed us to systematically compare and determine the strength and classification of the measured outcomes for all four experimentally characterized positions: S267, N271, G102, and Y146. Unsurprisingly, S267 demonstrated the highest functional rheostatic scores for all three substrates. On the other hand, N271 was a much “weaker” functional rheostat that only reached the rheostat threshold (0.5) for taurocholate and rosuvastatin. In addition, S267 and N271 both

exhibited neutral scores for surface expression demonstrating that mutations at these locations are well accepted and that NTCP can easily adjust its structure to accommodate changes in the amino acids, at least at these two positions. The most surprising outcome was G102 which reached the rheostatic threshold for all four experimental data sets, indicating that G102 is not a strong toggle position as initially thought based on evolutionary conservation. This further indicates that mutations at highly-conserved amino acid positions may show decreased transport and even some diminished expression, but the variants that are expressed still have enough function to be considered rheostatic. Lastly, as hypothesized, Y146 reached the neutral threshold for all data sets analyzed. This confirms that amino acid substitutions in flexible regions like protein loops and nonconserved positions are unlikely to greatly impact the function and expression of the protein. Although, there may be exceptions to this hypothesis such as if the mutation interrupts disulfide bonds or substrate binding.

Another aspect of protein mutation that was outside of the scope of this dissertation but should be considered in the future is protein interactions and their localization. Previous studies conducted in our laboratory confirmed that NTCP interacts with other transporters on the basolateral membrane of hepatocytes, such as OATP1B3 (Zhang et al., 2020). How these interactions occur has yet to be elucidated, but it is likely that some amino acid substitutions might impact these interactions. Given that transporters are simultaneously expressed in the hepatocyte naturally, this can further alter the function and expression of NTCP and therefore impact drug response in humans. Thus, the impact of mutations on protein-protein interactions and their localization would be important to investigate.

When considering all of the data presented in this dissertation, a few key conclusions can be drawn. First, rheostats do exist in the transmembrane protein NTCP. Second, rheostatic outcomes

are more complex than originally imagined. Third, it is currently impossible to predict these locations using the NTCP homology model and simulated energy scores using the Rosetta software suite. Thus, more investigation is needed to be able to predict these locations and their outcomes in the future. And finally, it was confirmed that mutations at “predictable” locations do not necessarily exhibit straightforward outcomes.

Chapter 6

Future Directions

We expected that by the end of this dissertation we would have: 1) determined that rheostat positions exist in transmembrane proteins and 2) gained an understanding of their general characteristics as well as their impact on the NTCP's structure and function. Both of these goals were completed. However, in addition, we had hoped to successfully predict rheostatic outcomes, specifically in terms of surface expression, within NTCP. However, the method we chose to pursue, homology modeling of NTCP based on bacterial homologs followed by structure-based energy simulations, was unable to predict changes in protein expression. Thus, my first proposed future direction would be to improve on the modeling of NTCP to determine if amended structure-based simulations could predict rheostatic outcomes. The bacterial homologs of ASBT (a family member of NTCP) allowed for modeling of NTCP, however, there may still be room to improve the model since the bacterial homologs only showed about 25% of sequence homology to human NTCP (Hu et al., 2011; Zhou et al., 2014). For example, homology modeling is fairly accurate for transmembrane domains but less so for regions outside of the membrane, like for loops (Fiser et al., 2000).

One way to ensure our modeling is accurate as well as improve it, would be to crystalize NTCP. However, this may prove to be a difficult task. Thus far, NTCP along with approximately 95% of other membrane bound proteins, have not been crystalized (Turkova and Zdrazil, 2019). This is mainly due to 1) limitations of *E. coli* that are normally used to overexpress the proteins, 2) difficulties recreating the membrane environment to ensure the native state is captured, and 3)

even ensuring the crystalizing conditions are optimal for the protein of interest (Newstead et al., 2008; Schlegel et al., 2010). Nevertheless, protein engineering and protein stability are being advanced and thus crystallization of membrane proteins may be attainable in the not so distant future (Tate and Schertler, 2009). An alternative widely used method to determine membrane protein structure that could be utilized is cryo-electron microscopy. The recent revolution in techniques has made this method more accessible.

Another future experimental approach would aid in the interpretation of the observed substrate-dependent results and test the hypothesis is correct that NTCP has at least two translocation pathways or binding pockets. In this dissertation, it was demonstrated that there were strong correlations between the transport of estrone-3-sulfate and rosuvastatin by the S267 variants, as well as a lack of correlation between taurocholate and either substrate. It was predicted that the strong correlation implied that estrone-3-sulfate and rosuvastatin may share similar translocation pathways. To test this hypothesis, I suggest the use of computational docking. In theory, I would use the NTCP homology model (or crystalized NTCP if available) along with, e.g., the UCSF DOCK software package (http://dock.compbio.ucsf.edu/DOCK_6/index.htm) to analyze the ligand binding pockets of the three substrates used in this dissertation. Previous docking studies on another hepatic transporter, OCT1, revealed that it has different binding pockets for at least three substrates within its translocation pathway (Boxberger et al., 2018). Thus, multiple docking sites are not outside the realm of possibility for NTCP, and there is an established method to test this hypothesis. Further, once the method is established, it would be possible to examine how mutations impact the docking of substrates, as with the S267 variants.

At this point, when considering the data presented in this dissertation, it is clear that more work is needed to completely understand rheostat positions and for patterns to emerge for future

rheostat predictions. But to see patterns and draw inferences about mutations in the future, more rheostatic positions need to be examined. Thus, I would first select more positions within NTCP while sampling different locations and conservation scores. I would choose additional locations within different areas of the protein such as amino acids in transmembrane domains or extracellular loops, towards or away from the translocation pathway, N- or C-terminal ends and conserved, nonconserved, or of intermediate conservation. Further, it would be beneficial to examine other clinically relevant polymorphic or variant locations within human NTCP such as S199R or R252H (Erlinger, 2015; Vaz et al., 2015; Li et al., 2019). I would also investigate human NTCP sequence data and create a multiple sequence alignment with only human NTCP sequences. I would then use this alignment to determine lesser-known locations where there are relevant mutations. Then I would focus my efforts on those positions after investigating more well-known polymorphisms and mutations.

In addition, it would be advantageous to characterize additional rheostatic positions in numerous other transporters. This would include multi-specific drug transporters like the organic anion transporting polypeptides (OATPs) as well as mono-specific proteins with narrow substrate specificities such as the Na^{+}/Ca^{2+} exchanger, for example. Examination of rheostat locations within additional transport proteins would allow us to determine if mutations in different proteins result in similar rheostatic outcomes thus furthering our understanding of rheostat positions outside of NTCP.

Next, once rheostat positions are fully understood in NTCP an individual protein, I would like to explore the expression and function of NTCP variants when the transport is co-expressed with other hepatic transporters. Previous studies in our laboratory have shown that OATP1B3 interacts with NTCP as well as other transporters (Zhang et al., 2020). Further, it is known that

NTCP forms homodimers (Bijsmans et al., 2012). These studies suggest that there is a complex interplay between NTCP and itself and other hepatic transporters. Thus, it is likely that NTCP variants would also impact these interactions. It would also be interesting to determine if the rheostatic outcomes are impacted by the co-expression of NTCP with other transporters. Further, these studies would enable us to elucidate which amino acids within NTCP are necessary for homo- and heterodimerization.

