



DISSERTATION / DOCTORAL THESIS

Titel der Dissertation / Title of the Doctoral Thesis

„Data-driven Molecular Modeling Studies with a Special
Focus on Hepatocellular Uptake Transporters“

verfasst von / submitted by

Mgr. Alzbeta Tuerkova, Bc.

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

Doktorin der Naturwissenschaften (Dr. rer. nat.)

Wien, 2020 / Vienna, 2020

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

A 796 610 449

Dissertationsgebiet lt. Studienblatt /
field of study as it appears on the student record
sheet:

Pharmazie

Betreut von / Supervisor:

Dr. Barbara Zdrazil, Priv.-Doz.

Preface

The computational research presented in this dissertation thesis was conducted from August 2017 until December 2020 in the Data Science & Computational Molecular Design research group at the Department of Pharmaceutical Chemistry, University of Vienna, under the supervision of Dr. Barbara Zdrazil, Priv.-Doz. Experimental work was performed by Réka Laczkó-Rigó, MSc (chemical synthesis and IC₅₀ measurements of 13-epiestrones), Prof. Gergely Szakács, and Dr. Csilla Özvegy-Laczka (OATP inhibition uptake experiments) from Medical University of Vienna, and Membrane Protein Research Group, Institute of Enzymology, Hungary, Dr. Marleen J. Meyer and Prof. Mladen V. Tzvetkov (generation of chimeric constructs, site-directed mutagenesis, and cellular uptake measurements of OCT1) from the University Medicine Greifswald, Germany. Additional computational work was performed by Dr. Sankalp Jain (OATP binary classification models) from National Institutes of Health, Maryland, US, Dr. Ulf Norinder (OATP conformal prediction) from Uppsala University, Sweden, and Brandon J. Bongers, MSc. and Prof. Gerard JP van Westen (OATP deep learning and proteochemometric models) from the University of Leiden, Netherlands.

Chapter 1 highlights the motivation of the study and introduces key research questions to be addressed in this thesis.

Chapter 2 provides theoretical aspects of the thesis. The biological background section describes the clinical relevance of solute carrier transporters with a special focus on hepatic organic anion transporting polypeptides and organic cation transporter 1. The methodological background covers cheminformatic (data integration, substructure mining) and structure-based (protein structure prediction, molecular docking, normal mode analysis) approaches used in this thesis. In addition, interconnection of the two different approaches is discussed in the review article included herein.

Chapter 3 contains a synopsis of results structured into six individual studies. Studies 1 to 4 have already been published in peer-reviewed journals. Studies 5 and 6 represent so-far unpublished data on the elucidation of steroids binding mechanism (Study 5) and prediction of novel OATP inhibitors (Study 6).

Chapter 4 provides a summary of the most essential findings and gives a future perspective for conducting follow-up studies.

Acknowledgements

First and foremost I am extremely grateful to my supervisor, Dr. Barbara Zdrazil, Priv.-Doz., for her invaluable advice, continuous support, patience during my PhD study, and for introducing me to the exciting field of data-driven computational design.

I would like to thank all the members of the Pharmacoinformatic Research Group for their help, support, as well as for undertaking numerous social activities. I would like to give a special thank to the ladies from the “Officemates” group - Claire, Ece, Eva, Florentina, Jennifer and Steffi - for a wonderful time spent in the office, and for being such great role models for women in computational chemistry.

My thanks goes to all collaborators who contributed to this study, namely Réka Laczkó-Rigó, MSc, Prof. Gergely Szakács, Dr. Csilla Özvegy-Laczka, Dr. Marleen J. Meyer, Prof. Mladen V. Tzvetkov, Dr. Sankalp Jain, Dr. Ulf Norinder, Brandon J. Bongers, MSc., and Prof. Gerard JP van Westen.

Next, I would like to express my gratitude to my family and friends, who took care about my work-life balance. Specifically, I would like to thank Anežka, Kája, Maruška, and Petra, who have always been here to support me through the hard times, or to have a chat & cup of coffee.

Last but definitely not least, I would like to thank Ivo for his love, understanding, encouragement in the past few years, and for being the best partner in both life and science.

• • •

My gratitude extends to the Austrian Science Fund (FWF, Grant P 29712) for the funding opportunity to undertake my studies at the Department of Pharmaceutical Chemistry, University of Vienna.

“The most important thing is to never stop questioning.”

— Albert Einstein

Declaration of Authorship

Hereby I declare, that the presented thesis is my original authorial work, which I have worked out by my own. All sources, references and literature used or excerpted during elaboration of this work are properly cited and listed in complete reference to the due source.

Vienna December 15, 2020

Alzbeta Tuerkova

Contents

I	MOTIVATION & AIMS	13
II	INTRODUCTION	19
1	Biological Background	21
1.1	Solute Carrier Transporters in Health and Disease	21
1.1.1	Hepatic organic anion transporting polypeptides	22
1.1.2	Organic cation transporter 1	23
2	Methodological Background	25
2.1	Cheminformatics as a Branch of Data Science	25
2.1.1	Automated integration of biomedical data	26
2.1.2	Substructure searches	29
2.2	Structure-based Molecular Modeling	30
2.2.1	Protein structure prediction	31
2.2.2	Molecular docking	33
2.2.3	Normal mode analysis as a tool to study protein dynamics	35
2.3	Current Advances in Studying Clinically Relevant Transporters of the Solute Carrier (SLC) Family by Connecting Compu- tational Modeling and Data Science	39
III	RESULTS	57
3	Synopsis of Results	59
3.1	Integrative Data Mining, Scaffold Analysis, and Sequential Binary Clas- sification Models for Exploring Ligand Profiles of Hepatic Organic Anion Transporting Polypeptides	61
3.2	Structural Dissection of 13-epiestrones Based on the Interaction with Hu- man Organic Anion-transporting Polypeptide, OATP2B1	77
3.3	A Ligand-based Computational Drug Repurposing Pipeline Using KNIME and Programmatic Data Access: Case Studies for Rare Diseases and COVID-19	88
3.4	Differences in Metformin and Thiamine Uptake between Human and Mouse Organic Cation Transporter OCT1: Structural Determinants and Potential Consequences for Intrahepatic Concentrations	109
3.5	Data-driven Ensemble Docking to Unravel Interactions of Steroid Analogs with Hepatic Organic Anion Transporting Polypeptides	123

3.6	Combining AI-driven and Structure-based Approaches to Identify Novel Inhibitors of Hepatic Organic Anion Transporting Polypeptides (OATPs)	183
3.6.1	Introduction	184
3.6.2	Material and Methods	185
3.6.3	Results & Discussion	188
3.6.4	Conclusions	197
IV	CONCLUDING DISCUSSION	199
V	SUPPLEMENTARY	227
VI	SCIENTIFIC OUTPUT	347
3.7	Publications	349
3.8	Selected Talks	351
3.9	Selected Posters	351
3.10	Awards	352
3.11	Teaching	352

List of Abbreviations

OATP	O rganic A nion T ransporting P olypeptide
ITC	I nternational T ransporter C onsortium
SLC	S o L ute C arrier
DAT	D op A mine T ransporter
SERT	S ERotonine T ransporter
LAT	L -type A mino acid T ransporter
ASCT	A SC amino acid T ransporter
OCT	O rganic C ation T ransporter
OAT	O rganic A nion T ransporter
OCTN	O rganic C ation T ransporter N ovel
MATE	M ultidrug T oxin E xtrusion protein
TMH	T rans M embrane H elix
MFS	M ajor F acilitator S uperfamily
ATP	A denosine T ri P hosphate
MPP	1- M ethyl-4- P henyl P yrдинium
TEA	T etra E thyl A mmonium
QSAR	Q uantitative S tructure- A ctivity R elationship
KNIME	K onsta N z I nformation M i N er
GLUT	G LUcose T ransporter
COVID-19	C ORona V irus D isease 20 19
IUPHAR	I nternational U nion of B asic & C linical P HARmacology
UniProtKB	U N I versal P RO T ein resource K nowledge B ase

PDB	P rotein D ata B ank
API	A pplication P rogramming I nterface
HTTP	H yper T ext T ransfer P rotocol
REST	R Epresentational S tate T ransfer
XML	e Xtensible M arkup L anguage
RDKIT	R ational D iscovery KIT
INCHI	I Nternational C hemical I dentifier
SMILES	S implified M olecular- I nterface L ine- E nter S ystem
ASCII	A merical S tandard C ode for I nformation I nterchange
MCS	M aximum C ommon S ubstructure
SMARTS	S Miles A Rbitrary T arget S pecification language
SB	S tructure- B ased
NMR	N uclear M agnetic R esonance
CRYO-EM	CRYO - E lectron M icroscopy
HTS	H igh- T roughput S equencing
AA	A mino A cid
SCR	S tructurally C onserved R egion
SVR	S tructurally V ariable R egion
PSIPRED	P osition- S pecific- I terative blast-based secondary structure P RE D iction
FUCP	FUC ose P ermease
CHARMM	C hemistry at HAR vard M olecular M echanics
DUD-E	D atabase of U seful D ecoys - E nhanced
AUC	A rea U nder the C urve
ROC	R eciever O perator C haracteristics
EF	E nrichment F actor
ADV	A uto D ock V ina
MWC	M onod- W yman- C hangus model
MD	M olecular D ynamics
NMA	N ormal M ode A nalysis

ENM	E lastic N etwork M odel
ANM	A nisotropic N etwork M odel
GNM	G aussian N etwork M odel
CDK	C hemistry D evelopment K it
RCSB	R earch C ollaboratory for S tructural B ioinformatics
AID	A ssay I D
CID	C ompound I D
CAS	C hemical A bstract S ervice
LabuteASA	L abute's A ccessible S urface A rea
SMR	M olecular R efractivity
TPSA	T opological P olar S urface A rea
URL	U niform R esource L ocator

Part I

MOTIVATION & AIMS

The liver is an organ accomplishing multiple physiological functions. [1] It converts various substances into the form that can be used by the organism, produces bile acids to promote lipid absorption, and regulates storage of glucose. The liver further represents a major site of drug metabolism and detoxification. [2] Hepatic transporters of the solute carrier (SLC) superfamily expressed at the basolateral membrane of human hepatocytes help maintain cellular homeostasis by regulating the transport of both endogenous substrates and xenobiotics. [3] Altered physiological function of these transporters can be attributed to their genetic variations or to the blockage of the transport by co-administered drugs. [4] Transporter-mediated drug-drug interactions are a relevant safety concern in the drug discovery process. An impaired function of liver cells might further evolve into severe medical consequences, such as hyperbilirubinemia (elevated levels of bilirubin in the blood) [5], or cholestasis (decreased or blocked flow of bile) [6].

Among others, hepatic organic anion transporting polypeptides (OATP1B1, OATP1B3, and OATP2B1), are polyspecific transporters with partially overlapping substrate or inhibitor profiles. [7] OATP-mediated drug-drug interactions with potential implications in hepatotoxicity are the reason why these transporters are listed among the most important transporters of emerging clinical importance in White Paper by the International Transporter Consortium (ITC) and also by U.S. Food and Drug Administration. [7] Examining novel therapeutic candidates for their potential interactions with hepatic OATPs has been recommended to be included in safety assessment strategies in the early-stage of a drug discovery pipeline. In addition, having new (selective) ligands to be used as tool compounds would enhance our understanding about the physiological role of these transporters. However, members of the OATP family are generally understudied and little is known about their molecular aspects of ligand recognition and selectivity.

In silico modeling approaches can be leveraged to gain an in-depth understanding of the OATP-ligand interactions. Structure-based studies can provide us with useful insights into the role of individual residues which might be responsible for selectivity switches. Up to this date, no experimentally-resolved structure of any OATP member has been released. Comparative modeling of OATP structures represents a challenge due to the fact that the possible structural templates share low sequence identity ($< 20\%$). Moreover, compound data for hepatic OATPs are heavily inconsistent and spread over diverse sources in the open domain.

In this thesis, we followed a holistic *in silico* approach, combining structure-based modeling with data-science (cheminformatics) techniques. First, integrative mining of OATP bioactivity data from public databases was performed. A thorough analysis of curated datasets helped to identify enriched substructures showing selective, dual-, or

pan-inhibitory activity against OATPs, as well as important physico-chemical features conferring general and/or preferred activity against OATP1B versus OATP2B subfamily. The trends delivered by R-group decomposition for a data set of steroidal compounds were further examined by newly measured 13-epiandrosterone derivatives. In addition, inhibitor datasets (uniting the compounds originating from public databases and from in-house measurements) served as a docking library to conduct molecular docking studies.

Next, available protein structures of structurally related Major Facilitator Superfamily (MFS) members were systematically analyzed with respect to their intrinsic dynamics. Normal mode analysis (NMA) provides us with insights into dominant protein motions which might have an impact on MFS transporters function. These findings prompted us to generate OATP structural models in distinct conformations based on templates with MFS fold. Ensemble docking followed by ligand enrichment calculations helped prioritize the best model per OATP transporter. Final models were used to establish a binding mode hypothesis for steroid analogs originating from the open domain and from in-house inhibition uptake experiments.

In order to further test the predictive power of the generated structural models, structure-based virtual screening of the ENAMINE Real database combined with different sorts of machine learning models was performed. Novel, potent inhibitors with different activity profiles across the three hepatic OATPs were identified by performing reuptake inhibition experiments of prioritized hits.

Next, a structure-based modeling approach was used for another hepatic transporter of pharmaceutical interest - Organic Cation Transporter 1 (OCT1). Chimeric constructs of human and mouse OCT1 identified a simultaneous replacement of transmembrane helices 2 (TMH2) and 3 (TMH3) to cause differences in the uptake of clinical substrates. Here, structural modeling of human and mouse OCT1 pointed an effect of coiled-coil interactions between TMH1 and TMH2 on the transport function.

The second aim of the thesis is to automate the *in silico* approaches deployed herein. For this purpose, multiple KNIME workflows were generated to conduct ligand-based studies. The applicability of KNIME workflows and programmatic data access is further examined for another emerging *in silico* method - computational drug repurposing. Here, workflow versatility is demonstrated by the help of two cases studies: (1) A rare disorder involving SLC transporters (here: GLUT-1 deficiency syndrome) and (2) a novel disease (here: COVID-19). In addition, NMA has been found as a powerful tool for large-scale exploitation of protein structural data to, e.g., determine conserved protein motions of an entire family of proteins. These findings can shape structure-based modeling studies, as shown for the case of hepatic OATPs.

Overall, structure-based modeling studies supplemented by the integration and analysis of ligand data provides comprehensive insights into OATP-ligand interactions and selectivity. The conclusions drawn here can help assist early-stage drug discovery by examining potential OATP-drug interactions or to design tool compounds to further elucidate the biological role of this class of transporters. In addition, human/mouse OCT1 studies revealed the role of tertiary interactions as an indirect mechanism which is capable to alter the function of hepatic transporters. Last but not least, leveraging open data using (semi-)automated workflows has been found to be useful approach to increase the confidence of structure-based modeling approaches applied herein.

Part II

INTRODUCTION

Chapter 1

Biological Background

1.1 Solute Carrier Transporters in Health and Disease

The human solute carrier (SLC) superfamily of membrane transporters consists of 446 members classified into 70 families on the basis of sequence identity¹. [8–11] SLC transporters mediate the transport of a broad spectrum of solutes across biological membranes, such as fatty acids and carbohydrates (SLC2 and SLC27 subfamilies) [12, 13], organic cations, anions, and zwitterions (SLC22 and SLCO subfamilies) [14, 15], bile acids (SLC10 subfamily) [16], or amino acids (SLC36 and SLC7 subfamilies) [17, 18].

Genetic studies have provided valuable insights into genetic polymorphism of SLCs. [10] Mutations in SLC genes can sometimes be linked to impairment of endogenous compound uptake. Malfunction of SLCs with a narrow substrate specificity can result in severe disorders. In this context, more than 100 SLC transporter-associated Mendelian diseases (known as monogenic disorders) have been identified. [19] For example, the dopamine (DAT, *SLC6A3* gene) and serotonin (SERT, *SLC6A4* gene) transporter, are recognized as relevant drug targets for neurological disorders. [20] Other SLCs, such as L-type amino acid transporter 2 (LAT, *SLC7A5* gene) [21], or ASC amino acid transporter (ASCT2, *SLC1A5* gene) [22], are largely overexpressed in different cancer cells and thus represent promising anticancer targets.

Many members of the SLC superfamily accept structurally diverse substrates, such as the SLC22 family of organic cation and anion transporters (OCTs, OATs, OCTNs), or the SLCO family (formerly known as SLC21). These polyspecific transporters enjoy considerable interest due to their essential role in the pharmacokinetics of xenobiotics. [23] Given their predominant expression in the liver and kidney, SLC22 and SLCO proteins

¹ $\geq 20\%$ of sequence identity is used as a threshold for SLC family classification.

regulate drug metabolism and elimination. [24] Therefore, assessing whether a novel drug candidate is a potential substrate or inhibitor of hepatic uptake transporters can provide safety measures to guide drug discovery and development.

The presented thesis is strongly oriented towards hepatic OATP-ligand interactions and subtype selectivity. In the following subsection we summarize current knowledge on hepatic OATPs (*SLCO1B1*, *SLCO1B3*, and *SLCO2B1* gene, respectively). More briefly, we describe another hepatic transporter under study - OCT1 (*SLC22A1* gene). Tissue expression, pharmacological importance, and substrate specificity of other clinically relevant transporters (OCT2, OCT3, OCTN1, OCTN2, MATE1, MATE2-K, OAT1, OAT2, OAT3, and OATP1A2) is covered by our review article (Section 2.3).

1.1.1 Hepatic organic anion transporting polypeptides

The hepatic organic anion transporting polypeptides - OATP1B1, OATP1B3, and OATP2B1 - are collectively expressed at the basolateral membrane of hepatocytes and are involved in the hepatobiliary transport of various compounds, such as bilirubin, bile salts, cholesterol phospholipids, hormones (and their conjugated forms), nutrients, toxins, and a wide range of prescription drugs. Hepatic OATPs are therefore increasingly recognized for the regulation of proper liver physiology, such as metabolism of bilirubin or enterohepatic circulation of bile salts. [5, 25] Defects in the expression and function of these transporters might lead to severe clinical consequences. For example, impaired uptake of bilirubin leads to elevated concentration of bilirubin in the blood, which in turn can result in the manifestation of Rotor syndrome. [26] Rotor syndrome is a rare, conjugated hyperbilirubinemia, induced by the simultaneous mutations in *SLCO1B1* and *SLCO1B3* genes.

OATPs exhibit broad substrate specificity with partially overlapping substrate/inhibitor profiles. Several specific compounds for OATP1B1 (e.g., pravastatin) [27], OATP1B3 (e.g., cholecystokinin octapeptide) [28], and OATP2B1 (e.g., erlotinib) [27], have been identified. Different ligand profiles across the three transporters might be attributed to the degree of their sequence similarities; while OATP1B1 and OATP1B3 share about 80% sequence identity, OATP2B1 is more evolutionary distant ($\sim 28\%$ sequence identity with OATP1B subfamily, Figure 1.1A). To date, molecular determinants conferring OATP subtype specificity remain to be elucidated.

OATPs are glycoproteins with 643-722 amino acids. Hydropathy analysis suggests 12 transmembrane helices (TMHs) interconnected by intra- and extra-cellular loops (Figure 1.1C). [29] A large extracellular domain between TMH9 and TMH10 contains a

disulfide-bonded pattern of eleven cysteine residues which resembles the kazal-type domain of serine protease inhibitors. [30] Other important structural features are the N-glycosylation sites in the extracellular loops 2 and 5 [31], phosphorylation sites at the N- and C-terminus [32], and the consensus sequence region spanning extracellular loop 3 and TMH6 region. Interestingly, this pattern (D-X-RW-[I,V]-GAWW-X-G-[F,L]-L, where X is a non-conserved residue) is fully conserved across the whole OATP superfamily (18 members, incl. OATPs from other species). [33] Based on the comparative modeling (Methods), OATPs possess Major Facilitator Superfamily (MFS) fold comprising multiple binding sites which are able to accommodate structurally unrelated compounds. [34] A rocker-switch model of transport was suggested for the MFS proteins (Figure 1.1D). [35] In this model, a transporter undergoes a transition from an outward-open state (i.e., extracellular-facing conformation) through intermediate states (i.e., occluded conformation) to an inward-open state (i.e., cytoplasm-facing conformation), where the substrate is released (Figure 1.1D). Early hypotheses about the transport mechanism were guided by the resolved structures of MFS members (such as lactose permease or glycerol-3-phosphate transport). A concentric, rigid-body motion, has been proposed for transporters. [36, 37] However, computational analyses have drawn the whole transition pathway as a cascade of multiple intermediate substates contributing to the allosteric mechanism of transport. [38] The transport is thought to be sodium- and ATP-independent. [39] Some studies suggest a role of pH which regulates OATP transport. [40, 41]

1.1.2 Organic cation transporter 1

OCT1 is predominantly localized at the sinusoidal membrane of human hepatocytes. OCT1 is a polyspecific transporter responsible for hepatic clearance of cationic drugs and other endogenous substrates (such as N-methyl quinine, MPP, aciclovir, ganciclovir, desipramine, TEA). [42] To give an example, OCT1 in connection with Multidrug and Toxin Extrusion 1 protein (MATE1, SLC47A1 gene) is implicated in metformin elimination. Therefore, risk assessment of potential drug-drug interactions ascribed to either OCT1 or MATE1 is of particular importance. [43] Similarly to hepatic OATPs, the physiological function of OCT1 might be altered by single nucleotide polymorphism (as shown, e.g., in case of metformin pharmacokinetics). [44] OCT1 is composed of 553 amino acids, possessing similar overall structural characteristics as predicted for OATPs, such as 12 membrane-spanning domains which constitute MFS fold. [45] Experimental evidence together with computational analyses suggested a whole interaction surface, rather than discrete binding sites, which accepts structurally diverse substrates. [46]

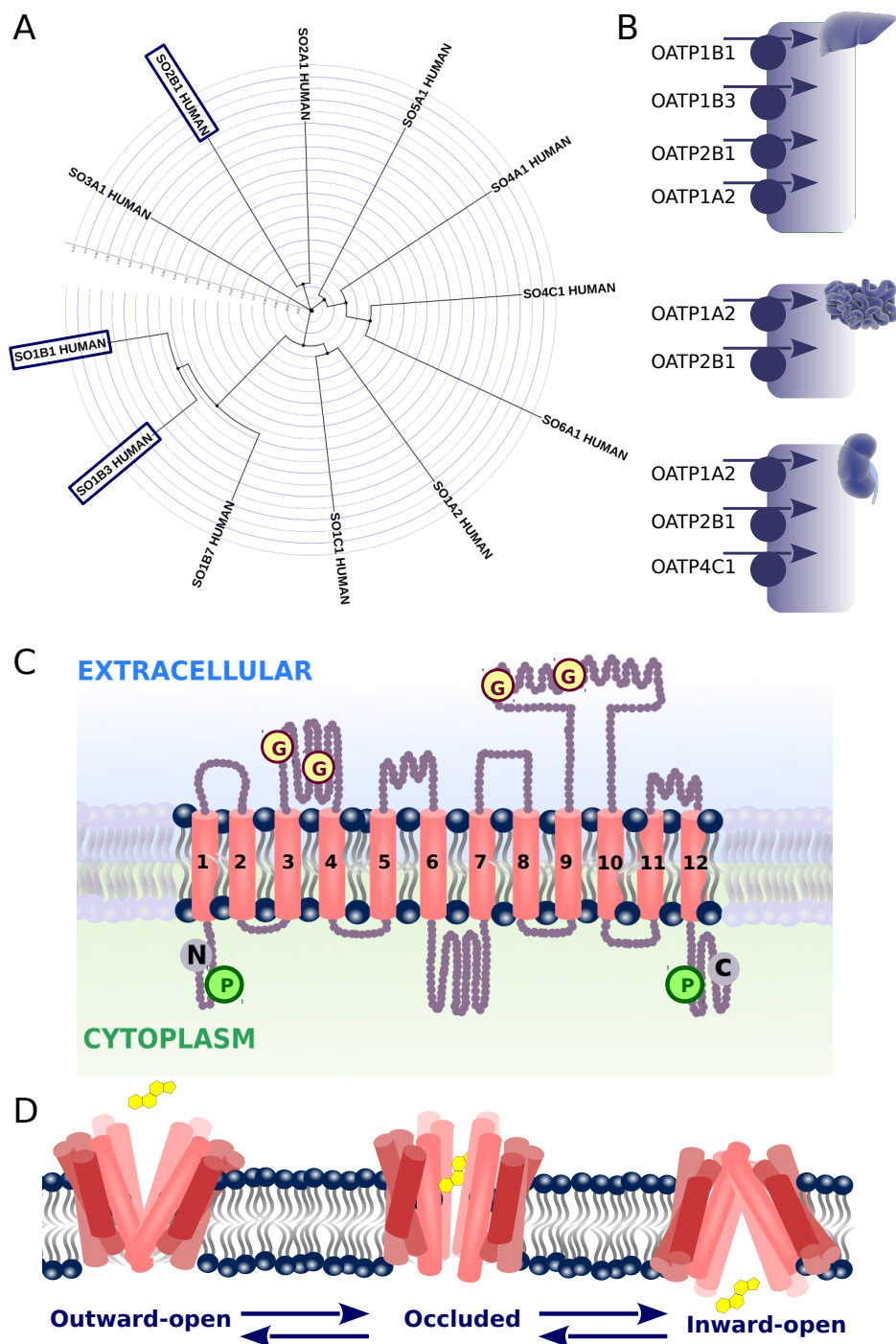


FIGURE 1.1: Phylogeny, tissue distribution, and structural aspects of hepatic OATPs. (A) Phylogenetic classification of human OATPs. Hepatic OATPs are highlighted. (B) Expression of human OATPs in liver, intestine, and kidney cells (from top to bottom). (C) Hydropathy analysis shows 12 transmembrane helices. Glycosylation ("G") and phosphorylation ("P") sites are located in the extra- and intra-cellular loops, respectively. (D) Rocker-switch mechanism of transport highlighting three major states - (1) outward-open, (2) occluded, and (3) inward-open.

Chapter 2

Methodological Background

The following chapter is conceived as an overview of methods used by the author of this thesis. The chapter provides the reader with theoretical essentials needed to understand the methodology applied herein. Specific computational protocols and technical details are described in the methodological sections of respective studies (Study 1-6 in Chapter 3).

2.1 Cheminformatics as a Branch of Data Science

Cheminformatics relates to data science in the sense that it extracts chemical information from small-molecule data. The term “cheminformatics” was originally associated with drug discovery [47]:

“Cheminformatics is the mixing of those information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and optimization.”

— F.K. Brown, 1998

Nowadays, cheminformatics covers a broad spectrum of different tasks, involving mining of chemical databases, data curation and standardization, molecular descriptor calculations, (sub)structure searches, chemical data visualization, exploration of chemical space, virtual screening, quantitative structure-activity relationships (QSAR), etc. [48] In an attempt to combine individual objectives into an united framework, different pipelining tools, such as Pipeline-Pilot [49] or Konstanz Information Miner (KNIME) [50], can be deployed. [51] In KNIME Analytics Platform, different operations on data (such as

data import/export, data transformation, training models, data visualizations) are accomplished via nodes. Each node consists of configuration parameters which can be adapted for a specific usage. The individual nodes are interconnected by input and output ports to constitute a workflow. In the course of this thesis, multiple KNIME workflows were developed to assist structure-based modeling studies. In Study 1 (Section 3.1) a KNIME workflow was created to integrate OATP bioactivity data from the public domain. The aim was to explore OATP ligand profiles with respect to enriched substructures and molecular features conferring OATP activity and/or selectivity. In Study 2 (Section 3.2) we expand on the scaffold analysis from Study 1 and perform structure-activity-relationship (SAR) by R-group decomposition. The workflow is applied for 13-epiestrones to define key molecular determinants driving OATP2B1 inhibition. In Study 3 (Section 3.3) we present a ligand-based drug repurposing pipeline in KNIME. The versatility of the workflow is demonstrated by using two case studies: (1) A rare disease involving clinically relevant glucose transporter 1 (here: GLUT-1 deficiency syndrome) and (2) a novel disease (here: COVID-19). In the following two subsections we give a brief overview on the biomedical data mining with a special focus on programmatic data access, data curation, and substructure searches. Specific workflow settings and related technical details can be found in the corresponding studies (Study 1-3).

2.1.1 Automated integration of biomedical data

The era of open data has reshaped drug discovery. [52] Nowadays, mining of chemical structures for the sake of data augmentation (e.g., to prepare a docking library or to train a machine learning model) is no longer its sole purpose. Instead, large-scale data analysis by using (semi)automated workflows enables to extract hidden patterns which might impact computer-aided drug discovery. [53] Moreover, programmatic access of biomedical data brings the ultimate benefit of workflow flexibility and reproducibility. The majority of databases used herein (e.g., ChEMBL [54], DrugBank [55], IUPHAR [56], UniProtKB [57], Protein Data Bank [58], Open Targets [59]) are providing targeted data access through an Application Programming Interface (API).

In this thesis, a triad of KNIME nodes is consecutively executed to fetch data through API requests. One concrete example is illustrated here - retrieval of available PDB structures on the basis of target UniProt IDs (here: Glucose transporter 1, UniProt ID=P11166):

(1) *Specification of an API request.* Here, the `join()` function in the “String Manipulation” node is used. The respective UniProt ID (“UniProt ID” column) is inserted

to a string as a variable (defined by the `$` character). The `strip()` function is used to remove blank space from the beginning and end of the UniProt ID variable:

```
join("http://www.uniprot.org/uniprot/",strip($Uniprot ID$),".xml")
```

Output of the “String Manipulation” node is a RESTful URL link:

```
http://www.uniprot.org/uniprot/P11166.xml
```

(2) Data retrieval from web services.

Here, the “GET request” node is executed and a data file for the requested target is appended to an output table. Additional columns list HTTP status code and content type (here: `application/xml`).

(3) Extraction of relevant properties from received files. Here, available PDB IDs are retrieved via a following XPath query:

```
/dns:uniprot/dns:entry/dns:dbReference[@type='PDB']/@id
```

The `dns` prefix corresponds to a specific namespace (here: `http://uniprot.org/uniprot`). The `@` character is used to specify certain XML attributes in the XPath query. Here, `dbReference[@type='PDB']` is forwarded to the XPath query to get all experimental structures by querying the `@id` (here: PDB ID) attribute.

The general procedure presented herein can be adapted to suit a specific API syntax or data format. Here, we have adapted a similar strategy to map the Open Target ID of a certain disease to available targets. For details the reader is pointed to Study 3 (Section 3.3). In addition, APIs were used to fetch ligand bioactivity data from DrugBank (Study 1, Section 3.1), ChEMBL (Study 3, Section 3.3), IUPHAR (Study 1 and 3 in Section 3.1 and 3.3), and PubChEM (Study 3 in Section 3.3). API calls were also used to query PDB to gather protein-ligand complexes in the framework of the drug repurposing pipeline (Study 3 in Section 3.3), or to check the correct stereochemistry of steroidal compounds (Study 5 in Section 3.6).

A prerequisite for merging ligand data from heterogeneous sources is to unify the representation of chemical data. [60] In this thesis, a structure curation procedure was built up on the basis of open cheminformatics software (RDKit [61] and Chemistry Development Kit [48]) and involves:

1. Removal of compound stereochemistry¹
2. Stripping off salts and other fragments
3. Neutralization of charges
4. Checking atomic clashes
5. Filtering out entries with unusual elements
6. Generating InChI, InChiKey, and canonical SMILES, from the standardized structures

For further details regarding compound standardization the reader is pointed to Study 3 (Section 3.3). InChI, InChiKey, and SMILES, are textual formats used to encode structural information.

InChI (International Chemical Identifier) is a molecular identifier composed of a sequence of individual layers encoding different types of information - the atoms, atom connectivity, tautomeric form, isotope type, stereochemistry, and electronic charge. [62] InChI does not have to contain all information, e.g the stereochemical layer can be omitted when performing 2D similarity searches. The InChI string is human-readable.

In contrast, the InChiKey is a hashed molecular representation derived from InChI. The InChiKey is composed of a fixed length of 27 characters divided into the five sections:

AAAAAAAAAAAAA-BBBBBBBFV-P

The first section (here: "A", 14 characters) corresponds to the core molecular constitution. The second section (here: "B", 8 characters) describes additional structural characteristics, if available (compound stereochemistry, isotopes, hydrogen positions, metal ligation). The third section (here: "F", single character) adds the flag - "S" for standard or "N" for non-standard InChI. The fourth section (here: "V", single character) corresponds to the version, and the last section (here: "P", single character) corresponds to the information about compound protonation/deprotonation. Hashed representation of the InChiKey makes the format non-human readable. The InChiKey is a fixed-length format and thus a more convenient identifier for, e.g., database searches. In the course of this thesis, the InChiKey was used to remove duplicated compounds from the dataset and to calculate median bioactivity values. For details regarding the

¹Compound stereochemistry is not needed for, e.g., 2D substructure searches or the calculation of topology-based descriptors.

bioactivity data curation and assignment of binary activity labels the reader is pointed to Study 1 (Section 3.1).

SMILES (Simplified Molecular-Input Line-Entry System) is a chemical format using ASCII strings. [63] Atoms are represented by element abbreviations. Apart from the common atoms (i.e., B, C, N, O, P, S, F, Cl, Br, I elements, atoms possessing no formal charge, atoms with normal valence and isotopes, and atoms which do not represent any chiral center), the individual elements have to be indicated in square brackets (e.g., [Ag] for silver). Atomic charges and hydrogens are explicitly included into the SMILES notation. If there is more than a single charge, the number of charges is written as a number (e.g., [NH4+] for ammonium). The “.”, “-”, “=”, “#”, “\$”, “:”, “/”, and “\” characters, respectively, are used to encode chemical bonds. If no bond is specified in the string, it is assumed to be a single bond (e.g., CCCO for propanol), although the character “-” can be used (e.g., C-C-C-O for propanol). Double bonds are encoded by the “=” (e.g., O=C=O for carbon dioxide), triple bonds by the “#” (e.g., C#N for hydrogen cyanide), quadruple bonds by the “\$” character (e.g., [Rh-](Cl)(Cl)(Cl)(Cl)\$[Rh-](Cl)(Cl)(Cl)Cl for octachlorodirhenate (III)), respectively. Dissociated structures are defined as the “.” character (e.g., CC(=O)[O-].CC(=O)[O-].O.[Cr+2] for chromium[II] acetate hydrate). Ring structures are indicated by the digit label to show the connectivity between non-adjacent atoms (e.g., C1CCCCC1 for cyclohexane). Aromaticity can be encoded in three different ways: (1) Kekulé notation of alternating single- and double- bonds (e.g., C1=CC=CC=C1 for benzene), (2) Using the “:” character (e.g., C1:C:C:C:C:C1 for benzene), (3) Writing the atoms within a ring as lower-case (e.g., c1ccccc1 for benzene). Branching in molecular topology is defined by the parentheses (e.g., CCC(=O)O for propionic acid). Stereochemistry can be included using the “\” or “/” directional character, respectively (e.g., F/C=C/F for trans- and F/C=C\F for cis-1,2-difluoroethylene). Configuration at the stereogenic center is encoded by the “@” or “@@” character, respectively. Isotopes are defined by a value of the isotopic mass placed before the atomic symbol.

Chemical compounds in the SMILES representation were extensively used for different substructure searches in the course of this thesis (Study 1-3 in the corresponding sections). The following subsection will briefly describe the theory of substructure searches.

2.1.2 Substructure searches

Substructure searching (also known as “molecule mining”) is a method governed by the “molecular similarity principle”: Structurally similar molecules tend to possess similar

molecular properties. [64] Molecular similarity can be explored by calculating molecular descriptors, or by using chemical fingerprints which encode the presence/absence of certain structural features (e.g., distinct functional groups). [65, 66] However, information about the specific molecular topology is generally not fully captured by these approaches. Rather, approaches based on graph theory can be deployed. Within graph matching approaches, the molecular structure is perceived as a graph composed of vertices (here: atoms) connected by edges (here: chemical bonds). Chemical searches based on molecular topology are well-suited to, e.g., differentiate between distinct structural isomers (e.g., dimethylpropane and n-pentane).

In this thesis, the Maximum Common Substructure (MCS) approach has been applied as a similarity metric. [67] Given two structures, A and B, if A is a subcomponent of B, then A is defined as a substructure of B, whereas B is defined as a superstructure of A. Substructure A is believed to be responsible for the shared biological activity between molecule A and B. When performing database searches on the basis of molecular topology, Smiles ARbitrary Target Specification language (SMARTS), can be used. SMARTS is an extension of SMILES which is exploited to specify sub structural patterns in chemical data. Structural query generated via SMARTS can contain wild card characters in order to achieve greater flexibility in the substructure searches.

2.2 Structure-based Molecular Modeling

Structure-based (SB) modeling is a computational technique to study target-ligand recognition by modeling 3D molecular interactions. SB modeling relies on the availability of a target 3D structure. Substantial advances in experimental methods of structural biology (such as X-ray crystallography, nuclear magnetic resonance spectroscopy, or Cryo-electron microscopy) have been made over the past decade, yielding 168,599 macromolecular structures deposited in Protein Data Bank (updated in September 2020). Despite advanced techniques for macromolecular structure determination, high-throughput sequencing (HTS) technologies are producing a significantly higher amount of novel protein sequences at an accelerated rate.

The domain of SLC transporters is heavily understudied, as evidenced by the fact that only 52 SLC transporters are structurally covered (177 PDB entries, updated in October 2020²). In such a case, computational prediction of protein structures might serve as an alternative approach to obtain a 3D model for a protein of interest. In this thesis, structural models for human and mouse Organic Cation Transporter 1 (OCT1, Study 4

²Number of experimentally-resolved structures was identified by using PDB RESTful webservices querying “SLC” keyword and subsequent analysis of gathered entries.

in Section 3.4) and OATP1B1, OATP1B3, and OATP2B1 (Study 5 in Section 3.6) were generated.

2.2.1 Protein structure prediction

Prediction of 3D protein structures from its primary sequence is ranked among the top 125 challenges of current scientific efforts. [69] Protein modeling methods are generally divided into three main categories: (1) *Ab-initio* methods, (2) homology modeling, and (3) protein threading methods.