We have also hypothesized, and have preliminary data suggesting, that the previously observed interactions between hepatic transporters is due to localization to certain membrane microdomains, such as lipid rafts. In addition, studies have suggested NTCP in particular can be found in lipid rafts and that if the hepatic microenvironment is altered, such as in the case of depleted cholesterol, there is an increase in NTCP transport (Molina et al., 2008). Thus, I would predict that mutations that impact the surface expression of NTCP would also further complicate NTCP's localization to the plasma membrane and as a result impact its interactions with other transporters. However, we would first need to completely characterize the localization of NTCP as well as understand the impact of certain disease states, like obesity, hepatitis, steatosis, etc. on its localization. Then, once the microenvironment and the factors that can impact NTCP expression and function are well characterized, we could investigate how mutations and rheostat positions further complicate NTCP's localization.

Another aspect we were unable to consider in the scope of this dissertation was the effect of mutations at rheostat locations on the infectivity of the Hepatitis B/D virus. It is known that NTCP acts as the receptor for this virus to infect hepatocytes (Yan et al., 2012). Further, it has been shown that if NTCP glycosylation is disrupted, HBV and HDV infection also decreases (Appelman et al., 2017). These results suggest that certain variants of NTCP, like Y146S, could

be resistant to hepatitis viral infection. Thus, I hypothesize that NTCP variants that demonstrated decreased or diminished surface expression, such as Y146S and N, are likely to have reduced viral infectivity. But in order to confirm the hypothesis that decreased variant surface expression could impact hepatitis infection, it would need to be further investigated.

Thus far, the studies we have completed and that I have suggested would be done *in vitro*. This is mainly because of the ease and flexibility of *in vitro* models when it comes to creating numerous variants. However, cell-based studies can only be taken so far. It is impossible to recreate an entire functioning system using only cell-based assays. Eventually, using an animal model to examine the effects of rheostatic positions in a whole organism would allow for a more complete understanding of their impact. Thus, I would select certain variants that showed unique and extreme functional and expression outcomes, such as selected G102 variants, and express them in a mouse model to determine their impact in a living system.

I anticipate the future directions I have proposed would allow us to fully comprehend the complexity of NTCP and rheostat positions and resolve questions we have been unable to investigate thus far. These future directions would allow us to resolve 1) the structure of NTCP, 2) the translocation of substrates through NTCP in the presence and absence of mutations, 3) the effect the microenvironment and disease states have on mutational outcomes, and vice versa, and 4) the impact of a rheostatic positions both *in vitro* and *in vivo*. Once these questions are addressed, we can look forward to the future of predictive pharmacogenomics. However, once we can accurately predict mutational outcomes in every protein, enzyme, transcription factor, etc., in a living system, we will still need to be able to efficiently apply this knowledge in real time in the clinic. Ideally, all this information would be culminated into a program of sorts that could be used in the clinic when a physician is considering prescribing a medication. This

program would act as an artificial human where a physician could input the patient's medical history, demographic information, genomic sequence, and even their medical concern. As a result, the program would determine the best course of treatment while accounting for all risk factors, the mechanism of action of the medication, and even possible adverse effects of the treatment. This program would also be beneficial to the pharmaceutical industry because it could be used in preclinical studies to decrease costs, eliminate the need for animal testing, and accurately predict human responses. While this may seem like a tremendous feat, I believe research, time, money, and extraordinary minds will make this hypothetical program a reality and ultimately improve approaches to medical care.

References

Adamski CJ and Palzkill T (2017) Systematic substitutions at BLIP position 50 result in changes in binding specificity for class A beta-lactamases. *BMC Biochem* **18:**2.

Alberts B, Johnson A, Lewis J, Raff M, Roberts K, and Walter P (2002) *Molecular Biology of the Cell*. Garland Science, New York.

Anwer MS and Stieger B (2014) Sodium-dependent bile salt transporters of the SLC10A transporter family: more than solute transporters. *Pflugers Arch* **466:**77-89.

Appelman MD, Chakraborty A, Protzer U, McKeating JA, and van de Graaf SF (2017) N-Glycosylation of the Na+-Taurocholate Cotransporting Polypeptide (NTCP) Determines Its Trafficking and Stability and Is Required for Hepatitis B Virus Infection. *PLoS One* **12:**e0170419.

Ashkenazy H, Abadi S, Martz E, Chay O, Mayrose I, Pupko T, and Ben-Tal N (2016) ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res* **44:**W344-350.

Berg JM, Tymoczko JL, and Stryer L (2002) *Biochemistry*. W H Freeman, New York.

Bijsmans IT, Bouwmeester RA, Geyer J, Faber KN, and van de Graaf SF (2012) Homo- and hetero-dimeric architecture of the human liver Na(+)-dependent taurocholate co-transporting protein. *Biochem J* **441:**1007-1015.

Binh MT, Hoan NX, Van Tong H, Sy BT, Trung NT, Bock CT, Toan NL, Song LH, Bang MH, Meyer CG, Kremsner PG, and Velavan TP (2019) NTCP S267F variant associates with decreased susceptibility to HBV and HDV infection and decelerated progression of related liver diseases. *Int J Infect Dis* **80:**147-152.

Bowman CM, Ma F, Mao J, and Chen Y (2020) Examination of Physiologically-Based Pharmacokinetic Models of Rosuvastatin. *CPT Pharmacometrics Syst Pharmacol*.

Boxberger KH, Hagenbuch B, and Lampe JN (2018) Ligand-dependent modulation of hOCT1 transport reveals discrete ligand binding sites within the substrate translocation channel. *Biochem Pharmacol* **156:**371-384.

Brittain HK, Scott R, and Thomas E (2017) The rise of the genome and personalised medicine. *Clin Med (Lond)* **17:**545-551.

Campitelli P, Swint-Kruse L, and Ozkan SB (2021) Substitutions at Nonconserved Rheostat Positions Modulate Function by Rewiring Long-Range, Dynamic Interactions. *Mol Biol Evol* **38:**201-214.

Chen R, Deng M, Rauf YM, Lin GZ, Qiu JW, Zhu SY, Xiao XM, and Song YZ (2019) Intrahepatic Cholestasis of Pregnancy as a Clinical Manifestation of Sodium-Taurocholate Cotransporting Polypeptide Deficiency. *Tohoku J Exp Med* **248:**57-61.

Chiang JY (2009) Bile acids: regulation of synthesis. *J Lipid Res* **50:**1955-1966.

Chiang JY (2013) Bile acid metabolism and signaling. *Compr Physiol* **3:**1191-1212.

Choi MK, Shin HJ, Choi YL, Deng JW, Shin JG, and Song IS (2011) Differential effect of genetic variants of Na(+)-taurocholate co-transporting polypeptide (NTCP) and organic anion-transporting polypeptide 1B1 (OATP1B1) on the uptake of HMG-CoA reductase inhibitors. *Xenobiotica* **41:**24-34.

Clamp M, Fry B, Kamal M, Xie X, Cuff J, Lin MF, Kellis M, Lindblad-Toh K, and Lander ES (2007) Distinguishing protein-coding and noncoding genes in the human genome. *Proc Natl Acad Sci U S A* **104:**19428-19433.