Ab-initio (de novo) methods can be used in case when no structural template for the protein of interest is available. The method includes exhaustive conformational searches following thermodynamic principles to identify low energy conformers. *Ab-initio* methods are primarily used for small peptides only, as the computational costs are too high to be applicable for standard proteins (> 120 amino acids). To date, *ab-initio* methods are mostly utilized for modeling and/or optimization of loop regions. [70–72]

Homology modeling methods rely on the availability of the structural template. In order to map a template structure to the unknown protein (“target”), sequence alignment is necessary. The sequence-to-structure relationship was first demonstrated by Chothia and Lesk in 1986. [73] The authors postulated that sequentially similar proteins tend to possess the same 3D structure. The template-target alignment consists of structurally conserved regions (SCRs) and structurally variable regions (SVRs). SCRs usually correspond to distinct structural motifs that are found e.g., in the hydrophobic core of the membrane proteins and/or in the binding sites. SVRs, on the contrary, usually correspond to loops which can have highly variable size and composition. The construction of the homology model comprises three consecutive steps: (1) first the protein “core” is modeled based on the backbone of SCRs, (2) loops joining the individual core segments are added, and (3) the side chains are modeled. In this dissertation thesis, the Modeller program was applied in Study 4 (Section 3.4), and Study 5 (Section 3.6). [74] The software works by satisfying spatial restraints, originating either from the template-target alignment and/or from the statistical analysis of various structural features (e.g., residue solvent accessibility, distance distribution of C-alpha atoms, side-chain dihedral angles). Individual restraints are defined as a probability density function (pdf), combined into a molecular (objective) function. For sake of model optimization, a combination of conjugated gradients method with molecular dynamics and simulated annealing is applied.

Homology modeling is a technique that is used for proteins with a high degree of homology ($\sim 70\%$ sequence identity). However, applying this method for proteins with low sequence identity (so called “twilight zone”, $< 20\%$ sequence identity) might lead

to low-quality models. [75] Therefore, predicting protein structures by threading (also known as fold recognition) might provide a viable alternative. The method relies on the fact that protein structure space is significantly smaller than the number of available sequences. [76] The structure prediction is done by statistical analysis of available experimental structures in relation to the target protein. The template-target fitness is evaluated by the scoring function, combining pairwise potential, environment fitness potential, mutational potential, gap penalty, hydrophobic burial, solvent accessibility, etc. In this thesis, the pGenThreader server [77] was used to predict structural templates for human and mouse OCT1 (Study 4 in Section 3.4) and hepatic OATPs (Study 5 in Section 3.6). The algorithm performs four consecutive steps: (1) the PSI-blast based secondary structure PREDiction (PSIPRED) method is used to predict the secondary structure of a target from its sequence. [77] (2) The profile-profile scoring scheme is generated to evaluate the target-template fitness. (3) A hydrophobic burial term is used to bias target-template alignment to the correct localization of hydrophobic residues. (4) Secondary structure dependent gap penalties are applied to revise a final alignment. In this thesis, Human Glucose 3 Transporter (GLUT3, PDB ID: 4zw9) [78] was identified among highly-ranked suitable templates for human (p-value ≤ 0.0001 , prediction score 103, 19.1% sequence identity) and mouse OCT1 (p-value ≤ 0.0001 , prediction score 109, 19.5% sequence identity). The final templates were selected according to several aspects, such as a high crystal resolution (1.5 Å) and the presence of an outward-occluded conformation, which is an optimal conformational state for investigating substrate binding. For methodological details the reader is pointed to the “Computational modeling” part in Study 4 (Section 3.4). Fucose transporter in an outward-open conformation (FucP, PDB ID: 3o7q) was predicted as a suitable template for OATP1B1 (p-value ≤ 0.0001 , prediction score 74, 14.5% sequence identity), OATP1B3 (p-value ≤ 0.0001 , prediction score 75, 15.6% sequence identity), and OATP2B1 (p-value ≤ 0.0001 , prediction score 93, 15.2% sequence identity), respectively. For methodological details the reader is pointed to the “Comparative modeling of hepatic OATPs” part in Study 5 (Section 3.6).

The quality of the structural models can be evaluated on the basis of either statistical potentials [79] or physics-based scores utilizing molecular mechanics force fields (e.g, effective force field based on the CHARMM parameters [80]). In this thesis, the MolProbit validation tool was used due to its applicability to membrane proteins (Study 4 in Section 3.4 and Study 5 in Section 3.6). [81] Validation analysis within the MolProbit tool consists of four consecutive steps: (1) Addition of hydrogen atoms. Asn/Gln/His flips are automatically corrected and -OH, -SH, and -NH₃ groups are rotationally optimized. (2) All-atom contact analysis by probing the amount of overlap between the non-bonded atoms. The “clashscore” generated upon contact analysis corresponds to the number of significant clashes (non-H-bond atom overlap > 0.4) per 1000 atoms.

(3) Ramachandran and rotamer analyses of backbone and side-chains, respectively. (4) Covalent-geometry analysis by checking the outliers of backbone bond-lengths and bond-angles. The final MolProbity score unites all individual quality metrics into a single value. The MolProbity score was used to filter out low-quality models for OCT1 (Study 4 in Section 3.4) and OATP1B1/OATP1B3/OATP2B1 (Study 5 in Section 3.6).

The quality of structural models can also be exemplified on the basis of enrichment of known ligands among a pool of inactive ligands and/or decoys. [82] Here, enrichment docking was employed in Study 5 (Section 3.6) to prioritize the best model for each hepatic OATP transporter. We followed a recommendation from the Directory of Useful Decoys (DUD-E) database and used the following ratio [83]:

$$1 \text{ (active)} : 36 \text{ (inactives)}$$

The area under the Curve (AUC) for Receive Operator Characteristic (ROC) plot was used as a metric to rank models from ensemble docking (see Study 5 in Section 3.6 for further details). [84] An enrichment factor of 1% (EF%1) of the docking database was used as an additional parameter to prioritize the final model. [85] For methodological details the reader is pointed to the “Preparation of docking library” and the “Enrichment docking for model prioritization” part of the Study 5 (Section 3.6).

2.2.2 Molecular docking

Molecular docking is a computational technique used to predict the optimal arrangement of ligands within a target binding site. [86] The general procedure consists of the following steps: (1) Grid coordinates for the search space are specified. (2) Different ligand positions, orientations, and conformations are sampled within a defined search space. (3) Calculated poses are evaluated according to the number of energetically favorable intermolecular interactions (“scoring function”). Scoring functions aim to approximate a standard chemical potential of the system. [87]

In Study 5 (Section 3.6) the entire transmembrane core was initially defined as putative binding site. The first round of enrichment docking calculations was done and structural models were ranked according to their AUC values. The top five models were retained and the contact surface area between the known ligands and residues was calculated. The calculated area allowed the identification of interaction “hot spots” in the transmembrane region. This information was used to re-define the search space for the second round of enrichment docking calculations into the restricted binding site. At the end, the top model per OATP transporter was prioritized on the basis of both AUC and EF1%

values. Final models were used to establish a binding mode hypothesis for OATP inhibitors/substates with a steroidal scaffold. In Study 6 (Section ??) possible interaction sites in OATP1B1, OATP1B3, and OATP2B1 transporters were mapped via a small molecule mapping server, FTMap. [88] An equivalent binding cavity was identified for all the three transporters (lined by TMH1, TMH2, TMH4, TMH5, TMH7, TMH8, and TMH11). Interestingly, the central binding site identified upon small molecule docking was also shown to be implicated into recognition of steroidal compounds in OATP1B1 and OATP1B3, as shown in Study 5 (Section 3.6). This region was defined as a binding site in Study 6 (Section ??).

AutoDock Vina (ADV, version 1.1.2.) was used to perform docking calculations in Study 5 (Section 3.6) and Study 6 (Section ??). [89] ADV uses a hybrid scoring function combining knowledge-based potentials and empirical scoring functions. The scoring function of ADV has been inspired by the X-score function. The function consists of distinct energetic terms, involving a weighted sum of steric interactions (attractive Gaussian function and repulsive terms), H-bond formation (where applicable), and hydrophobic effect. [90] Different ligand conformations are sampled by the stochastic Broyden-Fletcher-Goldfarb-Shanno method, built up on the basis of the quasi-Newton optimization method. [91] ADV uses a united-atom representation, i.e., only heavy atoms are included into the model. However, the user has to provide an input structure with a corrected protonation state (explicit hydrogens) to properly assign which functional groups can act as H-bond donors and/or acceptors. Protein and ligand structures were prepared via Autodock Tools scripts freely provided by the community (available at <http://autodock.scripps.edu/>). For concrete settings of docking runs the reader is pointed to Study 5 (Section 3.6) and Study 6 (Section ??).

Early docking experiments were guided by Fisher’s theory of the static protein structure (“lock”) and the ligand with fixed conformation (“key”). Although the shape complementarity approach introduced by the lock-and-key model is computationally fast, it completely omits the effect of dynamics accompanying protein-ligand interactions. The induced-fit model pioneered by Koshland [92, 93] assumes that the conformational change of a target is induced upon a ligand binding:

“As a glove changes shape when a hand slips into it, so an enzyme changes its conformation on binding a ligand.”

— D.E. Koshland, 1995

Such a model satisfies protein motions on a local scale, i.e., transitions between individual side-chain rotamers to adopt an optimal pose. An alternative model was introduced by Monod, Wymand, and Changeux (“MWC model”) and is based on the assumption

of so-called conformational selection. [94] In other words, a target samples an ensemble of different conformational substates under native state conditions and the ligand “selects” the protein conformer optimal for binding. A plethora of published studies point to a complex interplay between the two models which include protein flexibility. [95–97] In 2009 Bakan and Bahar performed an extensive study of conformational changes experimentally observed in the structures of three known enzyme targets (HIV-1 reverse transcriptase, p38 MAP kinase, and cyclin-dependent kinase 2). [98] The authors reported that the ligand is capable of recognizing a protein conformer which is intrinsically accessible under the equilibrium state (i.e., in the ligand-unbound form). The conclusions drawn from these studies point to the importance of so-called ensemble docking. [99] Here, the major challenge is to generate a structural ensemble and to interpret the docked poses. Conformational sampling suited for ensemble docking purposes has extensively been achieved via Molecular Dynamics (MD) approaches so far. [100–103] However, conventional MD simulations are not always an optimal choice due to the high computational costs, as well as the limited time scale to simulate and/or due to the force field errors. Here, we used Normal Mode Analysis (NMA) as an alternative approach to sample distinct template conformers (Study 5 in Section 3.6).

In the course of this thesis, an ensemble docking approach was used in Study 5 (Section 3.6). First, multiple OATP structures with various degrees of global (i.e., different backbone conformers) and local (i.e., different side-chain rotamers) flexibility was generated. The best poses were prioritized on the basis of the ligand enrichment, as described in the previous section. For methodological details the reader is pointed to Study 5 (Section 3.6).

2.2.3 Normal mode analysis as a tool to study protein dynamics

Structure-to-function relationship is a well-established paradigm in SB modeling. Yet, the structure itself becomes insufficient to explain a broad spectrum of biological activities. Specifically, a protein usually undergoes a conformational change to accomplish its biological function. Therefore, *dynamics* is a key factor which can bridge the gap between the protein structure and its function. In the field of SLC transporters, current studies reported that different functionalities across protein families sharing the same fold do not solely rely on the sequence (dis)similarity, but also on the differences in their conformational flexibility, which might confer substrate specificity. [104]

Normal mode analysis (NMA) using elastic network models (ENMs) is an analytical approach to study protein conformational dynamics. [105, 106]. ENMs provide an approximate representation of the potential energy function of a system, which is capable

of deriving the intrinsic dynamics near its equilibrium state. Beads in the network either represent individual atoms [107], or different sorts of coarse-grained entities, such as single nucleotides [108], residues (represented by C-alpha atoms) [109], or even the entire blocks of grouped residues (so called rigid-block models [110]). Each bead (here: network “node”) is a point in Cartesian space, connected to its spatial neighbors via uniform spring (here: network “edge”). Network edges are treated as harmonic restraints and are representatives of the bonded and non-bonded inter-residue interactions within a predefined cut-off. [111] The cut-off distance is usually set to 7.0 Å, based on the first coordination shell around residues observed in the experimentally-resolved structures. [112, 113] Using ENM-based NMA approach, protein dynamics can be interpreted by fluctuations in the network topology. Two types of ENMs are distinguished: (1) Gaussian network model (GNM), and (2) Anisotropic network model (ANM). Below we describe theoretical essentials behind GNM and ANM.

GNM is a basic model with a major utility to evaluate mean-square fluctuations and cross-correlations between the pairs of nodes (e.g., residues in the protein). In GNM, distance vector between residue i and j , R_{ij} , can be expressed as

$$\Delta R_{ij} = R_{ij} - R_{ij}^0 \quad (2.1)$$

where R_{ij} is the change in the distance between residue i and j from the initial distance R_{ij}^0 . In GNM, the inter-residue fluctuations are considered to be isotropic and Gaussian. The potential function of the entire network, V_{GNM} , can be then expressed as

$$V_{GNM} = \frac{\gamma}{2} \left[\sum_{i,j}^N \Gamma_{i,j} [(\Delta X_i - \Delta X_j)^2 + (\Delta Y_i - \Delta Y_j)^2 + (\Delta Z_i - \Delta Z_j)^2] \right] \quad (2.2)$$

,where γ is the force constant (usually set to be uniform for all interactions), Γ_{ij} is the Kirchoff matrix of the connectivities between residue i and j , as defined by

$$\Gamma_{ij} = \begin{cases} -1, & \text{if } i \neq j \text{ and } R_{ij} \leq r_c \\ 0 & \text{if } i \neq j \text{ and } R_{ij} > r_c \\ -\sum_{j,j \neq i} \Gamma_{ij} & \text{if } i = j \end{cases} \quad (2.3)$$

, where r_c is the cut-off for inter-residue contacts used in the GNM. Diagonal elements in the Kirchoff matrix are the total number of connections between individual nodes in the network. It can therefore be interpreted as the residue coordination number z , i.e., the

measure of the local packing density around a given residue. In a first approximation, mean square fluctuations, $(\Delta R_{ij})^2$ scales with $\Gamma^{-1} = 1/z$. [114]

Mean square residue fluctuations calculated from GNM have been found to be in good agreement with crystallographic B-factors. [113, 115, 116] One can make a connection between the theoretical residues fluctuations and calculated residues fluctuations via the following expression

$$B_i = \frac{8\pi^2}{3} \langle (\Delta R_i)^2 \rangle = \frac{8\pi^2 k_B T}{\gamma} [\Gamma^{-1}]_{ii} \quad (2.4)$$

Comparison of the fluctuations calculated by GNMs with B-factors deposited in PDB is a valid strategy to exemplify the predictive power of the computational models. [117] The major purpose of the methodology discussed herein is to derive functional motions of the proteins, i.e., those which are highly cooperative. Conformational transitions of the entire protein domains pave the substrate translocation pathway. Such modes of motions can be identified by converting the Kirchoff matrix into a product composed of three matrices: (1) the matrix of eigenvectors, U , (2) the diagonal matrix of eigenvalues, Λ , and (3) the transpose of the unitary matrix, U^T , as follows

$$\Gamma = U \Lambda U^T \quad (2.5)$$

Eigenvalues correspond to the mode *frequencies*, while the eigenvectors correspond to the mode *shapes*. Frequency of a mode i , λ_i , reflects the potential energy of the normal mode 1. The individual modes are ordered based on the ascending potential energy such as $\lambda_1 < \lambda_2 < \lambda_3 \dots$. In the context of protein functional motions, one seeks to unravel low-frequency (also called “soft”) modes, as these imply substantial conformational changes and thus are possessing the most prominent contribution to the overall protein dynamics. At the other end of the spectrum, the high-frequency modes do represent highly localized fluctuations, such as a loop motion within a binding site.

As an extension to the GNM, ANM permits to capture the directionality of the protein motions. In the ANM potential energy function, the $N \times N$ Kirchoff matrix is replaced by the $3N \times 3N$ Hessian matrix. The potential energy function, V_{ANM} , is then calculated as

$$V_{ANM} = \frac{\gamma}{2} \left[\sum_{i,j}^N (R_{ij} - R_{ij}^0)^2 H(r_c - R_{ij}) \right] \quad (2.6)$$

where H corresponds to the Hessian matrix (second partial derivative of the potential V), such that

$$H = \begin{bmatrix} H_{ii} & H_{ij} \\ H_{ji} & H_{jj} \end{bmatrix} \quad (2.7)$$

where $H_{i,j}$ element is a 3 x 3 matrix encoding the anisotropic orientation of nodes i and j . Each super element of the Hessian can be expressed as

$$H_{ij} = \begin{bmatrix} \frac{\partial^2 V_{ij}}{\partial x_i \partial x_j} & \frac{\partial^2 V_{ij}}{\partial y_i \partial x_j} & \frac{\partial^2 V_{ij}}{\partial z_i \partial x_j} \\ \frac{\partial^2 V_{ij}}{\partial y_i \partial x_j} & \frac{\partial^2 V_{ij}}{\partial y_i \partial y_j} & \frac{\partial^2 V_{ij}}{\partial y_i \partial z_j} \\ \frac{\partial^2 V_{ij}}{\partial z_i \partial x_j} & \frac{\partial^2 V_{ij}}{\partial z_i \partial y_j} & \frac{\partial^2 V_{ij}}{\partial z_i \partial z_j} \end{bmatrix} \quad (2.8)$$

The ENM-based NMA approach was originally used by Flory et al to study statistical mechanics of polymers. [118] In 1997, Bahar et al applied the same principles to protein structures in order to unravel their dynamics behavior. [113] Since then, NMA has been successfully applied to a broad range of biomolecular systems, such as different enzymes [119, 120], membrane proteins [121], protein complexes [122], and even supramolecular assemblies, such as ribosomes [123] or virus particles [124, 125]. Different biomolecular mechanisms have been elucidated using the NMA approach, such as allosteric modulations [126, 127], protein-protein interactions [128] or protein conformational transitions. [129] In recent times, considerable advances of ENMs have been made to model different non-equilibrium phenomena, such as protein transition pathways. [130] Capturing protein structures at different stages of the transport cycle can provide valuable insights into the transport mechanism, as demonstrated, e.g., for fucose transporter. [38]

Here, the motivation for applying the NMA approach was twofold; First, normal modes based on Gaussian Network Model (GNM) were calculated for all available Major Facilitator Superfamily (MFS) structures from Protein Data Bank (PDB). The idea behind was to classify the MFS proteins according to their dynamic landscape to derive distant relationships, as well as to uncover dominant protein changes which might be functionally relevant for the ligand recognition. For methodological details the reader is pointed to Study 5 (Section 3.6). Next, Anisotropic Network Models (ANM) were used to sample alternate protein conformations along the selected normal modes. The aim was to create a conformational ensemble for the MFS template (here: Fucose transporter in an outward-open conformation, PDB ID 3o7q), which provided a basis for generating

multiple OATP structural models for ensemble docking. For methodological details the reader is pointed to Study 5 (Section ??).

2.3 Current Advances in Studying Clinically Relevant Transporters of the Solute Carrier (SLC) Family by Connecting Computational Modeling and Data Science

TÜRKOVÁ, Alžběta; ZDRAZIL, Barbara. *Computational and Structural Biotechnology Journal*, **2019**, 17: 390-405.

In the following review article we outline the challenges and applications of combining data-science (cheminformatics) and structure-based modeling approaches to study ligand-transporter interactions.

Chapter 1 provides the biological context of clinically important transporters. Here, members of organic cation and anion transporter subfamilies (OCTs, OATs, and OATPs), as well as the Multidrug and Toxin Extrusion (MATE) transporters, are discussed with respect to their involvement in drug bioavailability and disposition, drug-drug interactions, and specific organ toxicities. Chapter 2 introduces the basic concepts of molecular modeling approaches, covering methods of ligand-based (QSAR, ligand-based pharmacophores) and structure-based (different approaches for protein structure prediction, loop modeling, validation of protein 3D models, molecular docking, structure-based pharmacophores) modeling, as well as molecular-dynamics-related methods. In Chapter 3 data availability on the both ligand- (i.e., bioactivity measurements) and protein- (i.e., experimentally resolved structures) level is demonstrated by studying the evolution of different computer-based methods used in published *in silico* studies. Furthermore, a protein threading approach is applied for the selected SLCs to show the availability of templates for the potential structure-based modeling studies. Chapter 4 gives a comprehensive overview of the individual ligand- and structure-based modeling approaches applied to uptake SLC transporters. The focus further lies in bridging the gap between the two approaches to gain an in-depth understanding of the crucial molecular determinants of transporter-ligand interactions. Chapter 5 covers the topic of modeling transporter selectivity at different pharmacological barriers. Illustrative use cases from the field of SLC transporters discussed herein were collected from literature. Chapter 6 opens up novel perspectives and challenges of computer-based studies on SLCs. Specifically, ensemble docking strategies are debated herein. Efficient sampling of protein conformational space is addressed accordingly. To conclude this review article, a

computational pipeline combining different molecular modeling methods with state-of-the-art data science approaches is proposed. The graphical abstract summarizes the major aspects of the review article (Figure 2.1)

B. Zdrazil conceived the design of the review article. A. Tuerkova gathered literature and performed protein structure predictions. A. Tuerkova and B. Zdrazil analyzed evolution of computer-based methods collected from literature. The manuscript was written by the contribution of both authors.