Clancy S (2008) Genetic Mutation, in: *Nucleic Acid Structure and Function* (Moss B ed), pp 187, Nature Education.

Claro da Silva T, Polli JE, and Swaan PW (2013) The solute carrier family 10 (SLC10): beyond bile acid transport. *Mol Aspects Med* **34:**252-269.

Collins FS and Fink L (1995) The Human Genome Project. *Alcohol Health Res World* **19:**190-195.

Creighton TE (1993) *Proteins: Structures and molecular properties*. Macmillan.

de Waart DR, Vlaming ML, Kunne C, Schinkel AH, and Oude Elferink RP (2009) Complex pharmacokinetic behavior of ezetimibe depends on abcc2, abcc3, and abcg2. *Drug Metab Dispos* **37:**1698-1702.

Deng M, Mao M, Guo L, Chen FP, Wen WR, and Song YZ (2016) Clinical and molecular study of a pediatric patient with sodium taurocholate cotransporting polypeptide deficiency. *Exp Ther Med* **12:**3294-3300.

Denson LA, Sturm E, Echevarria W, Zimmerman TL, Makishima M, Mangelsdorf DJ, and Karpen SJ (2001) The orphan nuclear receptor, shp, mediates bile acid-induced inhibition of the rat bile acid transporter, ntcp. *Gastroenterology* **121:**140-147.

Di Ciaula A, Garruti G, Lunardi Baccetto R, Molina-Molina E, Bonfrate L, Wang DQ, and Portincasa P (2017) Bile Acid Physiology. *Ann Hepatol* **16:**s4-s14.

Dong C, Zhang BP, Wang H, Xu H, Zhang C, Cai ZS, Wang DW, Shu SN, Huang ZH, and Luo XP (2019) Clinical and histopathologic features of sodium taurocholate cotransporting polypeptide deficiency in pediatric patients. *Medicine (Baltimore)* **98:**e17305.

Doring B, Lutteke T, Geyer J, and Petzinger E (2012) The SLC10 carrier family: transport functions and molecular structure. *Curr Top Membr* **70:**105-168.

Eloranta JJ, Jung D, and Kullak-Ublick GA (2006) The human Na+-taurocholate cotransporting polypeptide gene is activated by glucocorticoid receptor and peroxisome proliferator-activated receptor-gamma coactivator-1alpha, and suppressed by bile acids via a small heterodimer partner-dependent mechanism. *Mol Endocrinol* **20:**65-79.

Erlinger S (2015) NTCP deficiency: a new inherited disease of bile acid transport. *Clin Res Hepatol Gastroenterol* **39:**7-8.

Fenton AW, Page BM, Spellman-Kruse A, Hagenbuch B, and Swint-Kruse L (2020) Rheostat positions: A new classification of protein positions relevant to pharmacogenomics. *Med Chem Res* **29:**1133-1146.

Fiser A, Do RK, and Sali A (2000) Modeling of loops in protein structures. *Protein Sci* **9:**1753-1773.

Fowler DM and Fields S (2014) Deep mutational scanning: a new style of protein science. *Nat Methods* **11:**801-807.

Gray VE, Kukurba KR, and Kumar S (2012) Performance of computational tools in evaluating the functional impact of laboratory-induced amino acid mutations. *Bioinformatics* **28:**2093-2096.

Hagenbuch B (1997) Molecular properties of hepatic uptake systems for bile acids and organic anions. *J Membr Biol* **160:**1-8.

Hagenbuch B and Dawson P (2004) The sodium bile salt cotransport family SLC10. *Pflugers Arch* **447:**566-570.

Hagenbuch B and Meier PJ (1994) Molecular cloning, chromosomal localization, and functional characterization of a human liver Na+/bile acid cotransporter. *J Clin Invest* **93:**1326-1331.

Hallen S, Mareninova O, Branden M, and Sachs G (2002) Organization of the membrane domain of the human liver sodium/bile acid cotransporter. *Biochemistry* **41:**7253-7266.

Ho RH, Leake BF, Roberts RL, Lee W, and Kim RB (2004) Ethnicity-dependent polymorphism in Na+-taurocholate cotransporting polypeptide (SLC10A1) reveals a domain critical for bile acid substrate recognition. *J Biol Chem* **279:**7213-7222.

Ho RH, Tirona RG, Leake BF, Glaeser H, Lee W, Lemke CJ, Wang Y, and Kim RB (2006) Drug and bile acid transporters in rosuvastatin hepatic uptake: function, expression, and pharmacogenetics. *Gastroenterology* **130:**1793-1806.

Hodges AM, Fenton AW, Dougherty LL, Overholt AC, and Swint-Kruse L (2018) RheoScale: A tool to aggregate and quantify experimentally determined substitution outcomes for multiple variants at individual protein positions. *Hum Mutat* **39:**1814-1826.

Hu NJ, Iwata S, Cameron AD, and Drew D (2011) Crystal structure of a bacterial homologue of the bile acid sodium symporter ASBT. *Nature* **478:**408-411.

Huang L, Wang Y, and Grimm S (2006) ATP-dependent transport of rosuvastatin in membrane vesicles expressing breast cancer resistance protein. *Drug Metab Dispos* **34:**738-742.

Kalliokoski A and Niemi M (2009) Impact of OATP transporters on pharmacokinetics. *Br J Pharmacol* **158:**693-705.

Kalra A, Yetiskul E, Wehrle CJ, and Tuma F (2020) Physiology, Liver, in: *StatPearls*, Treasure Island (FL).

Kosoglou T, Statkevich P, Johnson-Levonas AO, Paolini JF, Bergman AJ, and Alton KB (2005) Ezetimibe: a review of its metabolism, pharmacokinetics and drug interactions. *Clin Pharmacokinet* **44:**467-494.

Kuhlman B and Baker D (2000) Native protein sequences are close to optimal for their structures. *Proc Natl Acad Sci U S A* **97:**10383-10388.

Landau M, Mayrose I, Rosenberg Y, Glaser F, Martz E, Pupko T, and Ben-Tal N (2005) ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res* **33:**W299-302.

Levy E, Spahis S, Sinnett D, Peretti N, Maupas-Schwalm F, Delvin E, Lambert M, and Lavoie MA (2007) Intestinal cholesterol transport proteins: an update and beyond. *Curr Opin Lipidol* **18:**310-318.

Li H, Deng M, Guo L, Qiu JW, Lin GZ, Long XL, Xiao XM, and Song YZ (2019) Clinical and molecular characterization of four patients with NTCP deficiency from two unrelated families harboring the novel SLC10A1 variant c.595A>C (p.Ser199Arg). *Mol Med Rep* **20:**4915-4924.

Li T and Chiang JY (2014) Bile acid signaling in metabolic disease and drug therapy. *Pharmacol Rev* **66:**948-983.

Liu R, Chen C, Xia X, Liao Q, Wang Q, Newcombe PJ, Xu S, Chen M, Ding Y, Li X, Liao Z, Li F, Du M, Huang H, Dong R, Deng W, Wang Y, Zeng B, Pan Q, Jiang D, Zeng H, Sham P, Cao Y, Maxwell PH, Gao ZL, Peng L, and Wang Y (2017) Homozygous p.Ser267Phe in SLC10A1 is associated with a new type of hypercholanemia and implications for personalized medicine. *Sci Rep* **7:**9214.

Malik MY, Jaiswal S, Sharma A, Shukla M, and Lal J (2016) Role of enterohepatic recirculation in drug disposition: cooperation and complications. *Drug Metab Rev* **48:**281-327.

Mann M and Jensen ON (2003) Proteomic analysis of post-translational modifications. *Nat Biotechnol* **21:**255-261.

Martin TA, Wu T, Tang Q, Dougherty LL, Parente DJ, Swint-Kruse L, and Fenton AW (2020) Identification of biochemically neutral positions in liver pyruvate kinase. *Proteins* **88:**1340-1350.

Mazin PV, Gelfand MS, Mironov AA, Rakhmaninova AB, Rubinov AR, Russell RB, and Kalinina OV (2010) An automated stochastic approach to the identification of the protein specificity determinants and functional subfamilies. *Algorithms Mol Biol* **5:**29.

McTaggart F, Buckett L, Davidson R, Holdgate G, McCormick A, Schneck D, Smith G, and Warwick M (2001) Preclinical and clinical pharmacology of Rosuvastatin, a new 3-hydroxy-3-methylglutaryl coenzyme A reductase inhibitor. *Am J Cardiol* **87:**28B-32B.

Meier PJ and Stieger B (2002) Bile salt transporters. *Annu Rev Physiol* **64:**635-661.

Meinhardt S, Manley MW, Jr., Parente DJ, and Swint-Kruse L (2013) Rheostats and toggle switches for modulating protein function. *PLoS One* **8:**e83502.

Miller M, Bromberg Y, and Swint-Kruse L (2017) Computational predictors fail to identify amino acid substitution effects at rheostat positions. *Sci Rep* **7:**41329.

Miller M, Vitale D, Kahn PC, Rost B, and Bromberg Y (2019) funtrp: identifying protein positions for variation driven functional tuning. *Nucleic Acids Res* **47:**e142.

Modi T and Ozkan SB (2018) Mutations Utilize Dynamic Allostery to Confer Resistance in TEM-1 beta-lactamase. *Int J Mol Sci* **19**.

Molina H, Azocar L, Ananthanarayanan M, Arrese M, and Miquel JF (2008) Localization of the Sodium-Taurocholate cotransporting polypeptide in membrane rafts and modulation of its activity by cholesterol in vitro. *Biochim Biophys Acta* **1778:**1283-1291.

Newstead S, Ferrandon S, and Iwata S (2008) Rationalizing alpha-helical membrane protein crystallization. *Protein Sci* **17:**466-472.

Nezasa K, Higaki K, Matsumura T, Inazawa K, Hasegawa H, Nakano M, and Koike M (2002) Liver-specific distribution of rosuvastatin in rats: comparison with pravastatin and simvastatin. *Drug Metab Dispos* **30:**1158-1163.

Ohnishi S, Hays A, and Hagenbuch B (2014) Cysteine scanning mutagenesis of transmembrane domain 10 in organic anion transporting polypeptide 1B1. *Biochemistry* **53:**2261-2270.

Pan W, Song IS, Shin HJ, Kim MH, Choi YL, Lim SJ, Kim WY, Lee SS, and Shin JG (2011) Genetic polymorphisms in Na+-taurocholate co-transporting polypeptide (NTCP) and ileal apical sodium-dependent bile acid transporter (ASBT) and ethnic comparisons of functional variants of NTCP among Asian populations. *Xenobiotica* **41:**501-510.

Patrick JE, Kosoglou T, Stauber KL, Alton KB, Maxwell SE, Zhu Y, Statkevich P, Iannucci R, Chowdhury S, Affrime M, and Cayen MN (2002) Disposition of the selective cholesterol absorption inhibitor ezetimibe in healthy male subjects. *Drug Metab Dispos* **30:**430-437.

Phillips KA, Deverka PA, Hooker GW, and Douglas MP (2018) Genetic Test Availability And Spending: Where Are We Now? Where Are We Going? *Health Aff (Millwood)* **37:**710-716.

Pinak M (2006) Enzymatic recognition of radiation-produced oxidative DNA lesion. Molecular dynamics approach, in: *Modern Methods for Theoretical Physical Chemistry of Biopolymers* (Starikov EB, Lewis JP, and Tanaka S eds), pp 191-210, Elsevier Science.

Prescott LF (1980) Kinetics and metabolism of paracetamol and phenacetin. *Br J Clin Pharmacol* **10 Suppl 2:**291S-298S.

Procko E (2020) The sequence of human ACE2 is suboptimal for binding the S spike protein of SARS coronavirus 2. *bioRxiv*.

Prueksaritanont T, Tang C, Qiu Y, Mu L, Subramanian R, and Lin JH (2002) Effects of fibrates on metabolism of statins in human hepatocytes. *Drug Metab Dispos* **30:**1280-1287.

Qiu JW, Deng M, Cheng Y, Atif RM, Lin WX, Guo L, Li H, and Song YZ (2017) Sodium taurocholate cotransporting polypeptide (NTCP) deficiency: Identification of a novel SLC10A1 mutation in two unrelated infants presenting with neonatal indirect hyperbilirubinemia and remarkable hypercholanemia. *Oncotarget* **8:**106598-106607.

Radivojac P Clark WT Oron TR Schnoes AM Wittkop T Sokolov A Graim K Funk C Verspoor K Ben-Hur A Pandey G Yunes JM Talwalkar AS Repo S Souza ML Piovesan D Casadio R Wang Z Cheng J Fang H Gough J Koskinen P Toronen P Nokso-Koivisto J Holm L Cozzetto D Buchan DW Bryson K Jones DT Limaye B

Inamdar H Datta A Manjari SK Joshi R Chitale M Kihara D Lisewski AM Erdin S Venner E Lichtarge O Rentzsch R Yang H Romero AE Bhat P Paccanaro A Hamp T Kassner R Seemayer S Vicedo E Schaefer C Achten D Auer F Boehm A Braun T Hecht M Heron M Honigschmid P Hopf TA Kaufmann S Kiening M Krompass D Landerer C Mahlich Y Roos M Bjorne J Salakoski T Wong A Shatkay H Gatzmann F Sommer I Wass MN Sternberg MJ Skunca N Supek F Bosnjak M Panov P Dzeroski S Smuc T Kourmpetis YA van Dijk AD ter Braak CJ Zhou Y Gong Q Dong X Tian W Falda M Fontana P Lavezzo E Di Camillo B Toppo S Lan L Djuric N Guo Y Vucetic S Bairoch A Linial M Babbitt PC Brenner SE Orengo C Rost B Mooney SD and Friedberg I (2013) A large-scale evaluation of computational protein function prediction. *Nat Methods* **10:**221-227.

Roden DM, McLeod HL, Relling MV, Williams MS, Mensah GA, Peterson JF, and Van Driest SL (2019) Pharmacogenomics. *Lancet* **394:**521-532.

Roscoe BP, Thayer KM, Zeldovich KB, Fushman D, and Bolon DN (2013) Analyses of the effects of all ubiquitin point mutants on yeast growth rate. *J Mol Biol* **425:**1363-1377.

Sankar PL and Parker LS (2017) The Precision Medicine Initiative's All of Us Research Program: an agenda for research on its ethical, legal, and social issues. *Genet Med* **19:**743-750.