Reprinted with permission from Elsevier.

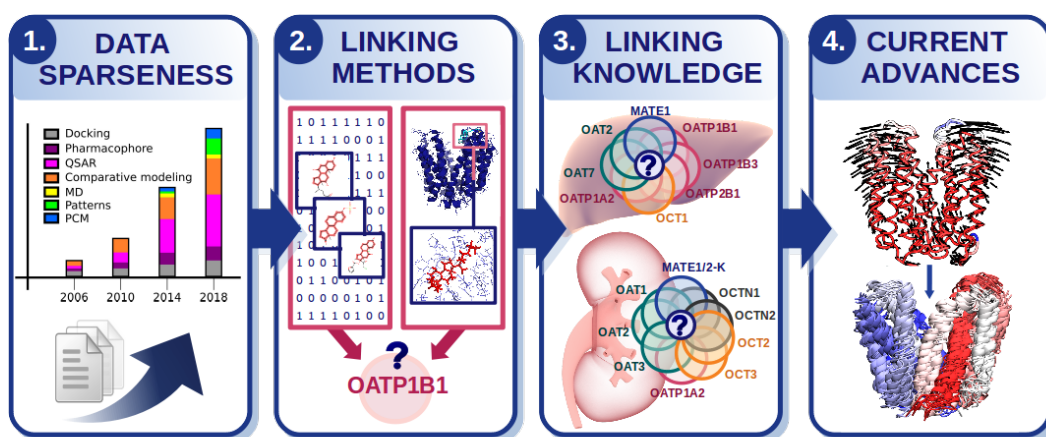


FIGURE 2.1: Graphical abstract of the review article. (1) Application of different *in silico* modeling methods over time. (2) Concrete examples of combining multiple methods to study compound-target interactions. (3) Application of different approaches to study selectivity profiling across the group of related transporters. (4) Innovative approaches are described here as a future outlook (e.g., ensemble docking in conjunction with normal mode analysis).



COMPUTATIONAL
AND STRUCTURAL
BIOTECHNOLOGY
JOURNAL

journal homepage: www.elsevier.com/locate/csbj



Mini Review

Current Advances in Studying Clinically Relevant Transporters of the Solute Carrier (SLC) Family by Connecting Computational Modeling and Data Science

Alžběta Türková, Barbara Zdrazil *

Department of Pharmaceutical Chemistry, Division of Drug Design and Medicinal Chemistry, University of Vienna, Althanstraße 14, A-1090 Vienna, Austria

ARTICLE INFO

Article history:

Received 23 November 2018

Received in revised form 28 February 2019

Accepted 1 March 2019

Available online 8 March 2019

Keywords:

SLC transporters

Uptake transporters

Molecular modeling

Drug-drug interactions

Selectivity

Conformational sampling

Fold-recognition

Open data

ABSTRACT

Organic anion and cation transporting proteins (OATs, OATPs, and OCTs), as well as the Multidrug and Toxin Extrusion (MATE) transporters of the Solute Carrier (SLC) family are playing a pivotal role in the discovery and development of new drugs due to their involvement in drug disposition, drug-drug interactions, adverse drug effects and related toxicity. Computational methods to understand and predict clinically relevant transporter interactions can provide useful guidance at early stages in drug discovery and design, especially if they include contemporary data science approaches. In this review, we summarize the current state-of-the-art of computational approaches for exploring ligand interactions and selectivity for these drug (uptake) transporters. The computational methods discussed here by highlighting interesting examples from the current literature are ranging from semiautomatic data mining and integration, to ligand-based methods (such as quantitative structure-activity relationships, and combinatorial pharmacophore modeling), and finally structure-based methods (such as comparative modeling, molecular docking, and molecular dynamics simulations). We are focusing on promising computational techniques such as fold-recognition methods, proteochemometric modeling or techniques for enhanced sampling of protein conformations used in the context of these ADMET-relevant SLC transporters with a special focus on methods useful for studying ligand selectivity.

© 2019 The Authors. Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Contents

1. Introduction: Clinically Relevant Transporters of the SLC Family and Implications in Drug Discovery	390
2. Molecular Modeling Approaches at a Glance	392
3. Data Sparseness as a Major Challenge in Computer-aided Drug Discovery & Implications to the Exploration of Uptake Transporters	393
4. Linking Ligand- and Structure-Based Modeling Methods for Studying Uptake Transporters	396
5. Selectivity Profiling: Linking Knowledge of Related Uptake Transporters	399
6. Accounting for Transporter Flexibility	400
7. Summary, Conclusions & Future Perspectives	402
Acknowledgements	402
References.	402

1. Introduction: Clinically Relevant Transporters of the SLC Family and Implications in Drug Discovery

Assessing a compounds' transporter pharmacology is an established paradigm in drug discovery and development, and efforts to document

clinically relevant interactions with transporters are systematically undertaken since at least a decade as demonstrated by the White Paper from the International Transporter Consortium from 2010 [111]. Such transporters broadly cover members of the ATP-binding cassette (ABC) transporter and the solute carrier (SLC) transporter superfamilies. Reviews covering the broader topic of modeling approaches on the SLC transporters are available from Colas et al. and Schlessinger et al. [20,102,103]. In this review, we are focusing on members of the

* Corresponding author.

E-mail address: barbara.zdrazil@univie.ac.at (B. Zdrazil).

Table 1

Tissue expression profiles of clinically relevant SLC transporters: predominant organs/tissues of expression are written in bold.

Human Transporter	Tissues of predominant expression	Reference
OCT1	Liver (sinusoidal membrane)	Motohashi, H., & Inui, K. I. [84]. AAPS J, 15(2), 581–588.
OCT2	Kidney (basolateral membrane)	Motohashi, H., & Inui, K. I. [84]. AAPS J, 15(2), 581–588.
OCT3	Placenta, testis, brain, lung, intestine etc.	Motohashi, H., & Inui, K. I. [84]. AAPS J, 15(2), 581–588.
OCTN1	Kidney (brush-border membrane), skeletal muscle , etc.	Motohashi, H., & Inui, K. I. [84]. AAPS J, 15(2), 581–588.
OCTN2	Kidney (brush-border membrane), liver , heart , etc.	Motohashi, H., & Inui, K. I. [84]. AAPS J, 15(2), 581–588.
MATE1	Kidney (brush-border membrane), liver (canalicular membrane), muscle, etc.	Motohashi, H., & Inui, K. I. [84]. AAPS J, 15(2), 581–588.
MATE2-K	Kidney (brush-border membrane)	Motohashi, H., & Inui, K. I. [84]. AAPS J, 15(2), 581–588.
OAT1	Kidney , skeletal muscle, brain and placenta	Roth et al. [97]. Br J Pharmacol, 165(5), 1260–1287.
OAT2	Liver , kidney	Roth et al. [97]. Br J Pharmacol, 165(5), 1260–1287.
OAT3	Kidney , brain	Roth et al. [97]. Br J Pharmacol, 165(5), 1260–1287.
OATP1A2	Brain, small intestine, kidney, testes, lung, Liver (cholangiocytes)	Roth et al. [97]. Br J Pharmacol, 165(5), 1260–1287.
OATP1B1	Liver (basolateral membrane)	Roth et al. [97]. Br J Pharmacol, 165(5), 1260–1287.
OATP1B3	Liver (basolateral membrane)	Roth et al. [97]. Br J Pharmacol, 165(5), 1260–1287.
OATP2B1	Liver (basolateral membrane), intestine, placenta, heart, mammary gland, brain	Roth et al. [97]. Br J Pharmacol, 165(5), 1260–1287.

SLC family which were identified to have major implications in the pharmacokinetics of drugs. Mostly these transporters are uptake transporters, only MATE1 and MATE2K - the Multidrug and Toxin Extrusion transporters (MATEs; belonging to SLC47A subfamily) - are representatives of efflux pumps that transfer substances out of cells (mainly expressed in kidney and liver cells). The other discussed transporters herein are uptake transporters belonging to the families of Organic Anion Transporting Polypeptides (OATPs; SLCO), Organic Anion Transporters (OATs; SLC22A), Organic Cation Transporters (OCTs; SLC22A), and Organic Carnitine Transporters (OCTNs; SLC22A).

The tissue expression of the different SLC members discussed in this review is quite heterogeneous and was already described extensively by others (summarized in Table 1) [84,97]. However, predominant organs of expression include liver, kidney, and to a lesser extent intestine. Pharmacological barriers in these organs are crucial for absorption, metabolism, distribution, and excretion of drugs and any impairment in the function of one of these transporters can therefore lead to adverse effects or toxicity by altered pharmacokinetics of the drug or food ingredient administered. For instance, treatment with the immunosuppressant cyclosporine can result in statin-induced myopathy when administered at the same time since statins are known to be substrates for OATP1B1. Inhibition of OATP1B1 by cyclosporine therefore leads to increased plasma concentration of different statins [86].

OATP1B1 and OATP1B3 are exclusively expressed in hepatocytes. Due to their high sequence similarity (~80%) they transport common substrates and are inhibited by common inhibitors. Nevertheless, there are some compounds (e.g. pitavastatin for OATP1B1 [41], cholecystokinin octapeptide CCK-8 for OATP1B3 [48]) which are selectively binding to only one of the two transporters with high affinity.

OATP2B1 is ubiquitously expressed, with highest expression levels at basolateral membranes of hepatocytes [97]. Due to its more distant relationship to OATP1B1 and OATP1B3 (~30% sequence identity to both), also substrate and inhibitor profiles are overlapping to a lesser extent [55,113].

Expression of OATP1A2 in liver cells has been observed in cholangiocytes, possibly accounting for the re-uptake of xenobiotics from bile [67]. In addition, common expression of OATP1A2 and OATP2B1 at the luminal membrane of intestinal absorptive cells is potentially implicated in drug-food interactions [28]. Specifically, it has been found that components of fruit juices (especially naringin) cause the inhibition of OATP1A2, which in turn affects oral bioavailability of fexofenadine (substrate of OATP1A2) [7].

OAT1, and OAT3 are expressed at the basolateral membrane of kidney cells, pointing to their concerted interplay in the uptake of substrates and related drug-drug interactions (DDIs) [16,29,44]. For example, it has been found that probenecid inhibits methotrexate transport by OAT1 and OAT3 [88]. OCT2 is expressed at the luminal

membrane of kidney cells [37] and its organ of highest expression is liver (rather than kidney).

OCTN1 and OCTN2 are highly expressed in kidney as well [76,124]. OCTN1 and OCTN2 are associated with several pathologies, such as inflammatory bowel disease, primary carnitine deficiency, diabetes, neurological disorders, and cancer [94]. Interestingly, in case of unresectable gastrointestinal stromal tumors treated with imatinib therapy, polymorphisms in OCTN1 and OCTN2 are associated with a prolonged time to progression in GIST patients receiving imatinib therapy [4]. In a different study, OCTN1-mediated uptake of cytarabine into tumor cells in a cohort of acute myeloid leukemia (AML) patients could be related to reduction in the development of resistance to chemotherapy [27].

Human MATE proteins have been shown to be widely distributed across different body tissues, including liver, skeletal muscles, testis, and kidney. Efflux- and uptake-activity of human MATEs have been observed to be pH-dependent [66]. Nevertheless, in renal cells MATE1 serves as an efflux pump, mediating the transport of the substances from kidney into urine. The OCT2-MATE1 interplay in uptake and effluxing compounds in the kidney shows that also little related transporters can be conjointly involved in clinically relevant DDIs. MATE2-K is the human kidney-specific MATE2 active splice variant of MATE2 [84]. In addition, the ubiquitously expressed variant of MATE, MATE2-B, has also been detected. Nevertheless, this variant probably does not have any functional activity [76].

For being able to assess a compounds' risk to interact with any of these ADMET-relevant transporters, having predictive in silico models for all of them at hand would be useful at an early stage in the drug discovery pipeline. As will be discussed in the following sections, such efforts are complicated by limitations in data availability in the domain of ADMET-relevant SLC transporters. Other challenges include the overlapping substrate or inhibitor specificities across these transporters which are arising from the promiscuous nature of these proteins [111], as well as from the substrate/inhibitor promiscuity.

In this review article, data science and computational modeling approaches for unravelling ligand-transporter interactions for the important class of clinically relevant SLC transporters are discussed. In chapter 2, a general overview on important ligand- and structure-based modeling methods which are traditionally being used in computer-aided drug discovery is provided. Next, challenges arising from data sparseness on both the ligand and protein side in the field of clinically relevant SLC transporters are discussed (chapter 3). Chapter 4 is providing more details about different ligand-based (LB) and structure-based (SB) modeling methods that have been used in the context of clinically relevant SLC transporters and discusses insights that were delivered by the different studies. The emphasis is further put on the combination of LB and SB methods to provide a more comprehensive picture of ligand recognition. In chapter 5, studies and different

methods providing insights into transporter selectivity are highlighted. Finally, the challenge of including information on transporter flexibility into the SB-modeling procedure is discussed (chapter 6) and some recent developments in the field are highlighted (such as the use of elastic network models for conformational sampling). In conclusion, a computational workflow for studying ligand interactions with clinically relevant SCL transporters is proposed, interconnecting ligand data integration, data curation and analysis, as well as LB and SB modeling techniques in order to come up with binding mode hypotheses.

2. Molecular Modeling Approaches at a Glance

Computational approaches have become a standard paradigm in area of preclinical drug discovery [78]. Molecular modeling is an interdisciplinary field that incorporates theoretical concepts and efficient computational algorithms to study chemical phenomena [65]. The main idea is to use approximative mathematical models while being able to predict behaviour of chemical systems as closely as possible [129]. In the field of computer-aided drug design (CADD), molecular modeling methods are generally divided into two categories: (1) Ligand-based (LB) and (2) Structure-based (SB) modeling. LB modeling is also known as 'indirect drug design', since in the modeling approach protein information is not taken into account. On the other hand, SB molecular modeling techniques enable to study protein-ligand complexes at the molecular level. SB drug design is also termed 'direct drug design' since in these cases 3D information of the target protein or a structural model built on basis of (phylogenetically or structurally) a related protein template is included into the modeling process.

In terms of LB modeling, a system of interest is represented mostly as a statistical model incorporating knowledge on the associated compounds (substrate or inhibitor data) possessing experimentally determined activities against a particular target. Compound series can be used to derive an abstract representation of important molecular features being crucial for a binding event. This approach is known as LB pharmacophore modeling and it is especially useful to overcome difficulties in aligning and searching for structurally little related compounds which persist similar in terms of chemical features (so-called "scaffold-hopping" concept) [45,128]. Another widely used approach is Quantitative Structure-Activity Relationship (QSAR) which is based on correlating physico-chemical properties (or other representations of the chemical structure, such as molecular fingerprints of compounds) to their biological activity [23]. QSAR modeling and related qualitative approaches (such as binary classification models) have currently become especially popular due to the observed increase of the deployment of machine learning and deep learning algorithms in CADD. Especially classification models (models being trained to distinguish compounds into e.g. 'active' and 'inactive' class) can subsequently be used for virtual screening of chemical databases with the aim to detect new (potentially active) compounds. Qualitative models can also aid in the interpretation of ligand profiles which might trigger a compound's activity against a particular target [113]. Beyond the conventional SAR techniques, 3D-QSAR modeling provides a natural extension to the classical QSAR formalism by superimposing 3D ligand structures to retrieve 3D-based ligand features [116].

SB molecular modeling approaches are significantly hampered by the limited number of experimentally resolved structures for membrane transporters. So far (updated on January 2019), the number of the available crystal structures for membrane proteins in Protein Data Bank (PDB) did not exceed ~3.5% of all deposited entries. Such a small fraction of resolved membrane proteins is primarily caused by problems when overexpressing membrane proteins in bacteria [101], as well as by the obstacles accompanying crystallization of membrane proteins [89]. These include, for example, finding optimal conditions for crystallization [87], as well as difficulties to account for a membrane-like environment which in turn is being crucial for crystallizing the native form of membrane proteins. However, substantial progress has been

made by e.g. improving protein engineering to increase the stability of a protein of interest [110]. The latter issue is commonly solved by using micelle detergents to solubilize membrane proteins for purification purposes [106]. Except for X-ray crystallization, solid-state NMR techniques can be used to resolve structure of membrane proteins [63]. In this methodology, membrane nanodiscs have successfully been applied as membrane-mimetics [98]. If the crystal, NMR, or recently also Cryo-EM structure of a respective target is not available, 3D structural models can be built "from scratch" upon basic thermodynamic principles [39]. These methods are known as *ab initio* (de novo) structural predictions and are rather rarely used, mainly due to the challenges to efficiently sample the whole conformational space at a feasible time scale [51]. In most of the cases, 3D structural models are created on basis of sequence-homologous proteins with known structure ("homology modeling" [52]), or more evolutionary distant proteins with conserved fold ("fold-recognition methods" or "protein threading" [75]). Fitting of a target sequence onto the 3D coordinates of a sequentially- or structurally- related protein ("template") is further accompanied by a global geometry optimization to satisfy structural restraints imposed by internal coordinates, as well as local optimization to remove steric clashes and/or reduce the noise caused by poorly modeled side chain rotamers [62,99,119]. This procedure is commonly applicable for the transmembrane core of membrane proteins. However, it might not be sufficient for modeling intra- and extra-cellular domains, which are mostly consisting of large loop regions [32]. Loop modeling is a non-trivial step in protein structure prediction, since the loops are intrinsically disordered regions, often requiring enhanced conformational sampling [33]. In addition, loops are usually indeterminate parts in the crystallization process due to their high conformational flexibility and thus low electron density in X-ray diffraction patterns. Homology modeling is subsequently limited by incomplete sequence alignment because of missing loop regions. Modeling of extremely short segments (<3 amino acids) can be satisfied by geometrical constraints of their bond lengths and angles. Template-based loop construction by using a database of known structural fragments (not necessarily with identical sequence) is a common modeling approach for medium-size loops (~10 amino acids). In general, energy minimization combined with MD and simulated annealing is advisable to refine modeled loops. For longer loops (~25 amino acids), de novo coarse grained modeling methods (employing e.g. united residue models) have successfully been applied [49].

Furthermore, molecular dynamics (MD) simulations with enhanced sampling techniques can additionally be integrated to the 3D model building procedure especially to refine low-confidence regions, such as flexible loops (as discussed above) and less structurally-conserved parts of the protein [34]. When performing MD simulations on membrane proteins, one should also account for the substantial role of phospholipid membranes which spatially restrain a protein's 3D structure [82]. The computationally less demanding approach is to treat membrane environment as a mean-field continuum model which replaces explicit protein-lipid interactions by effective interactions being a priori included in force-field parameters of a membrane-protein system [31]. These approaches have become particularly useful for the simulations of large time scale events, such as protein folding, albeit for the cost of lacking all-atom representation of a simulated system. The lack of high-resolution accuracy by using implicit membrane models can be corrected by the explicit representation of lipid bilayers in the simulation box. However, running MD simulations with explicit lipid bilayers might be unfeasible for biologically relevant time scales.