Schirris TJ, Ritschel T, Bilos A, Smeitink JA, and Russel FG (2015) Statin Lactonization by Uridine 5'-Diphospho-glucuronosyltransferases (UGTs). *Mol Pharm* **12:**4048-4055.

Schlegel S, Klepsch M, Gialama D, Wickstrom D, Slotboom DJ, and de Gier JW (2010) Revolutionizing membrane protein overexpression in bacteria. *Microb Biotechnol* **3:**403-411.

Schonhoff CM, Ramasamy U, and Anwer MS (2011) Nitric oxide-mediated inhibition of taurocholate uptake involves S-nitrosylation of NTCP. *Am J Physiol Gastrointest Liver Physiol* **300:**G364-370.

Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, and Serrano L (2005) The FoldX web server: an online force field. *Nucleic Acids Res* **33:**W382-388.

Soulele K and Karalis V (2019) On the population pharmacokinetics and the enterohepatic recirculation of total ezetimibe. *Xenobiotica* **49:**446-456.

Stieger B (2010) Role of the bile salt export pump, BSEP, in acquired forms of cholestasis. *Drug Metab Rev* **42:**437-445.

Stieger B (2011) The role of the sodium-taurocholate cotransporting polypeptide (NTCP) and of the bile salt export pump (BSEP) in physiology and pathophysiology of bile formation. *Handb Exp Pharmacol***:**205-259.

Stieger B, Meier Y, and Meier PJ (2007) The bile salt export pump. *Pflugers Arch* **453:**611-620.

Tate CG and Schertler GF (2009) Engineering G protein-coupled receptors to facilitate their structure determination. *Curr Opin Struct Biol* **19:**386-395.

Tungtur S, Schwingen KM, Riepe JJ, Weeramange CJ, and Swint-Kruse L (2019) Homolog comparisons further reconcile in vitro and in vivo correlations of protein activities by revealing over-looked physiological factors. *Protein Sci* **28:**1806-1818.

Turkova A and Zdrazil B (2019) Current Advances in Studying Clinically Relevant Transporters of the Solute Carrier (SLC) Family by Connecting Computational Modeling and Data Science. *Comput Struct Biotechnol J* **17:**390-405.

Uversky V (2013) Posttranslational Modification, in: *Brenner's Encyclopedia of Genetics* (Maloy S and Hughes K eds), pp 425-430.

Vaz FM, Paulusma CC, Huidekoper H, de Ru M, Lim C, Koster J, Ho-Mok K, Bootsma AH, Groen AK, Schaap FG, Oude Elferink RP, Waterham HR, and Wanders RJ (2015) Sodium taurocholate cotransporting polypeptide (SLC10A1) deficiency: conjugated hypercholanemia without a clear clinical phenotype. *Hepatology* **61:**260-267.

Vogenberg FR, Isaacson Barash C, and Pursel M (2010) Personalized medicine: part 1: evolution and development into theranostics. *P T* **35:**560-576.

Wolters H, Elzinga BM, Baller JF, Boverhof R, Schwarz M, Stieger B, Verkade HJ, and Kuipers F (2002) Effects of bile salt flux variations on the expression of hepatic bile salt transporters in vivo in mice. *J Hepatol* **37:**556-563.

Wu T, Swint-Kruse L, and Fenton AW (2019) Functional tunability from a distance: Rheostat positions influence allosteric coupling between two distant binding sites. *Sci Rep* **9:**16957.

Yan H, Zhong G, Xu G, He W, Jing Z, Gao Z, Huang Y, Qi Y, Peng B, Wang H, Fu L, Song M, Chen P, Gao W, Ren B, Sun Y, Cai T, Feng X, Sui J, and Li W (2012) Sodium taurocholate cotransporting polypeptide is a functional receptor for human hepatitis B and D virus. *Elife* **1:**e00049.

Yates CM and Sternberg MJ (2013) The effects of non-synonymous single nucleotide polymorphisms (nsSNPs) on protein-protein interactions. *J Mol Biol* **425:**3949-3963.

Ye K, Vriend G, and AP IJ (2008) Tracing evolutionary pressure. *Bioinformatics* **24:**908-915.

Zhang Y, Ruggiero M, and Hagenbuch B (2020) OATP1B3 Expression and Function is Modulated by Coexpression with OCT1, OATP1B1, and NTCP. *Drug Metab Dispos* **48:**622-630.

Zhou X, Levin EJ, Pan Y, McCoy JG, Sharma R, Kloss B, Bruni R, Quick M, and Zhou M (2014) Structural basis of the alternating-access mechanism in a bile acid transporter. *Nature* **505:**569-573.

Appendices

I. Appendix A: Supporting Tables and Figures for Chapter 3

Supporting Table 3-1. Statistical analysis of substrate uptake by wildtype NTCP and S267 variants using multiple comparisons

S267 variants	Adjusted P values Taurocholate transport	Adjusted P values Taurocholate transport	Adjusted P values Taurocholate transport
WT vs. A	0.9878	0.1837	0.0635
WT vs. C	<0.0001	<0.0001	0.0001
WT vs. D	0.0796	0.9999	0.9996
WT vs. E	<0.0001	0.0003	<0.0001
WT vs. F	<0.0001	0.0002	<0.0001
WT vs. G	0.1190	0.1423	0.0141
WT vs. H	0.9997	0.0899	<0.0001
WT vs. I	0.0016	0.0110	<0.0001
WT vs. K	<0.0001	0.0008	0.0002
WT vs. L	<0.0001	0.0177	0.0323
WT vs. M	0.0001	0.0307	<0.0001
WT vs. N	<0.0001	0.9993	0.9997
WT vs. P	<0.0001	<0.0001	<0.0001
WT vs. Q	0.2623	0.3553	0.0983
WT vs. R	<0.0001	0.0003	<0.0001
WT vs. T	0.9997	0.1270	0.0004

WT vs. V	0.9996	0.1475	0.0009
WT vs. W	0.0005	<0.0001	<0.0001
WT vs. Y	0.0047	<0.0001	<0.0001

Adjusted P values for each comparison were calculated using Dunnett's multiple comparisons test in GraphPad Prism version 8. Significance was set at 0.05 and significant differences are indicated in bold.

Supporting Table 3-2. Statistical analysis of initial substrate uptake normalized for surface expression multiple comparisons

S267 variants	Adjusted P values Taurocholate transport	Adjusted P values Taurocholate transport	Adjusted P values Taurocholate transport
WT vs. A	0.9993	0.4934	0.1990
WT vs. C	0.0062	<0.0001	<0.0001
WT vs. D	0.0284	0.9991	0.9943
WT vs. E	0.1533	0.0744	0.0493
WT vs. F	0.0002	<0.0001	<0.0001
WT vs. G	0.7724	0.2669	0.0337
WT vs. H	0.9832	0.9949	0.0002
WT vs. I	0.7712	0.4995	0.0011
WT vs. K	0.0021	0.1286	0.0798
WT vs. L	0.0006	0.1014	0.1104
WT vs. M	0.0017	0.0725	0.0004
WT vs. N	0.0004	0.9996	0.9998
WT vs. P	<0.0001	0.0030	0.0002
WT vs. Q	<0.0001	<0.0001	<0.0001
WT vs. R	0.0001	0.0128	0.0028
WT vs. T	0.9879	0.1712	0.0015
WT vs. V	0.9999	0.4063	0.0070
WT vs. W	0.0008	0.3074	0.0006
WT vs. Y	0.0054	0.0150	<0.0001

Adjusted P values for each comparison were calculated using Dunnett's multiple comparisons test in GraphPad Prism version 8. Significance was set at 0.05 and significant differences are indicated in bold.

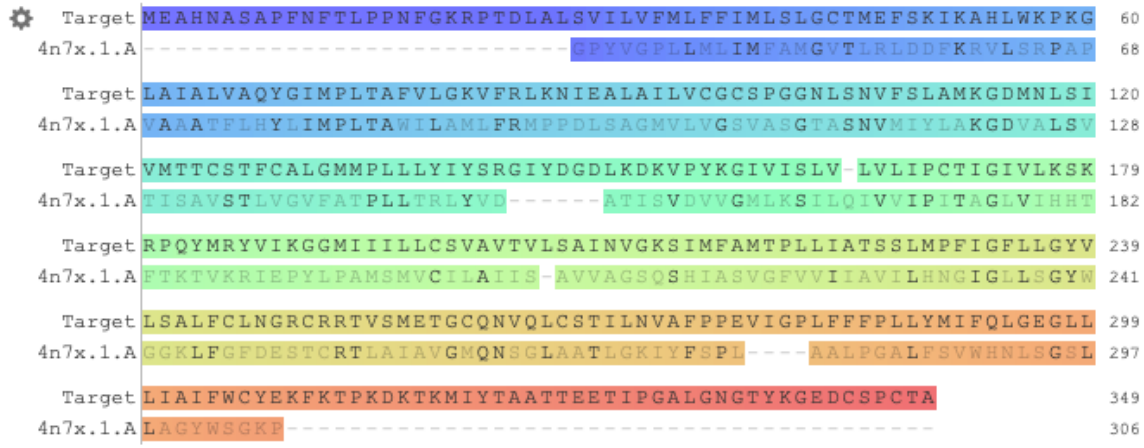
Supporting Table 3-3. Pearson and Spearman coefficients calculated for various correlation studies

	Pearson		Spearman	
	Corr. coeff.	p-value	Corr. coeff.	p-value
Taurocholate v Estrone-3-Sulfate (excluding outliers)	0.9321	**<0.0001**	0.8265	**0.0002**
Taurocholate v Rosuvastatin (excluding outliers)	0.9143	**<0.0001**	0.7000	**0.0048**
Estrone-3-Sulfate v Rosuvastatin (excluding outliers)	0.9509	**<0.0001**	0.8702	**<0.0001**
Taurocholate v Rosetta Inward	0.0537	0.8273	0.1561	0.5233
Estrone-3-Sulfate v Rosetta Inward	-0.1361	0.5785	-0.1193	0.6266
Rosuvastatin v Rosetta Inward	-0.1510	0.5372	0.0754	0.7589
Taurocholate v Rosetta Outward	-0.1091	0.6565	0.2825	0.2413
Estrone-3-Sulfate v Rosetta Outward	0.0652	0.7908	0.1789	0.4636
Rosuvastatin v Rosetta Outward	0.0662	0.7877	0.2491	0.3037
Taurocholate v Rosetta Inward Minus Outward	0.1532	0.5313	-0.0737	0.7643
Estrone-3-Sulfate v Rosetta Inward Minus Outward	-0.2095	0.3892	-0.3018	0.2093
Rosuvastatin v Rosetta Inward Minus Outward	-0.2273	0.3494	-0.1737	0.4770
Surface Expression v Rosetta Inward	-0.3237	0.1765	-0.2466	0.3088
Surface Expression v Rosetta Outward	0.3183	0.1842	0.3010	0.2105
Surface Expression v Rosetta Inward Minus Outward	-0.6364	**0.0034**	-0.5186	**0.0229**
Surface Expression v FoldX Inward Minus Outward	-0.4446	0.0968	-0.3861	0.1551
Rosetta (Inward-Outward) v FoldX (Inward-Outward)	0.5787	**0.0238**	0.5714	**0.0286**

Correlation scores from comparison figures (Figures 3-4, 3-8, and Supporting Figures 3-3, 3-4, 3-5) were calculated in GraphPad Prism 8 using either the Pearson correlation or the nonparametric Spearman correlation. Significance was set at 0.05 and significant differences are indicated in bold.