Quality assessment of 3D protein models is required to distinguish the native protein structure from the physically non-relevant states [40]. For this purpose, several scoring functions, including statistics-based, knowledge-based, physics-based, or their combinations, have been developed. To give an example, ProQM is a statistics-(learning-) based method using Support Vector Machines (SVM) models which are trained on the known structures to predict correct structural

features of membrane proteins, such as membrane topology or conserved structural motifs [95]. Another example is C-score which estimates the confidence of target-template alignments based on the fold-recognition methods [127]. TM-score function is a metric of 3D similarity between two proteins when performing structural alignment [131].

The next step in SB modeling is to apply docking algorithms to iteratively search for preferred orientations of a ligand molecule relatively to the protein binding site(s) [68]. Subsequently, protein-ligand complexes can be ranked on basis of scoring functions (knowledge-based [85], physics/force field-based [35], machine-learning [2], and/or empirical [12]), estimating the likelihood of all possible binding poses with respect to the energetically (un)favorable intermolecular interactions. Molecular docking can be performed by using different strategies, ranging from rigid docking, where the protein structure is kept fixed and only the ligand's conformational space is explored [9], to induced-fit docking, where the protein local backbone movements are allowed to adjust the proper ligand-protein binding pose [117]. More sophisticated and computationally demanding methods are treating the whole protein structure as flexible [123]. Docking screens are typically evaluated on basis of score accuracy (by comparing the predicted binding affinities to the experimental ones, if available [8]), enrichment factor (by checking if the docking screens are able to discriminate between known actives and known inactives/decoys [46]), or on basis of prospective validation (by measuring e.g. IC50 values [47]). Traditional docking approaches can be complemented by more accurate free energy calculations not only to estimate absolute free energy of binding [24], but also to study ligand selectivity profiles [3]. Approximations of binding free energies are given by methods like MM/GBSA (Molecular Mechanics/Generalized Born Surface Area), as originally described by Kollman et al. [59]. However, free energy perturbation (FEP) provides better estimates of free energy of ligand binding, and it offers the possibility to directly evaluate the impact of mutated functional groups of the ligand from the energetic point of view [118]. On the protein side, biochemical mutational studies can be informed by docking exercises and vice versa as demonstrated for the interleukin 8-gene [22].

Following the same strategy as in case of LB pharmacophores, SB pharmacophore models can be created upon projection of the important pharmacophoric features of a target-ligand complex to an abstract representation [107]. In the recent past, new techniques combining LB and SB approaches have become popular. For example, proteochemometric modeling (PCM) can outperform traditional QSAR modeling by simultaneous evaluation of the similarity of ligands and targets [115]. The great advantage here is the possibility to extrapolate the activities of known ligands against known targets to novel targets without knowing their 3D structure.

3. Data Sparseness as a Major Challenge in Computer-aided Drug Discovery & Implications to the Exploration of Uptake Transporters

Chapter 2 provides an overview of main computational techniques which are traditionally being used in the drug discovery pipeline. When it comes to investigations on clinically relevant SLC transporters, however, the direct application of above-mentioned methods is far from being trivial. The main obstacle hampering computational studies is caused by data sparseness on both ligand and protein levels in the domain of uptake transporters. In addition, the promiscuous nature of uptake transporters being able to bind both endogenous compounds and xenobiotics [56], considerable transporter flexibility which accompanies translocation processes (such as “rocker-switch” mechanism as proposed for Major Facilitator Superfamily members) [120], as well as the likely existence of multiple binding sites [43,70,74,125], turns all modeling efforts into even more challenging tasks.

From a ligand's perspective, modeling studies are strongly impeded by the inconsistent and mostly insufficient number of high-quality bioactivity data which is spread over different data sources in the open

domain. For LB studies on uptake transporters, single-point percentage inhibition data has often become the only source for QSAR modeling [1,14,19,55,58,60,122]. Such models trained on single-point inhibition data can achieve high accuracies, particularly if the measurements were retrieved following the same experimental protocols, as demonstrated by e.g. the classification models by Karlgren et al. [55] (with accuracies between 73% and 92% for the different models). However, even in the case of almost identical assay protocols, other parameters such as the substrate concentration might to a significant extent change the final value of the read out (e.g. percentage inhibition value). In case of the studies by Ahlin et al. [1] and Chen et al. [19] which both measured reuptake inhibition for OCT1 inhibitors by using the same probe substrate (4–4-dimethylaminostyryl-*N*-methylpyridinium, abbreviated as ASP⁺), a different substrate concentration (2 μ M in Chen et al. and 1 μ M in Ahlin et al., respectively) as well as inhibitor concentration (20 μ M in Chen et al. and 50 μ M in Ahlin et al.) led to a percentage of around 14% of all overlapping compounds (measured in both papers) being differently classified by the different studies (by using a cutoff of 50%). Obviously, the study by Chen et al. was much more rigorous in assigning the label “inhibitor”, since many conflicting annotations with Ahlin et al. turned out to be rather false negatives (data not shown).

In most of these cases, the activity cut-off for binary classification into actives and inactives was set to 50% [1,19,55,60,122]. In some studies, however, the authors applied more stringent activity criteria for setting a threshold by removing weak actives from the data set, e.g. in the study by van de Steeg et al. [114]. Here, the activity cut-off for OATP1B1 inhibitors was defined as $\geq 60\%$, while OATP1B1 non-inhibitors were defined as $\leq 40\%$. It concludes that all “grey zone” data points (the weak actives) in the range (40;60) were excluded from the dataset. On one hand, removal of weak actives can be beneficial to reduce the noise caused by the variations in the experimental measurements [114]. On the other hand, complete exclusion of weak actives might decrease the applicability domain of such models, e.g. when attempting to predict the activity of compounds which are structurally closely related to those from the “grey zone”. For establishing LB pharmacophore models for MATE1 [126] and OCT2 inhibitors [125], classification of inhibition/substrate data into binary classes was done via manual literature searches with the aim to extract recommended activity thresholds proposed by the authors of primary literature sources. For deciding upon the cutoffs for percentage inhibition data in order to generate predictive classification models for OATP1B1, OATP1B3, and OATP2B1 inhibition from diverse data source we have chosen the same strategy [113].

Structurally and pharmacologically distinct compounds can also be partitioned into discrete clusters based on their molecular features, as applied for e.g. the classification of OCT2 inhibitors/non-inhibitors [58]. After the clustering step, different activity cut-offs (inhibitory effect $\geq 50\%$ or $\geq 75\%$) were probed to examine which clusters contained the highest fraction of the inhibitors based on the applied cut-off. Specifically, by applying the more stringent cut-off ($\geq 75\%$) the inhibition patterns of identified clusters became more pronounced. Distinct OCT2 clusters were subsequently used for deriving several independent SAR analyses to explain complementary inhibitory mechanisms of OCT2 inhibitors.

Full dose-response curve data has been exploited to much lesser extent. Examples include IC50 measurements for LB studies on MATE1, MATE2-K [5], and OCT2 [122], Km values for OATP1B1 substrate pharmacophore models [17], as well as Ki values for pharmacophore modeling of OCT2 stereoselective binding [79].

If the amount of bioactivity data is not sufficient for model development, direct usage of categorical annotations, such as “substrate”, “non-substrate”, “inhibitor”, or “non-inhibitor”, can be applied. For this purpose, Drugbank (containing a collection of marketed or FDA-approved drugs [121]), or Metrabase (primarily containing substrates for OCT1, OATP1A2, OATP1B1, OATP1B3, and OATP2B1 [72]) can serve as rich, open sources for LB modeling. For example,

Table 2

Number of unique compounds/bioactivities from open domain databases. In case of Metrabase and IUPHAR, only numerical bioactivity values have been extracted from the data bases (categorical annotations were discarded).

Transporter	UNIPROT ID	CHEMBL	Metrabase	IUPHAR	Transportal	Total number of unique compounds
OCT1	O15245	290/437	230/607	1/1	44/86	307
OCT2	O15244	126/221	No entries	1/1	67/120	144
OCT3	O75751	37/44	No entries	1/1	28/44	54
OCTN1	Q9H015	26/33	No entries	No entries	11/20	26
OCTN2	O76082	67/96	No entries	No entries	6/10	68
MATE1	Q96FL8	70/139	No entries	3/4	31/55	86
MATE2-K	Q86VL8	44/55	No entries	3/4	23/46	60
OAT1	Q4U2R8	111/205	No entries	1/1	74/132	123
OAT2	Q9Y694	32/39	No entries	No entries	32/39	50
OAT3	Q8TCC7	113/180	No entries	No entries	68/102	131
OATP1B1	Q9Y6L6	1993/2566	307/752	1/1	61/139	2052
OATP1B3	Q9NPD5	1972/2469	249/408	3/3	45/95	2015
OATP1A2	P46721	70/100	54/96	2/2	17/24	75
OATP2B1	O94956	252/392	232/461	0/0	21/46	294

substrate annotations from Metrabase were retrieved for QSAR and PCM modeling to predict OCT1, OATP1A2, OATP1B1, OATP1B3, and OATP2B1 substrates [104]. It has to be pointed out, that using manual activity annotations for binary classification modeling seems to be quite error-prone since it is not clear how data curators decided upon assignment of activity annotations in certain cases. As demonstrated in our current LB studies on hepatic OATPs [113], an extensive comparison of Metrabase annotations for OATP1B1, OATP1B3, and OATP2B1 (non-)substrates and (non-)inhibitors with numerical bioactivity measurements from ChEMBL revealed activity misclassifications for Metrabase data up to 74%. On the other hand, categorical annotations can still be utilized in developing accurate computational models when e.g. performing selective fusion of more independent classifiers and therefore increasing the confidence in the model's predictability [104].

With increasing efforts of making bioactivity data publicly available to the scientific community, new challenges are arising. Nowadays, it is no longer only access to data but the proper usage of data which can provide a competitive advantage in drug discovery. Finally, filtering out high-quality data is essential as well as making use of the possibilities to interconnect different types of data, including data originating from diverse sources. In the light of those efforts, data analytics platforms for creating automated data workflows, such as the Konstanz Information Miner (KNIME [11]) or Pipeline Pilot [108], have become particularly useful. Notably, integrative data mining (i.e., data fusion from different sources) can enrich existing data sets by not only the size in enumerated compounds, but also by obtaining novel scaffolds which can lead to an expansion of the available chemical space as demonstrated recently for hepatic OATPs [113]. Moreover, merging data from multiple independent bioactivity measurements (Km, IC50, EC50, Ki, percentage inhibition data) for a single compound can significantly increase the confidence of bioactivity data. When it comes to QSAR modeling, it is usually not recommended to mix compound data originating from different bioactivity end-points [53,61]. On the other hand, binary classification (e.g. separating substrates from non-substrates) should be independent from a certain assay or experimental protocol used [80]. Another benefit when considering multiple bioactivity measurements is the ability to rationally decide upon activity thresholds for binary label assignment (e.g. active/inactive) by studying the distribution of bioactivities within the data set [113].

As recently demonstrated [113], above mentioned pipelining tools are quite handy for semi-automatically fetching ligand data from different open data sources, such as ChEMBL [10], UCSF-FDA Transportal [83], IUPHAR [90], and Drugbank [121] and Metrabase [72]. It has to be emphasized that ligand data originating from different sources might be highly inconsistent with respect to their structural format used. To overcome this issue, applying a standardization protocol, as

e.g. by Atkinson (available at <https://wwwdev.ebi.ac.uk/chembl/extra/francis/standardiser/>), has proven to be useful for mapping data from different sources.

From Table 2 and Table 3 it becomes clear that the most comprehensive data set is currently available for OATP1B1 and OATP1B3, with ChEMBL being detected as the most prominent source (1993 and 1972 unique compounds, respectively). This trend is reflected by the high number of LB modeling studies with a special focus on OATP1B1: QSAR/classification models [6,54,55,60,104,105,113,114], PCM [14,104], and LB pharmacophore model [17].

In addition to this quite restricted availability of compound bioactivity data for uptake transporters which limits over all the applicability domain of respective LB models, SB modeling efforts are strongly impeded by the lack of crystal structures in this domain. Only MATE1 bacterial templates are available (NorM from *Vibrio cholerae* 21, pdb id: 3mkt; pMATE from *Pyrococcus furiosus*, pdb id: 3vvn). Both of the available crystal structures share approximately 24% identity with human MATE1. Except for this case, there are no available homologs for the other clinically relevant SLC transporters. These drawbacks can be diminished by the use of fold-recognition methods to search for structurally related analogues with conserved fold. The rationale behind is that a protein's secondary structure should have been conserved to a higher extent during evolution than its amino acid sequence [52]. An overview of available crystal structures which were predicted as potentially useful templates for SB modeling of ADMET relevant SLC transporters is provided in Fig. 1. Predictions were performed by using pGenThreader prediction server [71]. It appears interesting that all predicted templates (except for chain L from respiratory complex I) belong

Table 3

Number of unique compounds with categorical activity annotations from the open domain.

Transporter	UNIPROT ID	Metrabase	Drugbank	Total number of unique compounds
OCT1	O15245	504	69	470
OCT2	O15244	No entries	56	56
OCT3	O75751	No entries	29	29
OCTN1	Q9H015	No entries	27	27
OCTN2	O76082	No entries	54	54
MATE1	Q96FL8	No entries	6	6
MATE2-K	Q86VL8	No entries	1	1
OAT1	Q4U2R8	No entries	111	111
OAT2	Q9Y694	No entries	36	36
OAT3	Q8TCC7	No entries	77	77
OATP1B1	Q9Y6L6	375	75	385
OATP1B3	Q9NPD5	338	44	345
OATP1A2	P46721	111	61	107
OATP2B1	O94956	352	33	338

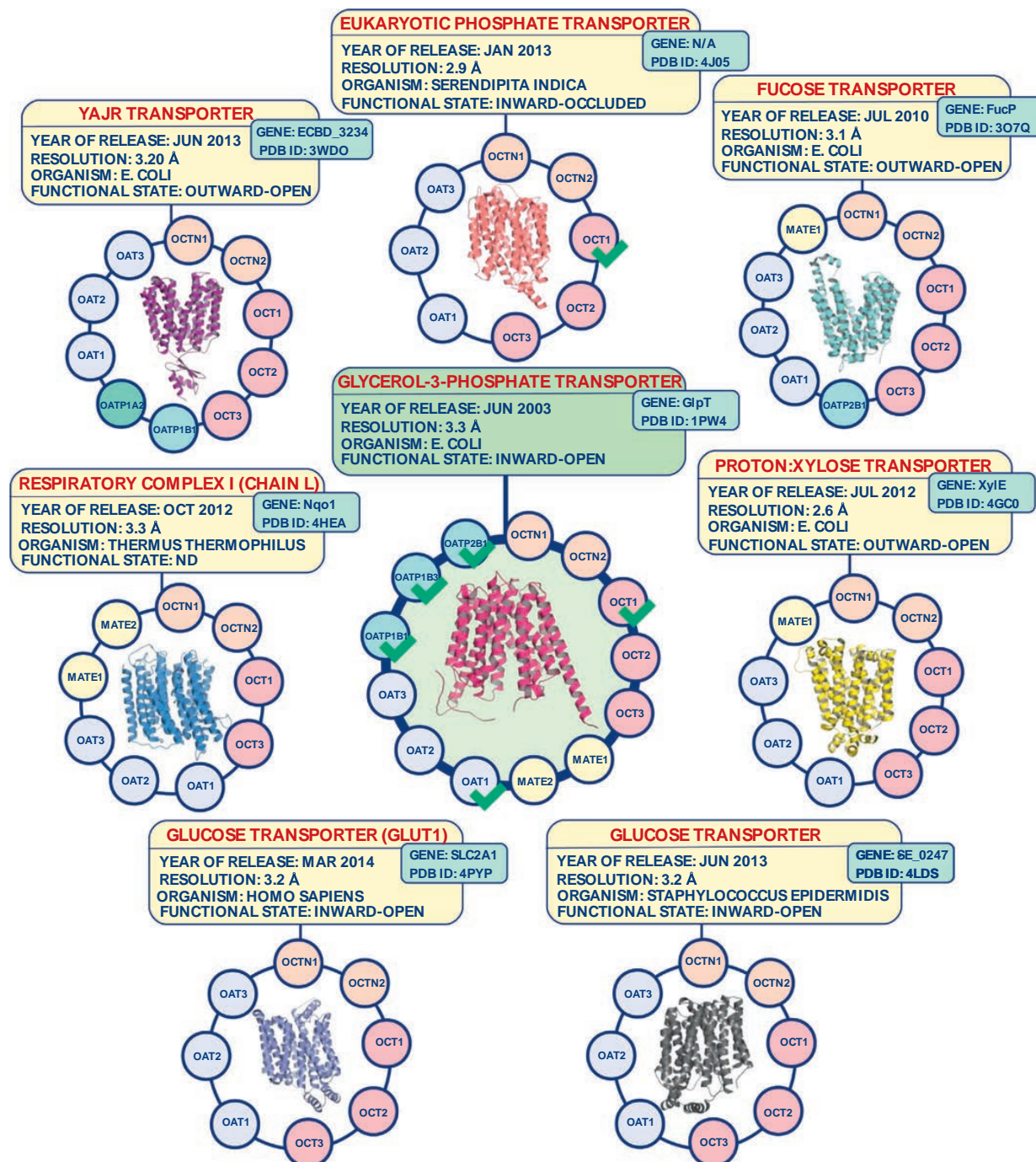


Fig. 1. Predicted structural templates for ADMET relevant SLC transporters by using pGenThreader ($p < 0.001$). Templates already used in structure-based modeling for a respective uptake transporter are indicated by a green check mark.

to the Major Facilitator Superfamily (MFS) [92]. All protein structures are built up of twelve transmembrane helices. It is interesting to note that sequence similarity among discussed transporters and the detected template structures is generally rather low (10–25%), which obviously slowed down the generation of comparative models on basis of these templates compared to other protein families (numbers of published studies per year including comparative modeling for

clinically relevant SLC transporters is depicted in Fig. 2). As visible from Fig. 1, different templates are reflecting different functional states of the transporters (inward-open, occluded, outward-open). Out of these templates, Glycerol-3-Phosphate transporter (pdb id: 1pw4) was the most abundantly used template, specifically for building computational models for OATP1B1 [69], OATP1B3 [36,73,77], OATP2B1 [57,77], OAT1 [93,112], and OCT1 [13]. The popularity of

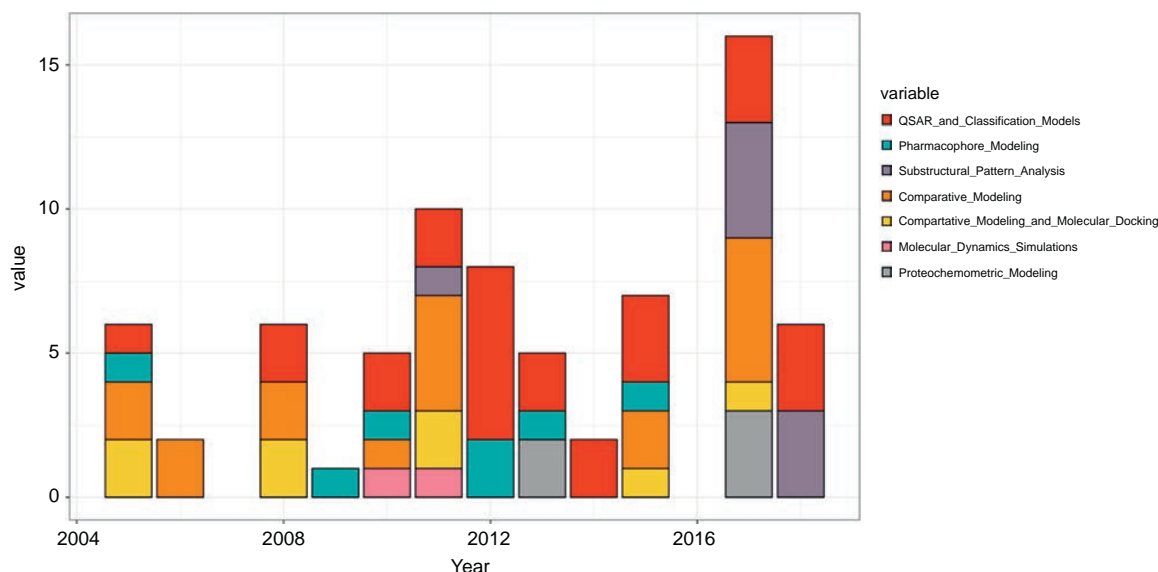


Fig. 2. Time evolution of different computer-based methods used in published in silico studies on clinically relevant SLC transporters. abscissae....year; ordinate....numbers of published studies.

Glycerol-3-Phosphate transporter as a structural template for this class of transporters can mainly be attributed to the fact that it appeared as the first available template in 2003.

4. Linking Ligand- and Structure-Based Modeling Methods for Studying Uptake Transporters

Natural increase in available ligand bioactivity data and structural templates also led to an increase of published computational studies on clinically relevant SLC transporters over the years. Fig. 2 shows the time evolution of different computational methods that have been used in the context of uptake transporter modeling. In general, we can observe an increase in in silico studies over the years for both LB and SB approaches. Both approaches have been used early on and we can observe peaks in the emergence of QSAR/classification models and new homology models in 2011/12. In 2017, one publication reported the establishment of new comparative models for OCT1, OCT2, OCT3, OCTN1, and OCTN2, which is also visible as a peak from Fig. 2. Interestingly, MD simulations were integrated only twice into the process of clinically relevant SLC transporter modeling, namely for OAT1 and MATE1 (in 2011/12). With the recent increase in available comparative models for uptake transporters, a revival of MD-based methods can be expected especially for performing enhanced conformational sampling of distinct transporter states in the future (see also chapter 5). Naturally, studies including PCM modeling appeared later in literature, since they incorporate information from the ligand and protein side. Thus, for such a technique to be effective it requires a minimum amount of ligand data and ideally knowledge on potential protein binding sites.

Also, from a static perspective, LB and SB methods have been used to approximately the same extend (e.g. 26 studies on QSAR/classification modeling; 26 studies including comparative modeling). If a fair amount of compound bioactivity data for a particular transporter is available (at least a few hundred unique compounds), LB approaches can deliver quicker and more comprehensive insights into important molecular features driving compound affinity towards a particular target than SB approaches. On the other hand, traditional SAR-based methods do not account for the hypothetical presence of distinct binding sites. Such drawbacks can be diminished by sub-structural pattern analyses such as studying pharmacological profiles of enriched scaffolds in the data

set. The hereby retrieved congeneric SAR series are likely to interact with the same binding pocket and differences in pharmacological activities are likely triggered by subtle modifications at the side chain level [113]. Thus, such compound series are useful collections for subsequent structure-based docking studies.

In addition, LB pharmacophore models can be useful in detecting important pharmacophoric ligand features which can complement SB docking studies into comparative models. An interesting pharmacophore-based modeling approach where in vitro and in silico ('IVIS') methods are combined has been adopted by Diao et al. to map pharmacophoric features of human OCTN2 inhibitors [25,26]. This methodology has subsequently been used also by Astorga et al. to study the inhibitory profiles of human MATE1 [5]. Specifically, the IVIS procedure aims to iteratively build 3D-pharmacophore models, further used for database screening and subsequent experimental testing of new hits. Afterwards, high-affinity detected compounds are used to rebuild initial pharmacophore hypotheses to perform another round of database screening, and so on. For human MATE1 inhibitors, a common-features pharmacophore has initially been developed by merging pharmacophoric features of both high- and low-affinity MATE1 inhibitors [25]. The aim of mixing high- and low-affinity compounds for building a single pharmacophore model is to detect the minimal essential properties which are crucial for the effective interaction with MATE1. The iterative procedure has finally led to a pharmacophore with two hydrophobic features, one H-bond acceptor and one ionizable (cationic) feature. It is noteworthy to mention that the generation of quantitative pharmacophores has been strongly dependent on the probe substrates used in the in vitro measurements. Specifically, when using 4-4-dimethylaminostyryl-N-methyl-pyridinium instead of 1-methyl-4-phenylpyridinium, the final pharmacophore model was composed of three hydrophobic features, two H-bond acceptors, and three excluded volumes, spatially arranged in strikingly different configuration from the original one. These findings suggest that human MATEs might contain multiple substrate (and potentially inhibitor) binding sites. Therefore, for developing a complete computational model of MATE inhibitory mechanisms it would be required to build up multiple pharmacophore models per distinct binding site. For OCTN2 inhibitors, the common pharmacophore model revealed three hydrophobic

Table 4

List of available ligand- and structure-based molecular modeling studies on uptake transporters discussed in this review.

Paper	Transporter	Method	Description	Results
Tanihara et al. [132]	MATE1	Binary classification modeling	Identification of key features for inhibitory mechanism	Cationic charge is crucial for MATE1 inhibition.
Diao et al. [25]	MATE1	Bayesian machine learning modeling	Identification of key features for inhibitory mechanism	Six-membered rings including nitrogen are important for MATE1 inhibition.
Astorga et al. [5]	MATE1	IVIS pharmacophore modeling	Iterative identification of new MATE1 inhibitors by using pharmacophore-based virtual screening	Two hydrophobic features, H-bond acceptor and cationic feature have occurred in final pharmacophore model for MATE1.
Zhang et al. [130]	MATE1	Structural model building, molecular dynamics simulation	Topology of human MATE1 transporter, stability of constructed structural model	Human MATE1 transporter consists of 12 TM which have a functional role; 13th TM is not required for the transport.
Wittwer et al. [122]	MATE1	Binary classification modeling	Identification of key features for inhibitory mechanism	Cationic charge, molecular weight, and lipophilicity are important features for MATE1 inhibition.
Xu et al. [126]	MATE1	Combinatorial pharmacophore modeling	Studying multiple inhibitory mechanisms of MATE1 inhibitors	Four different binding sites (two competitive, one non-competitive and one mixed inhibition binding site) were identified for MATE1.
Xu et al. [126]	MATE1	Structural model building, molecular docking	Elucidate the evidence of multiple binding sites from combinatorial pharmacophore model	Four different binding sites (two competitive, one non-competitive and one mixed inhibition binding site) were identified for MATE1.
Astorga et al. [5]	MATE2-K	IVIS pharmacophore modeling	Iterative identification of new MATE2-K inhibitors by using pharmacophore-based virtual screening	Two hydrophobic features, H-bond acceptor and cationic feature have occurred in final pharmacophore model.
Perry et al. [93]	OAT1	Structural model building	Identification of critical residues important for OAT1 transport function	Importance of aromatic amino acid at position 230 has been discovered.
Truong et al. [133]	OAT1, OAT3	QSAR modeling	Comparison of interactions of antiviral drugs with OAT1 versus OAT3	Number of H-bond donors (alcohols and amides) and total polar surface area have triggered a preferred inhibitory activity towards OAT1.
Tsigelny et al. [112]	OAT1	Molecular dynamics simulation	Investigations of dynamics events accompanying OAT1 transport	Tilting mechanism of two hemi-domains is crucial for the initialization of transport process.
Soars et al. [134]	OAT1, OAT3	QSAR modeling	Comparison of inhibitor features between OAT1 and OAT3	OAT1 and OAT3 inhibitors have statistically significant inhibitory profiles.
Bednarczyk et al. [135]	OCT1	LB pharmacophore modeling	Identification of important pharmacophoric features for OCT1 inhibition	Three hydrophobic and one positive ionizable feature are important features for OCT1 inhibition.
Moaddel et al. [79]	OCT1	LB pharmacophore modeling	Studying stereoselective recognition of OCT1 transporters	One positive ionizable feature, one hydrophobic, and two H-bond acceptor features are important for OCT1 inhibition.
Ahlin et al. [1]	OCT1	QSAR modeling	Identification of molecular features being important for inhibitory activity	H-bond donors, lipophilicity, cationic charge positively correlate with OCT1 inhibition.
Badolo et al. [6]	OCT1	QSAR modeling	Identification of molecular features being important for inhibitory activity	Topological polar surface area negatively correlates with OCT1 inhibition.
Shaikh et al. [104]	OCT1, OATP1B1, OATP1B3, OATP2B1	QSAR/PCM modeling, substructural analysis	Identification of important molecular features and structural fragments for substrate activity against reported transporters	Developed models were used for prediction of substrate propensity for blood-brain barrier transporters.
Dakal et al. [22]	OCT1, OCT2, OCT3, OCTN1, OCTN2	Structural model building	Multiscale structural models construction for OCT transporters	Constructed structural models for OCTs share close structural similarity with GLUT3 transporter (pdb id: 5c65).
Chen et al. [19]	OCT1	Structural model building, molecular docking	Identification of critical residues important for OCT1 activity; virtual screening for sake of detecting new OCT1 inhibitors	D474 is important for ligand binding; detection of two distinct binding sites in translocation channel.
Boxberger et al. [13]	OCT1	Structural model building, molecular docking	Identification of critical residues important for OCT1 activity	Identification of three distinct binding sites based on the presence of critical residues (W218, Y222, T226, I443, I447, Q475).
Kido et al. [58]	OCT2	QSAR modeling	Identification of molecular determinants for OCT2 inhibitors	Suggestion of multiple binding sites for OCT2 transporter.
Suhre et al. [136]	OCT2	2D-QSAR modeling, Comparative Molecular Field Analysis (CoMFA)	Identification of molecular determinants of OCT2 substrates and inhibitors	Hydrophobicity, steric factor, and number of rotatable bonds were identified as important features for OCT2 inhibition.
Wittwer et al. [122]	OCT2	QSAR modeling	Identification of molecular determinants of OCT2 inhibitors	Occurrence of both zwitterionic and basic functional groups is important for OCT2 inhibition.
Xu et al. [125]	OCT2	Combinatorial pharmacophore modeling	Studying multiple inhibitory mechanism of OCT2 inhibitors	Four distinct pharmacophore hypotheses, corresponding to the competitive inhibition (one hypothesis), non-competitive inhibition by occlusion (two hypotheses), and one mixed inhibition pattern, have been identified for OCT2 inhibitors.
Diao et al. [26]	OCTN2	IVIS pharmacophore modeling	Iterative identification of new OCTN2 inhibitors by using pharmacophore-based virtual screening	Three hydrophobic and one positive ionizable feature are important for OCTN2 inhibition.
Diao et al. [25]	OCTN2	IVIS pharmacophore modeling, Bayesian modeling	Iterative identification of new OCTN2 inhibitors by using pharmacophore-based virtual screening	Two hydrophobic features, one H-bond donor, and positive ionizable feature are important for OCTN2 inhibitors; aromatic and tertiary-amine groups have also been detected via Bayesian modeling.
Mandery et al. [73]	OATP1A2,	Structural models	Comparison of structural determinants for	K361 and K399 are highly conserved residues across

(continued on next page)

Table 4 (continued)

Paper	Transporter	Method	Description	Results
	OATP1B3, OATP2B1	construction	ligand activity among OATP family	OATP family; K361 is pointing towards the translocation pore; variable loop located within a translocation pore differs in terms of crucial residues for respective targets (R58 and S62 in OATP1B3, Q58 and P62 in OATP1A2, and S64 in OATP2B1).
Chang [17]	OATP1B1	LB pharmacophore modeling	Detection of pharmacophoric features for OATP1B1 substrates	Two H-bond acceptors and two or three hydrophobic features are important for OATP1B1 substrates.
Badolo et al. [6]	OATP1B1, OATP1B3	QSAR modeling	Identification of molecular features being important for OATP1B1/1B3 inhibition	Lipophilicity, polarity, lower base pKa, higher number of H-bond acceptors, and higher molecular weight correlate with OATP1B inhibition.
Soars et al. [105]	OATP1B1	QSAR modeling	Identification of molecular features being important for OATP1B1 inhibition	Low number of aromatic bonds (<7), lipophilicity, and hydrogen-bonding potential are important for OATP1B1 inhibition.
Karlgrén et al. [54]	OATP1B1	QSAR modeling	Virtual screening for detecting new OATP1B1 inhibitors	Lipophilicity, larger molecular weight, larger polar surface area
Karlgrén et al. [55]	OATP1B1, OATP1B3, OATP2B1	QSAR modeling	Comparison of molecular determinants for ligand activity among hepatic OATPs	Lipophilicity and polar surface area are general features for OATP inhibition; OATP2B1 inhibitors are less dependent on polarity than OATP1B1/1B3 inhibitors.
Van de Steeg et al. [114]	OATP1B1	Bayesian modeling	Identification of molecular features being important for OATP1B1 inhibition	Conjugated-bond systems, (hetero)cycles with acceptor/donor atoms inside or outside the rings, molecular weight, molecular surface area, lipophilicity, number of rings, number of rotatable bonds, number of H-bond acceptors are important for OATP1B1 inhibition.
Bruyn et al. [14]	OATP1B1, OATP1B3	PCM modeling	Comparison of molecular determinants for ligand activity for OATP1B1 and OATP1B3 transporter	Lipophilicity, absence of cationic charge, number of ringbonds, presence of an anionic functional group, molecular volume, and substantial number of H-bond acceptors are important for general OATP1B inhibition; low number of aromatic bonds correlates with OATP1B1 inhibition, whereas higher lipophilicity and moderate number of H-bond donors corresponds with OATP1B3 inhibition.
Kotsampasakou et al. [60]	OATP1B1, OATP1B3	QSAR modeling	Comparison of molecular determinants for ligand activity for OATP1B1 and OATP1B3 transporters; virtual screening to search for new OATP1B ligands	Number of H-bond donors and acceptors, LogP, molecular refractivity, topological surface area, molecular weight, number of rotatable bonds, topological radius, topological diameter, topological shape, global topological charge index, have been used to develop models for OATP1B1 and OATP1B3.
Türkova et al. [113]	OATP1B1, OATP1B3, OATP2B1	Substructural analysis, QSAR modeling	Comparison of molecular determinants for ligand activity among hepatic OATPs	Lipophilicity, molecular weight, number of atoms, molecular refractivity, and flexibility are important features for general OATP inhibition; OATP2B1 inhibitors tend to be more planar than OATP1B1/1B3 inhibitors.
Li et al. [69]	OATP1B1	Structural model construction, molecular docking	Exploring the importance of selected amino acids from TM2 on the uptake of Estrone-3-sulphate	D70 and F73 are involved in the interaction with substrates; two distinct binding sites (low- and high-affinity site) for Estrone-3-sulphate have been identified.
Hong et al. [137]	OATP1B1	Structural model construction, molecular docking	Exploring the importance of selected amino acids from TM11 on the uptake of prototypic substrates	Importance of negative charge at position 596 for OATP1B1 uptake.
Glaeser et al., [36]	OATP1B3	Structural model construction, molecular docking	Identification of important amino acids on OATP1B3 transport function	Importance of positive charge at position 41, importance of R580 residue on OATP1B3 transport.
Meier-Abt et al. [77]	OATP1B3, OATP2B1	Structural model construction, molecular docking	Comparison of important amino acids on OATP1B3 and OATP2B1 transport function	R181 might contribute to the OATP1B substrate specificity, while H579 is hypothesized to be crucial for OATP2B family; conservation of H-bonds patterns, as well as helix-breaking residues (proline and glycine patterns), have also been detected.
Gui and Hagenbuch [38]	OATP1B1, OATP1B3	Structural model construction, molecular docking	Comparison of important amino acids on OATP1B1 and OATP2B1 transport function	TM10 is pronounced to drive the differences between OATP1B1 and OATP1B3.
Khuri et al. [57]	OATP2B1	Structural model construction, molecular docking, QSAR modeling	Identification of molecular features being important for OATP2B1 inhibition; virtual screening for new OATP2B1 inhibitors	OATP2B1 inhibitors are lipophilic.

features and one positive ionizable feature [26]. This IVIS-based pharmacophore hypothesis was partially confirmed by the upcoming study on OCTN2, showing that two hydrophobic features, one H-bond donor, as well as one positively ionizable feature are likely driving OCTN2 inhibitory activity [25].

As already introduced in chapter 3, SB modeling is complicated by the lack of native family member templates. Exclusively human and rabbit MATE1 structural models have been constructed on basis of sequence-homologous templates retrieved by the BLAST algorithm

[126,130]. The first structural model for human MATE1 has been built upon the sequence similarity with NorM crystal structure from *Vibrio cholerae* (pdb id: 3mkt, 35.6% sequence similarity) [130]. After structural model generation, molecular dynamics (MD) simulation has been performed to test the stability of generated human/rabbit MATE1 structural model. In parallel, MD simulations have been performed for the NorM template structure to see if the conformational dynamics of NorM crystal and derived homology models remains conserved. A 50 ns long production MD confirmed overall stability.

Interestingly, several helices in the human MATE1 homology model (for example, TM6 and TM9) reoriented and assumed opposite tilt angles when compared to the template. Obviously, a partial closing of the translocation pore was happening for the homology model. This interesting use case highlights the fact that each transporter ortholog might possess its internal dynamics, which cannot always easily be captured by the template structure.

Other reported SB studies rely on the structural similarity with MFS members. For details describing studies for a respective uptake transporter, see Table 4. In the recent past, a multiscale approach for 3D structural modeling of seven human OCTs (OCT1, OCT2, OCT3, OCTN1, OCTN2, OCT6, and FLIPT1) has been published. This study [21]. This study [144] introduces a comprehensive modeling pipeline for tertiary structure prediction, starting from the comparative sequence alignment, and fold-recognition 3D model building combined with ab-initio modeling performed via I-TASSER [145]). In addition, post-translation modifications of functionally relevant structural motifs (e.g. phosphorylation, ubiquitination, and/or glycosylation sites) were predicted via bioinformatic tools (PhosphoSitePlus available at <http://www.phosphosite.org/> and NetNGlyc 1.0 server available at <http://www.cbs.dtu.dk/services/NetNGlyc/>). An integrative approach for 3D structure prediction is followed by the comprehensive evaluation of the modeled structures based on different metrics, including sequence identity between the target and template, query coverage, and consensus Z-score of the top threading programs. Structural analysis of generated models has revealed that the 3D structural models generated here share structural similarity with the human glucose transporter GLUT3 (pdb id: 5c65). Visual inspection of the obtained models further implies 2-pseudofold symmetry, as well as the hypothesis about two distinct functional states (inward- and outward-open). These observations were fully supported by the structural superposition of 3D generated models of OCTs with the GLUT3 transporter. This use case again demonstrates that phylogenetically unrelated transporters (since OCTs belong to SLC22A and GLUTs belong to SLC2A subfamily) can share the same fold.