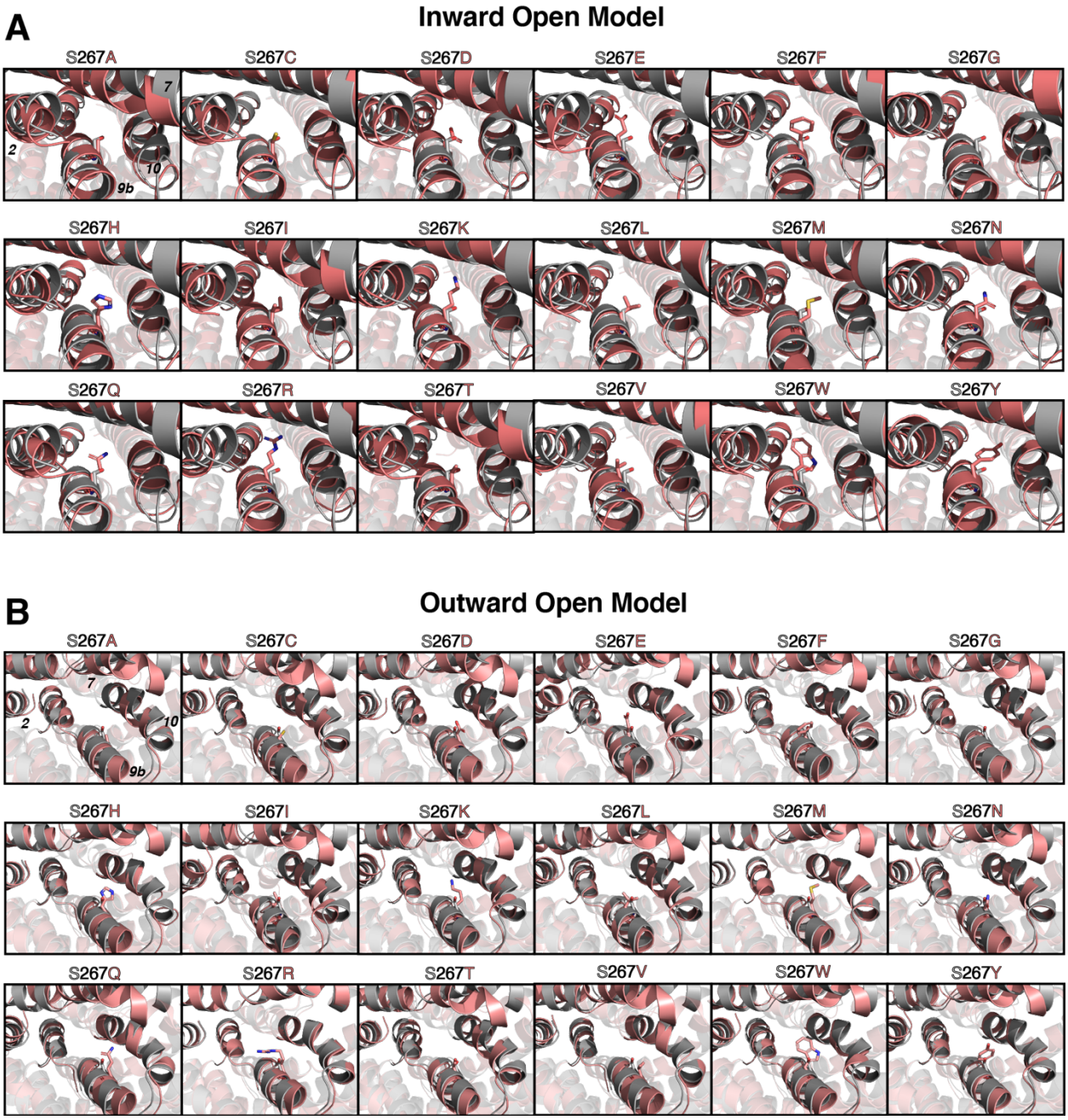
A

B



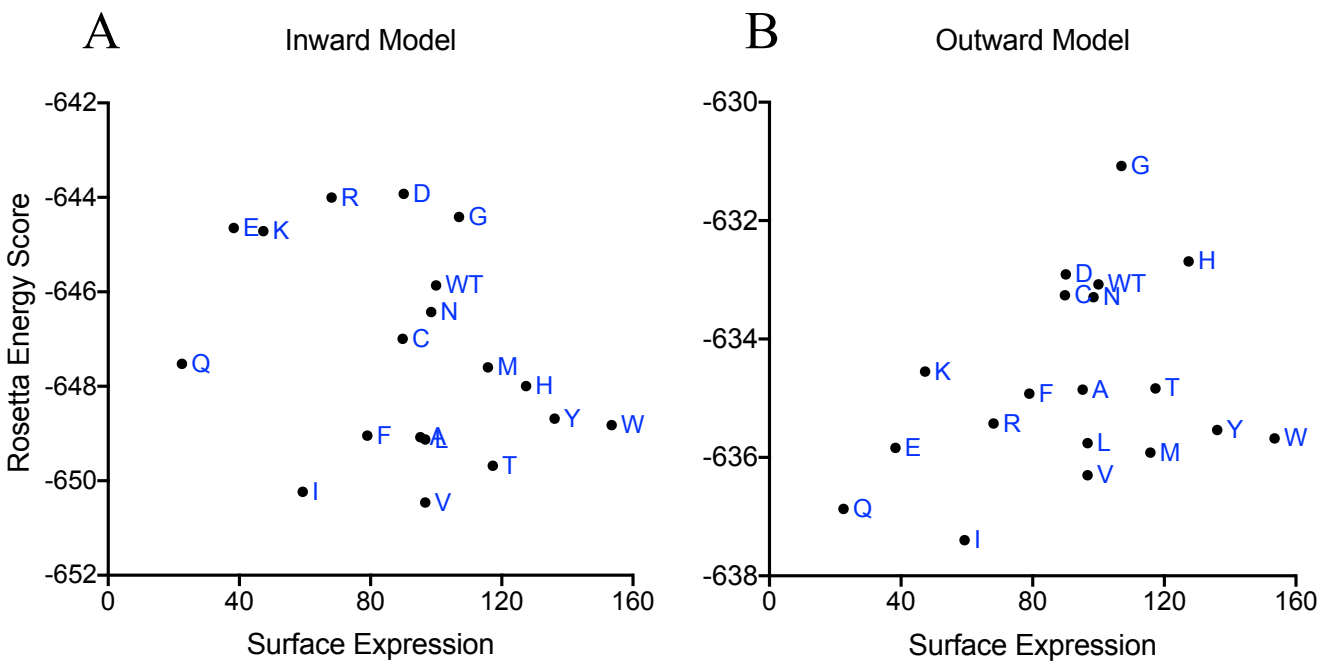
Supporting Figure 3-1. Pairwise sequence alignments of human NTCP against the templates used for comparative modeling.

Top: Alignment to *Yersinia frederiksenii* ASBT ($ASBT_{Yf}$, 26% sequence identity to NTCP, PDB ID 3zuy), used to model the outward-open conformation of human NTCP. *Bottom:* Alignment to *Neisseria meningitidis* ASBT ($ASBT_{Nm}$, 25% sequence identity to NTCP, PDB ID 4n7x), used to model the inward-open conformation of human NTCP. *This figure was completed by collaborators*



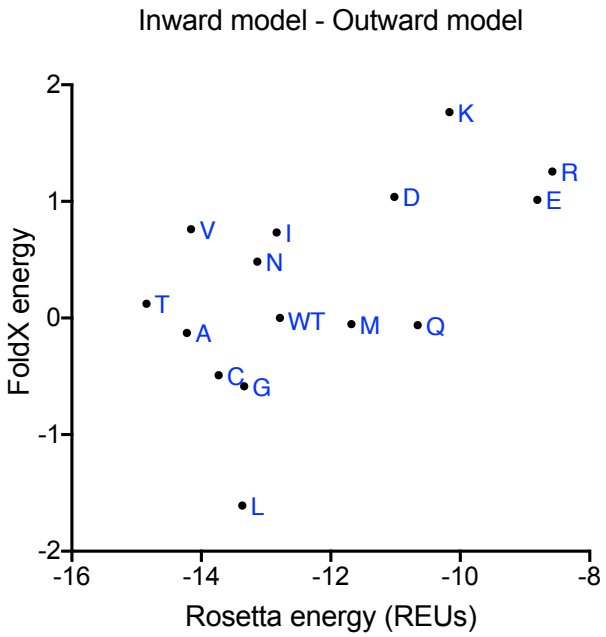
Supporting Figure 3-2. Sequence variation at S267 position leads to alternative local packing.

Sequence variation at S267 position leads to alternative local packing. Structural details are shown from WT in gray and 18 mutant structures in salmon color (proline was excluded) at the 267 position. *A.* Models in the inward-open conformation. *B.* Models in the outward-open conformation. *This figure was completed by collaborators*