Two very promising ways of integrating ligand and protein information are so-called combinatorial pharmacophores as well as PCM modeling approaches. Combinatorial pharmacophore modeling was first reported for OCT2 inhibitors in 2013 [125]. In general, this approach represents a multi-step combinatorial scheme to generate a set of diverse LB pharmacophore models including the available information about their binding modes: First, generated pharmacophore models with identical pharmacophoric features in a close spatial arrangement are grouped in order to reduce the large pool of potential hypotheses and a combinatorial approach is employed to test all possible combinations of different pharmacophore hypotheses. The main idea behind the combinatorial pharmacophores is to study how different pharmacophoric patterns are corresponding to (potential) multiple binding mode hypotheses of uptake transporters. For this purpose, different sub-categories of reference inhibitors are being used - (1) competitive inhibitors (i.e., binding to orthosteric binding site), (2) occluding inhibitors (i.e., noncompetitive inhibitors, occluding the substrate binding site and locking the conformation transformation of the transporter), and (3) allosteric inhibitors (i.e., modulating the transporter's function by binding to the different - allosteric - binding site). A use case on MATE1-OCT2 selectivity profiling is presented in chapter 5.

Proteochrometric modeling (PCM) is conceptualized as an advanced extension to the conventional QSAR-based modeling by simultaneous considerations of the similarity between multiple ligands and multiple targets [115]. PCM modeling can thus be categorized as a method at the interface between ligand- and structure-based modeling. The two-dimensional structural sequence information can be integrated into the PCM model either as a whole amino acid sequence, or the pre-selection of key residues (e.g. those occurring in the binding pocket and/or other conserved residues) can be performed. ([64,93]; [138]). Protein

sequences can then be reduced to a more abstract representation by calculating the Z-scales which are corresponding to the principal components of multi-property matrices combining different physico-chemical properties, such as lipophilicity, volume, and polarity for respective residues [64,115]. In case of uptake transporters discussed herein, PCM modeling was used for the investigations of structural determinants between OATP1B1 and OATP1B3 [14], as well as in a recent study by Shaikh et al. for investigating transporter substrates of OCT1, OATP1A2, OATP1B1, OATP1B3, and OATP2B1 [104].

5. Selectivity Profiling: Linking Knowledge of Related Uptake Transporters

Since uptake transporters are often co-expressed at pharmacological barriers and generally transport a wide variety of pharmaceutical agents, it is of medical interest to increase the understanding of the interplay of such related transporters. A prominent example are hepatic OATPs – OATP1B1, OATP1B3, and OATP2B1 – which are responsible for e.g. bile acids uptake (such as taurocholic acid), but also pharmaceuticals, hormones etc. into hepatocytes ([139], [140]). It is only insufficiently understood to date, how ligand activity and selectivity towards one of the three transporters is achieved. Such knowledge could not only pave the way for functional studies on these transporters (by the use of truly selective tool compounds), but would also increase our knowledge on critical compound/drug properties associated with the onset of clinically relevant drug-drug interactions.

On principle, studies on determining factors for selectivity can include knowledge from the ligand side (QSAR/classification modeling, pharmacophore modeling), the protein side (comparative modeling, molecular docking, virtual screening, MD simulations), or both (PCM modeling or combinations of the latter approaches).

In case of OATP1B1, OATP1B3, and OATP2B1, comparative QSAR modeling has been performed by Karlgren et al. already in 2012 [55] which identified important molecular features for general OATP inhibition (vs. non-inhibition): higher lipophilicity, molecular weight and polarity. Just recently our group identified additional features discriminating hepatic OATP inhibitors from non-inhibitors: higher polarizability, molecular refractivity (corresponding to the distribution of charge over a molecule's surface), and flexibility (expressed as a higher number of rotatable bonds) [146]. Development of in silico models for individual OATP transporters by Karlgren et al. [55] has identified certain differences between the OATP1B and OATP2B subfamily. In contrast to OATP1B1 and OATP1B3, OATP2B1 inhibitory activity has been negatively correlated with nonpolar- and total- surface area, proposing that OATP2B1 inhibitors might be less dependent on polarity than OATP1B1 and OATP1B3 inhibitors. In addition, in our current study we could highlight additional properties to be responsible for OATP1B1 and OATP1B3 versus OATP2B1 inhibition (the latter seem to be more planar, whereas OATP1B members tend to possess a large number of amide bonds) [113].

Another way to explore ligand (and potentially selectivity) profiles is an enrichment analysis in substructures among actives of one target of interest vs the other(s). Again, for hepatic OATPs, this methodology led to a list of enriched scaffolds possessing a certain activity profile (i.e., OATP1B1 selective inhibition, OATP1B1/OATP1B3 dual inhibition, OATP1B1/OATP1B3/OATP2B1 pan-inhibition). As an outcome, e.g. the pravastatin-like scaffold showed a preferential inhibitory activity for OATP1B1 (over OATP1B3 and OATP2B1) and the cyclosporine-like scaffold accounted for OATP1B1/OATP1B3 dual inhibition. Interestingly, the steroidal scaffold has been found to be enriched in the actives of all three hepatic OATPs. Here depending on the side-chain variations, preferred activity towards one of the targets might be achieved [113].

Pharmacophore modeling is also interesting for studying ligand selectivity across different species. To give an example, human- and rabbit-OCT2 pharmacophore models indicate that despite the similarity

of most of the pharmacophoric features (reflected by 83% sequence identity of these two OCT2 variants), there is a difference in the spatial arrangement of hydrogen bonding features [109].

Finally, even ligand profiles and selectivity among uptake transporters of different families might be of interest, in particular if they are commonly expressed at the same pharmacological barrier. A way to tackle this is comparing pharmacophore hypotheses generated for the two targets of interest, like in the case of OCT2 and MATE1, which both are playing a significant role in renal disposition and toxicity (König et al. [141]). It has been shown that charge distribution was one of the important factors, favoring the inhibitory activity of one transporter with respect to the other. Specifically, OCT2 inhibitors comprise both zwitterionic and basic functional groups, whereas MATE1 inhibitors are less enriched with basic groups and do not necessarily contain zwitterionic groups [122].

An even more comprehensive understanding of OCT2-MATE1 selectivity profiling was delivered by a combinatorial pharmacophore-based approach [125,126], as introduced in chapter 4. Since MATE1 and OCT2 are commonly expressed in the kidney, it is interesting to learn about their interplay and selectivity switches to better understand transporter-mediated drug distribution and drug elimination processes (König et al., 2011). A combinatorial pharmacophore model approach developed for both OCT2 [125] and MATE1 [126] can therefore reveal which features are shared and which ones are unique for just one of these two transporters. The latter can give hints for selectivity switches at the ligand level. For OCT2, combinatorial pharmacophore modeling has revealed four distinct pharmacophoric hypotheses [125]. An aromatic feature was included in all four hypotheses, suggesting the essential role of pi-pi interactions in the OCT2-ligand recognition. In addition, a cationic charge has appeared in three out of four pharmacophoric hypotheses, which corresponds to previous findings [122]. Xu et al. also compared the molecular weights for the inhibitors matching different pharmacophoric features which provided additional insights into the constitution and/or size of distinctive binding site(s) within the transporter. Following the same strategy as for OCT2, the authors studied multiple inhibitory mechanisms of MATE1 ligands in a follow-up paper [126]. The model reveals significant importance of aromatic rings, as well as hydrophobicity to induce MATE1 inhibition. When compared to the combinatorial model for OCT2 inhibition, it becomes obvious that one of the hypotheses was the same in both transporters, thus proposing one common binding mode hypothesis which can accommodate a substantial number of dual MATE1 and OCT2 inhibitors.

In the future more sophisticated methods, such as multi-label classification might come into play, depending on the availability of compound data with consistent bioactivity measurements for targets under study. Such methods were recently used for studying selectivity profiles of ABC transporters [81].

In addition to the above discussed LB approaches for studying selectivity, the molecular basis for selectivity is delivered by the protein structure, and more specifically by subtle differences in residues interacting with the ligand during binding and transport. To give an illustrative example, attempts to understand selectivity among hepatic OATPs at a molecular level are discussed. Transmembrane regions for both OATP1B3 and OATP2B1 were built on basis of templates from the MFS family, namely glycerol-3-phosphate (pdb id: 1pw4) and lactose permease (pdb id: 1pv6) from *Escherichia coli* [77]. The model quality has subsequently been validated via docking of the cardiac glycoside digoxin into the putative translocation channel. Based on the 3D models and multiple sequence alignment across OATP family members, the analyses suggest that the pore-facing residue R181 might contribute to the substrate specificity of OATP1B transporters, as this residue is fully conserved across the OATP1B family. In analogy, H579 is hypothesized to be crucial for binding of ligands to members of the OATP2B family and it is found at a spatially adjacent position to R181 [77]. Another SB modeling effort for understanding commonalities and differences of the more closely related hepatic transporters OATP1B1 and

OATP1B3 (~80% sequence identity) led to the construction of a series of chimeric proteins between OATP1B3 and 1B1 [38]. The aim here was the determination of structural domains and/or residues responsible for substrate selectivity of OATP1B3, specifically for CCK-8. Homology modeling and molecular docking led to binding mode hypotheses which were further validated experimentally. When replacing TM10 in OATP1B3 with TM10 of OATP1B1 a dramatically reduced degree of CCK-8 transport was observed, indicating that TM10 is indeed crucial for recognition and/or translocation of CCK-8. Using site-directed mutagenesis, key residues for substrate binding namely, Y537, S545, and T550 in TM10 were identified [38].

Using ligand and protein information in conjunction for selectivity profiling can be conducted by using PCM modeling. This technique outperforms conventional QSAR models and can be used to virtually screen for selective compounds that are solely active on a single member of a subfamily of targets [115]. In the field of clinically relevant SLC transporters, PCM modeling was first undertaken for investigating chemical features favoring OATP1B1 and OATP1B3 inhibition [14]. Performing multiple sequence alignment for OATP1B1, OATP1B3, OATP2B1, and OATP1A2 aided in identifying most conserved regions. Further, critical protein residues were prioritized on basis of SB modeling studies previously done for hepatic OATPs [73,77]. This step again demonstrates the usefulness of combining different computational approaches. PCM models were developed for OATP1B inhibition (2-class classification model, i.e., 'OATP1B inhibitor' or 'OATP1B non-inhibitor'), and individual OATP1B1/1B3 inhibition (4-class classification model, i.e., 'OATP1B dual inhibitor', 'OATP1B dual non-inhibitor', 'OATP1B1 selective inhibitor', or 'OATP1B3 selective inhibitor'). When looking at protein properties, only limited conclusions could be deduced from this study. The limited interpretability of target information is caused by the fact that only two proteins were included into the PCM modeling (van Westen et al., 2012). As a future perspective, the authors suggest to apply the PCM procedure to more OATP proteins, as there are bioactivity measurements for 22 OATP isoforms available in ChEMBL. In conclusion, the developed 2- and 4-class classification models united the important molecular features reported from previous computational studies [6,17,55,105], but also provided new information about OATP1B1/1B3 inhibition (e.g. high number of ring bonds).

Pharmacological profiles can also be investigated for a whole group of simultaneously expressed transporters located at the same pharmacological barrier. For example, Shaikh et al. combined QSAR and PCM modeling to perform an extensive exploration of substrate interactions for 13 clinically relevant efflux and uptake transporters, including OCT1, OATP1B1, OATP1B3, OATP2B1, and OATP1A2 [104]. The motivation behind such an extensive study is to predict transport across major pharmacological barriers as a whole by a sequence of computational models. Thus, consensus models by applying various machine learning techniques were finally constructed and the developed models were used to predict the substrate propensity of compounds for blood-brain barrier (BBB) transport.

6. Accounting for Transporter Flexibility

Since substrate translocation via transporters requires the protein to adapt different conformational states, protein flexibility has to be taken into account also when performing SB modeling [30]. So far, most of the docking studies on uptake transporters were performed by considering only a single transporter conformation (e.g., OAT1 [93], OCT1 [13], OATP1B1 [69], OATP1B3 [38]). However, building a structural model on basis of a single conformation can inherently bias subsequent docking screens since e.g. some ligands might not fit into a narrow binding pocket of a particular conformation. Khuri et al. attempted to capture OATP2B1 transporter dynamics by building multiple OATP2B1 comparative models based on seven templates considering different conformational states: inward-open, outward-open, and occluded [57]. In general, multiple comparative models can together provide a

more comprehensive representation of a ligand binding event [100]. However, combining drastically different conformational states (such as outward-open with inward) is risky, since allosteric modulations rather happen upon subtle changes in protein conformation and slight rotameric variations of side chains [96].

An orthogonal approach to the selection of multiple template structures for sampling protein conformations is the generation of template protein ensembles by using MD simulations. For lactose permease (pdb code: 1pv6), a representative of the Major Facilitator Superfamily which to a high degree relates to uptake transporters by secondary

structure, both all-atom [42,91] and coarse-grained [50] simulations have been performed for elucidating ligand binding and even transport events. However, direct usage of MD simulations to mimic the translocation process of template-based comparative models is only rarely used since the majority of available template structures cover only transmembrane regions. For example, available templates for OATP1B1 cover <400 amino acids out of 691 in total. Running MD simulations with an incomplete target structure could lead to artifacts. Furthermore, the limited time scale for all-atom simulations (hundreds of nanoseconds to a few microseconds), as well as the choice of including or omitting the lipid

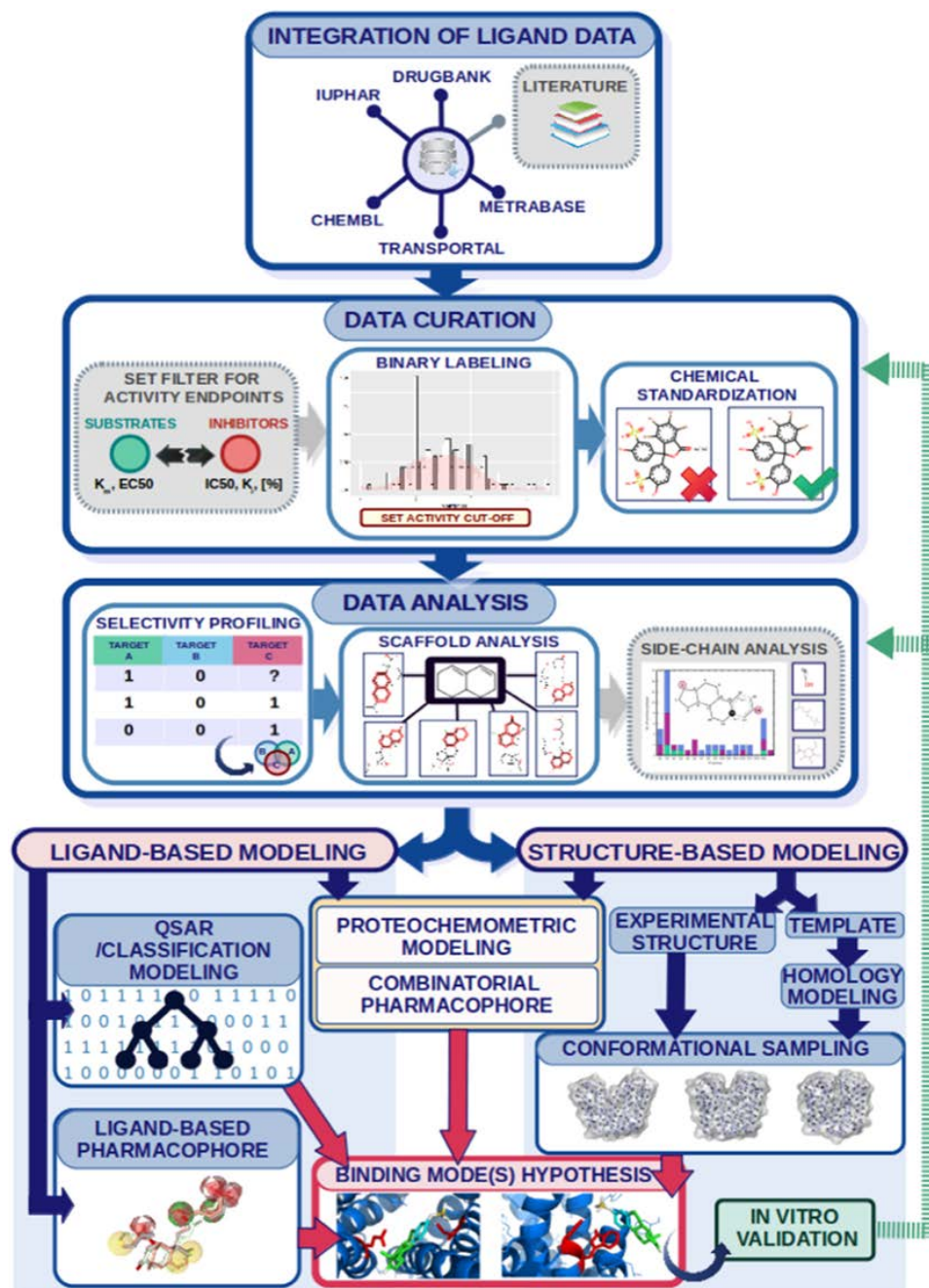


Fig. 3. Proposed computational workflow for studying ligand interactions with ADMET-relevant SLC transporters. Results from in vitro validation can be inputted to the stages of data curation and data analysis and subsequently be used for a new iteration of modeling.

bilayer into the simulated system, can heavily affect the correctness and interpretability of the MD simulations. Contemporary simulation techniques, such as stochastic Monte Carlo, or MD simulations with enhanced sampling (e.g. by applying replica-exchange methods) can be used for structural refinement of extra- and intracellular domains.

To date, enhanced sampling techniques have been used only once for building a complete structural model for any of the discussed uptake transporters. As shown in a MD simulation study on OAT1 [112], modeling of the complete OAT1 transporter structure was divided into two stages: First, only the transmembrane region was modeled based on a template with high secondary structure similarity (glycerol-3-phosphate transporter; pdb code: 1pw4). At the second step, the extracellular domain was iteratively sampled by simulated annealing, while the transmembrane region was kept restrained. Results of this study indicate that the structural refinement by applying enhanced sampling methods could significantly improve existing structural models for uptake transporters which in turn would enhance the understanding of functional aspects of the transport mechanisms.

Only recently, reduced representation methods, such as normal mode analysis by applying elastic network models, can be used to overcome shortcomings arising from the high computational demands of conventional MD simulations ([142,143]). For example, elastic network models have been used for the structurally-related fucose transporter (pdb code: 3o7q) to study its molecular basis for allosteric modulations [18]. Furthermore, by building comparative models in different functional states (inward- and outward-open conformation), elastic network models were capable to reproduce the whole translocation pathway of this transporter.

For multiple ensemble docking, normal mode simulations have been shown to be particularly useful to e.g. detect a biologically relevant conformation of dopamine D3 receptor, which has subsequently been prospectively validated by the existing dopamine D3 crystal structure [15]. One might argue that the generation of multiple conformations for a template structure might inherently bias the construction of structural models, since they can possess its internal dynamics which probably cannot be completely captured by the template. However, as shown and discussed in case of the human MATE1 homology model [130], the overall conformational stability between the template and the derived MATE1 homology models remained unchanged.

As a conclusion, the use of normal mode simulations in structure-based modeling studies can potentially improve conformational sampling when modeling uptake transporters. This in turn can lead to more accurate docking poses with the aim to better understand ligand-protein binding events and potentially selectivity switches.

7. Summary, Conclusions & Future Perspectives

ADMET-related SLC transporters are proteins of emerging interest in the framework of preclinical drug design. As demonstrated herein - by collecting available ligand and protein information from the open domain - data sparseness resulted in quite limited understanding of these transporter to date. Other factors complicating effective exploration of this class of proteins is their promiscuous nature, with potentially multiple binding sites, as well as overlapping substrate- and inhibitor profiles.