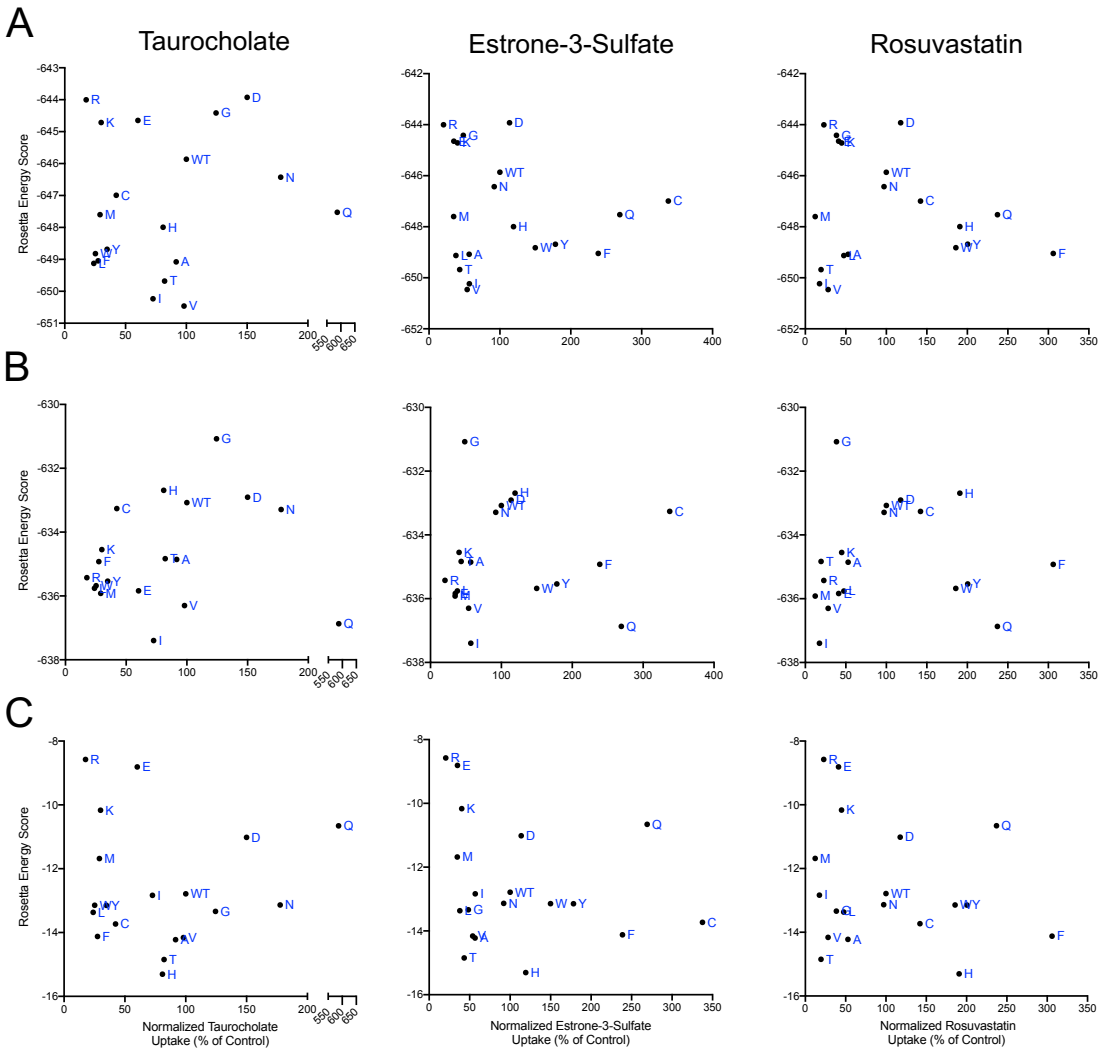
Supporting Figure 3-3. Correlation of Rosetta energies and quantification of surface expression levels.

Correlation of Rosetta energies and quantification of surface expression levels. Energies calculated using Rosetta are plotted against the percent surface expression (Figure 3-2B) of each variant. Individual points are labeled with letters to indicate their amino acid replacements; wildtype is indicated as "WT". *A.* Energy for the inward-open model, *B*. energy for the outward-open model.



Supporting Figure 3-4. Comparison of Rosetta and FoldX calculations.

For each mutation, the energy differences between the Inward open and Outward open models are plotted, as calculated by Rosetta and FoldX. The FoldX values are centered at zero because all energy differences are calculated relative to the WT model for that conformation (*i.e.,* there is no background energy difference between the Inward open and Outward open starting points, since both are set to zero). By contrast, Rosetta assigns a value of -12.8 to WT because it scores the Inward open conformation more favorably than the Outward open. Calculated energies are correlated with one another ($p<0.03$). Points are excluded for proline (Rosetta did not yield reasonable energies) and for aromatic amino acids (FoldX did not yield reasonable energies). *This figure was completed by collaborators*



Supporting Figure 3-5. Correlation of Rosetta energy scores and normalized initial substrate uptake.

Inward-open model *A*., outward-open model *B*. and inward-open minus outward-open model ***C*.** energy scores calculated using the Rosetta software suite are plotted against normalized substrate uptake values from Figure 3-3. No correlations were observed.

II. Appendix B: Supporting Tables for Chapter 4

Supporting Table 4-1. Pearson and Spearman Correlation Values calculate for various 271 position comparisons

	Pearson		Spearman	
	Value	p-value	Value	p-value
Taurocholate v Rosetta Inward	-0.08100	0.7342	0.02556	0.9148
Estrone-3-Sulfate v Rosetta Inward	-0.08745	0.7139	-0.1699	0.4738
Rosuvastatin v Rosetta Inward	-0.1733	0.4649	-0.2256	0.3390
Taurocholate v Rosetta Outward	-0.2838	0.2253	-0.1414	0.5522
Estrone-3-Sulfate v Rosetta Outward	-0.3026	0.1947	-0.3278	0.1582
Rosuvastatin v Rosetta Outward	-0.2894	0.2158	-0.2526	0.2826
Taurocholate v Rosetta Inward Minus Outward	0.04426	0.8530	0.1248	0.6001
Estrone-3-Sulfate v Rosetta Inward Minus Outward	0.04620	0.8466	-0.04211	0.8601
Rosuvastatin v Rosetta inward Minus Outward	-0.03923	0.8696	-0.1263	0.5957
Surface Expression v Rosetta Inward	-0.1782	0.4522	-0.09474	0.6912
Surface Expression v Rosetta Outward	-0.1075	0.6519	-0.05113	0.8305
Surface Expression v Inward Minus Outward	-0.1204	0.6130	-0.1609	0.4980
Taurocholate v Estrone-3-Sulfate	0.8802	**<0.0001**	0.8541	**<0.0001**
Taurocholate v Rosuvastatin	0.8807	**<0.0001**	0.8150	**<0.0001**
Estrone-3-Sulfate v Rosuvastatin	0.9527	**<0.0001**	0.9564	**<0.0001**

Pearson (linear) and Spearman (rank order- nonparametric) correlation scores from Figures 4-5, 4-6, and 4-7 calculated using GraphPad Prism 8. Significant correlations ($p>0.05$) are indicated in bold.

Supporting Table 4-2. Values used in RheoScale calculations

	Values					
Experiment	Maximum	Maximum override	Minimum	Minimum override	Dead	Recommended # of Bins
Surface Expression	160.7	-	1*	1	1	9
TCA Initial Uptake	587.1	400	2.8	-	2.8	5
E3S Initial Uptake	337.6	-	3.0	-	3.0	6
Rosuvastatin Initial Uptake	306.1	-	0.1	2	2	10

RheoScale calculator parameters used in Table 4-2. Maximum and minimum values correspond to the maximal and minimal values for the corresponding data set. If the maximum value was greater than 400, as was the case for a TCA initial uptake values, the maximal value was overridden to 400. If the minimal value was less than 2, the value was overridden to 2, as seen with the rosuvastatin uptake. For surface expression, multiple variants were not expressed or expressed at very low levels, giving values near or below 0. These number were replaced by a value of 1, thus the asterisks indicate the lowest arbitrary value of surface expression which corresponds to variants that are “dead” or not expressed on the cell surface. Recommended # of bins relates to the number of bins the calculator recommends depending on the data sets and their error. These values serve as a suggestion and are often changed to more properly fit the data.