As demonstrated by discussed examples of molecular modeling and data analysis herein, new emerging technologies are on the rise also for these targets being particularly hard to unlock. Especially, data integration techniques and data analysis can lead to useful hypothesis about interesting SAR series at the beginning of an in silico study. Further, combining LB and SB methods seems to be an effective strategy, especially when it comes to selectivity profiling (like in the case of PCM modeling), or the exploration of knowledge about multiple binding sites (like in the case of combinatorial pharmacophores). In general, inclusion of in vitro experiments is a must especially for SB methods, e.g. to test the established binding mode hypotheses. In return, those

in vitro measurements will lead to an increase in data points for a particular target, which can further be explored by statistical methods (such as machine learning approaches). For SB approaches, it would be interesting to include more systematically conformational sampling of protein conformations and multiple template structures into the comparative modeling step. It will be interesting to then compare results to those from docking into a single static template.

We are proposing a general workflow for in silico modeling of clinically relevant SLC transporters (see Fig. 3) which makes use of all available molecular modeling approaches and combines them with timely data science approaches (as far as the available data allows the different methods).

With the aim to develop safer medicine, it is of extraordinary importance to increase our understanding of the molecular basis of respective transporter-ligand interactions on an individual level (transporter by transporter), but also from a global point of view (how they act in concert). During evolution, transporters were optimized to help metabolizing chemical matter that we were exposed to. Obviously, they found a way to transport a wide variety of chemically distinct compounds. Understanding the interplay of clinically relevant transporters in transporting chemical matter, as well as the mechanisms underlying transporter selectivity, can help to unravel potential drug-drug or drug-food interactions which will finally lead to safer drugs in the future.

Acknowledgements

We acknowledge financial support provided from the Austrian Science Fund (FWF), grant no. P 29712.

References

- Ahlin G, Karlsson J, Pedersen JM, Gustavsson L, Larsson R, Matsson P, et al. Structural requirements for drug inhibition of the liver specific human organic cation transport protein 1. *J Med Chem* 2008;51:5932–42. <https://doi.org/10.1021/jm8003152>.
- Ain QU, Aleksandrova A, Roessler FD, Ballester PJ. Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. *Wiley Interdisc Rev* 2015;5:405–24. <https://doi.org/10.1002/wcms.1225>.
- Aldeghi M, Heifetz A, Bodkin MJ, Knapp S, Biggin PC. Predictions of ligand selectivity from absolute binding free energy calculations. *J Am Chem Soc* 2017;139:946–57. <https://doi.org/10.1021/jacs.6b11467>.
- Angelini S, Pantaleo MA, Ravegnini G, Zenesini C, Cavrini G, Nannini M, et al. Polymorphisms in OCTN1 and OCTN2 transporter genes are associated with prolonged time to progression in unresectable gastrointestinal stromal tumours treated with imatinib therapy. *Pharmacol Res* 2013;68(1):1–6. <https://doi.org/10.1016/j.phrs.2012.10.015>.
- Astorga B, Ekins S, Morales M, Wright SH. Molecular determinants of ligand selectivity for the human multidrug and toxin extruder proteins MATE1 and MATE2-K. *J Pharmacol Exp Ther* 2012;341:743–55. <https://doi.org/10.1124/jpet.112.191577>.
- Badolo L, Rasmussen LM, Hansen HR, Sveigaard C. Screening of OATP1B1/3 and OCT1 inhibitors in cryopreserved hepatocytes in suspension. *Eur J Pharm Sci* 2010;40:282–8. <https://doi.org/10.1016/j.ejps.2010.03.023>.
- Bailey DG, Dresser GK, Leake BF, Kim RB. Naringin is a major and selective clinical inhibitor of organic anion-transporting polypeptide 1A2 (OATP1A2) in grapefruit juice. *Clin Pharmacol Ther* 2007;81:495–502. <https://doi.org/10.1038/sj.cpt.6100104>.
- Ballante F. Protein-ligand docking in drug design: Performance assessment and binding-pose selection. In: Mavroumoustakos T, Kellici TF, editors. *Rational drug design: methods and protocols*. New York, NY: Methods in Molecular Biology; 2018. p. 67–88. https://doi.org/10.1007/978-1-4939-8630-9_5.
- Baumgartner MP, Camacho CJ. Choosing the optimal rigid receptor for docking and scoring in the CSAR 2013/2014 experiment. *J Chem Inf Model* 2016;56:1004–12. <https://doi.org/10.1021/acs.jcim.5b00338>.
- Bento AP, Gaulton A, Hersey A, Bellis LJ, Chambers J, Davies M, et al. The ChEMBL bioactivity database: an update. *Nucleic Acids Res* 2014;42:D1083–90. <https://doi.org/10.1093/nar/gkt1031>.
- Berthold MR, Cebron N, Dill F, Gabriel TR, Köster T, Meini T, et al. KNIME - the Konstanz information miner: version 2.0 and beyond. *SIGKDD Explor News* 2009;11:26–31. <https://doi.org/10.1145/1656274.1656280>.
- Böhm H-J. Prediction of binding constants of protein ligands: a fast method for the prioritization of hits obtained from de novo design or 3D database search programs. *J Comput Aided Mol Des* 1998;12:309–23. <https://doi.org/10.1023/A:1007999920146>.

- [13] Boxberger KH, Hagenbuch B, Lampe JN. Ligand-dependent modulation of hOCT1 transport reveals discrete ligand binding sites within the substrate translocation channel. *Biochem Pharmacol* 2018;156:371–84. <https://doi.org/10.1016/j.bcp.2018.08.028>.
- [14] Bruyn TD, Westen GJPV, IJzerman AP, Stieger B, Witte de P, Augustijns PF, et al. Structure-based identification of OATP1B1/3 inhibitors. *Mol Pharmacol Mol* 2013. <https://doi.org/10.1124/mol.112.084152> 112.084152.
- [15] Carlsson J, Coleman RG, Setola V, Irwin JJ, Fan H, Schlessinger A, et al. Ligand discovery from a dopamine D₃ receptor homology model and crystal structure. *Nat Chem Biol* 2011;7:769–78. <https://doi.org/10.1038/nchembio.662>.
- [16] Cha SH, Sekine T, Fukushima JI, Kanai Y, Kobayashi Y, Goya T, et al. Identification and characterization of human organic anion transporter 3 expressing predominantly in the kidney. *Mol Pharmacol* 2001;59:1277–86.
- [17] Chang C. Comparative pharmacophore modeling of organic anion transporting polypeptides: a meta-analysis of rat Oatp1a1 and human OATP1B1. *J Pharmacol Exp Ther* 2005;314:533–41. <https://doi.org/10.1124/jpet.104.082370>.
- [18] Chang S, Li K, Hu J, Jiao X, Tian X. Allosteric and transport behavior analyses of a fucose transporter with network models. *Soft Matter* 2011;7:4661–71. <https://doi.org/10.1039/C0SM01543A>.
- [19] Chen EC, Khuri N, Liang X, Stecula A, Chien H-C, Yee SW, et al. Discovery of competitive and noncompetitive ligands of the organic cation transporter 1 (OCT1; SLC22A1). *J Med Chem* 2017;60:2685–96. <https://doi.org/10.1021/acs.jmedchem.6b01317>.
- [20] Colas C, Ung PM-U, Schlessinger A. SLC transporters: structure, function, and drug discovery. *Medchemcomm* 2016;7:1069–81. <https://doi.org/10.1039/C6MD00005C>.
- [21] Dakal TC, Kumar R, Ramotar D. Structural modeling of human organic cation transporters. *Comput Biol Chem* 2017;68:153–63. <https://doi.org/10.1016/j.cmbiolchem.2017.03.007>.
- [22] Dakal TC, Kala D, Dhiman G, Yadav V, Krokhotin A, Dokholyan NV. Predicting the functional consequences of non-synonymous single nucleotide polymorphisms in IL8 gene. *Sci Rep* 2017;7(1). <https://doi.org/10.1038/s41598-017-06575-4> 6525.
- [23] Dastmalchi S, Hamzeh-Mivehroud M, Sokouti B, Hamzeh-Mivehroud M, Sokouti B. Quantitative structure – activity relationship : a practical approach. CRC Press; 2018. <https://doi.org/10.1201/9781351113076>.
- [24] Deng N, Flynn WF, Xia J, Vijayan RSK, Zhang B, He P, et al. Large scale free energy calculations for blind predictions of protein–ligand binding: the D3R grand challenge 2015. *J Comput Aided Mol Des* 2016;30:743–51. <https://doi.org/10.1007/s10822-016-9952-x>.
- [25] Diao L, Ekins S, Polli JE. Quantitative structure activity relationship for inhibition of human organic cation/carnitine transporter. *Mol Pharm* 2010;7:2120–31. <https://doi.org/10.1021/mp100226q>.
- [26] Diao L, Ekins S, Polli JE. Novel inhibitors of human organic cation/carnitine transporter (hOCTN2) via computational modeling and in vitro testing. *Pharm Res* 2009;26:1890–900. <https://doi.org/10.1201/s11095-009-9905-3>.
- [27] Drenberg CD, Gibson AA, Pounds SB, Shi L, Rhinehart DP, Li L, et al. OCTN1 is a high-affinity carrier of nucleoside analogs. *Cancer Res* 2017;77(8):2102–11. <https://doi.org/10.1158/0008-5472.CAN-16-2548>.
- [28] Dresser GK, Bailey DG, Leake BF, Schwarz UI, Dawson PA, Freeman DJ, et al. Fruit juices inhibit organic anion transporting polypeptide-mediated drug uptake to decrease the oral availability of fexofenadine. *Clin Pharmacol Ther* 2002;71:11–20. <https://doi.org/10.1067/mcp.2002.121152>.
- [29] Enomoto A, Takeda M, Shimoda M, Narikawa S, Kobayashi Yukari, Kobayashi Yasuna, et al. Interaction of human organic anion transporters 2 and 4 with organic anion transport inhibitors. *J Pharmacol Exp Ther* 2002;301:797–802.
- [30] Erickson JA, Jalaie M, Robertson DH, Lewis RA, Vieth M. Lessons in molecular recognition: the effects of ligand and protein flexibility on molecular docking accuracy. *J Med Chem* 2004;47:45–55. <https://doi.org/10.1021/jm030209y>.
- [31] Feig M. Implicit membrane models for membrane protein simulation. *Methods Mol Biol* 2008;443:181–96. https://doi.org/10.1007/978-1-59745-177-2_10.
- [32] Fiser A, Do RK, Sali A. Modeling of loops in protein structures. *Protein Sci* 2000;9:1753–73.
- [33] Fiser A, Sali A. ModLoop: automated modeling of loops in protein structures. *Bioinformatics* 2003;19:2500–1. <https://doi.org/10.1093/bioinformatics/btg362>.
- [34] Flohil JA, Vriend G, Berendsen HJC. Completion and refinement of 3-D homology models with restricted molecular dynamics: application to targets 47, 58, and 111 in the CASP modeling competition and posterior analysis. *Proteins* 2002;48:593–604. <https://doi.org/10.1002/prot.10105>.
- [35] Genheden S, Ryde U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opin Drug Discovery* 2015;10:449–61. <https://doi.org/10.1517/17460441.2015.1032936>.
- [36] Glaeser H, Mandery K, Sticht H, Fromm MF, König J. Relevance of conserved lysine and arginine residues in transmembrane helices for the transport activity of organic anion transporting polypeptide 1B3. *Br J Pharmacol* 2010;159:698–708.
- [37] Gorboulev V, Ulzheimer JC, Akhoundova A, Ulzheimer-Teuber I, Karbach U, Quester S, et al. Cloning and characterization of two human polyspecific organic cation transporters. *DNA Cell Biol* 1997;16:871–81. <https://doi.org/10.1089/dna.1997.16.871>.
- [38] Gui C, Hagenbuch B. Amino acid residues in transmembrane domain 10 of organic anion transporting polypeptide 1B3 are critical for cholecystokinin Octapeptide transport[†]. *Biochemistry* 2008;47:9090–7. <https://doi.org/10.1021/bi8008455>.
- [39] Hardin C, Pogorelov TV, Luthy-Schulten Z. Ab initio protein structure prediction. *Curr Opin Struct Biol* 2002;12:176–81.
- [40] Heim AJ, Li Z. Developing a high-quality scoring function for membrane protein structures based on specific inter-residue interactions. *J Comput Aided Mol Des* 2012;26:301–9. <https://doi.org/10.1007/s10822-012-9556-z>.
- [41] Hirano M, Maeda K, Shitara Y, Sugiyama Y. Contribution of OATP2 (OATP1B1) and OATP8 (OATP1B3) to the hepatic uptake of pitavastatin in humans. *J Pharmacol Exp Ther* 2004;311:139–46. <https://doi.org/10.1124/jpet.104.068056>.
- [42] Holyoake J, Sansom MSP. Conformational change in an MFS protein: MD simulations of LacY. *Structure* 2007;15:873–84. <https://doi.org/10.1016/j.str.2007.06.004>.
- [43] Hoshino Y, Fujita D, Nakanishi T, Tamai I. Molecular localization and characterization of multiple binding sites of organic anion transporting polypeptide 2B1 (OATP2B1) as the mechanism for substrate and modulator dependent drug–drug interaction. *Med Chem Commun* 2016;7:1775–82. <https://doi.org/10.1039/C6MD00235H>.
- [44] Hosoyamada M, Sekine T, Kanai Y, Endou H. Molecular cloning and functional expression of a multispecific organic anion transporter from human kidney. *Am J Physiol* 1999;276:F122–8.
- [45] Hu Y, Stumpfe D, Bajorath J. Recent advances in scaffold hopping: miniperspective. *J Med Chem* 2017;60:1238–46. <https://doi.org/10.1021/acs.jmedchem.6b01437>.
- [46] Huang N, Shoichet BK, Irwin JJ. Benchmarking sets for molecular docking. *J Med Chem* 2006;49:6789–801. <https://doi.org/10.1021/jm0608356>.
- [47] Irwin JJ. Community benchmarks for virtual screening. *J Comput Aided Mol Des* 2008;22:193–9. <https://doi.org/10.1007/s10822-008-9189-4>.
- [48] Ismail MG, Stieger B, Cattori V, Hagenbuch B, Fried M, Meier PJ, et al. Hepatic uptake of cholecystokinin octapeptide by organic anion-transporting polypeptides OATP4 and OATP8 of rat and human liver. *Gastroenterology* 2001;121:1185–90.
- [49] Jamroz M, Kolinski A. Modeling of loops in proteins: a multi-method approach. *BMC Struct Biol* 2010;10:5. <https://doi.org/10.1186/1472-6807-10-5>.
- [50] Jewel Y, Dutta P, Liu J. Exploration of conformational changes in lactose permease upon sugar binding and proton transfer through coarse-grained simulations. *Proteins* 2017;85:1856–65. <https://doi.org/10.1002/prot.25340>.
- [51] Jothi A. Principles, challenges and advances in ab initio protein structure prediction. *Protein Pept Lett* 2012;19:1194–204.
- [52] Kaczanowski S, Zielenkiewicz P. Why similar protein sequences encode similar three-dimensional structures? *Theor Chem Acc* 2010;125:643–50. <https://doi.org/10.1007/s00214-009-0656-3>.
- [53] Kallioikoski T, Kramer C, Vulpatti A, Gedeck P. Comparability of mixed IC50 data – a statistical analysis. *PLoS One* 2013;8. <https://doi.org/10.1371/journal.pone.0061007>.
- [54] Karlgren M, Ahlin G, Bergström CAS, Svensson R, Palm J, Artursson P. *in vitro* and *in silico* strategies to identify OATP1B1 inhibitors and predict clinical drug–drug interactions. *Pharm Res* 2012;29:411–26. <https://doi.org/10.1007/s11095-011-0564-9>.
- [55] Karlgren M, Vildhede A, Norinder U, Wisniewski JR, Kimoto E, Lai Y, et al. Classification of inhibitors of hepatic organic anion transporting polypeptides (OATPs): influence of protein expression on drug–drug interactions. *J Med Chem* 2012;55:4740–63. <https://doi.org/10.1021/jm300212s>.
- [56] Kell DB, Dobson PD, Bilsland E, Oliver SG. The promiscuous binding of pharmaceutical drugs and their transporter-mediated uptake into cells: what we (need to) know and how we can do so. *Drug Discov Today* 2013;18:218–39. <https://doi.org/10.1016/j.drudis.2012.11.008>.
- [57] Khuri N, Zur AA, Wittwer MB, Lin L, Yee SW, Sali A, et al. Computational discovery and experimental validation of inhibitors of the human intestinal transporter OATP2B1. *J Chem Inf Model* 2017;57:1402–13. <https://doi.org/10.1021/acs.jcim.6b00720>.
- [58] Kido Y, Matsson P, Giacomini KM. Profiling of a prescription drug library for potential renal drug–drug interactions mediated by the organic cation transporter 2. *J Med Chem* 2011;54:4548–58. <https://doi.org/10.1021/jm2001629>.
- [59] Kollman PA, Massova I, Reyes C, Kuhn B, Huo S, Chong L, et al. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc Chem Res* 2000;33:889–97. <https://doi.org/10.1021/ar000033j>.
- [60] Kotsampasakou E, Brenner S, Jäger W, Ecker GF. Identification of novel inhibitors of organic anion transporting polypeptides 1B1 and 1B3 (OATP1B1 and OATP1B3) using a consensus vote of six classification models. *Mol Pharm* 2015;12:4395–404. <https://doi.org/10.1021/acs.molpharmaceut.5b00583>.
- [61] Kramer C, Kallioikoski T, Gedeck P, Vulpatti A. The experimental uncertainty of heterogeneous public K(i) data. *J Med Chem* 2012;55:5165–73. <https://doi.org/10.1021/jm300131x>.
- [62] Krieger E, Joo K, Lee Jinwoo, Lee Jooyoung, Raman S, Thompson J, et al. Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: four approaches that performed well in CASP8. *Proteins* 2009;77(Suppl. 9):114–22. <https://doi.org/10.1002/prot.22570>.
- [63] Ladizhansky V. Applications of solid-state NMR to membrane proteins. *Biochimica et Biophysica Acta (BBA) - proteins and proteomics*. Biophys Canada 2017;1865:1577–86. <https://doi.org/10.1016/j.bbapap.2017.07.004>.
- [64] Lapinsh M, Prusis P, Lundstedt T, Wikberg JES. Proteochemometrics modeling of the interaction of amine G-protein coupled receptors with a diverse set of ligands. *Mol Pharmacol* 2002;61:1465–75.
- [65] Leach A. *Molecular modelling: principles and applications*. 2nd. ed. Harlow, England ; New York: Pearson; 2001.
- [66] Lechner C, Ishiguro N, Fukuhara A, Shimizu H, Ohtsu N, Takatani M, et al. Impact of experimental conditions on the evaluation of interactions between multidrug and toxin extrusion proteins and candidate drugs. *Drug Metab Dispos* 2016;44:1381–9. <https://doi.org/10.1124/dmd.115.068163>.
- [67] Lee W, Glaeser H, Smith LH, Roberts RL, Moekel GW, Gervasini G, et al. Polymorphisms in human organic anion-transporting polypeptide 1A2

- (OATP1A2): implications for altered drug disposition and central nervous system drug entry. *J Biol Chem* 2005;280:9610–7. <https://doi.org/10.1074/jbc.M411092200>.
- [68] Lengauer T, Rarey M. Computational methods for biomolecular docking. *Curr Opin Struct Biol* 1996;6:402–6.
- [69] Li N, Hong W, Huang H, Lu H, Lin G, Hong M. Identification of amino acids essential for Estrone-3-sulfate transport within transmembrane domain 2 of organic anion transporting polypeptide 1B1. *PLoS One* 2012;7:e36647. <https://doi.org/10.1371/journal.pone.0036647>.
- [70] Liu X, Huang J, Sun Y, Zhan K, Zhang Z, Hong M. Identification of multiple binding sites for substrate transport in bovine organic anion transporting polypeptide 1a2. *Drug Metab Dispos* 2013;41:602–7. <https://doi.org/10.1124/dmd.112.047910>.
- [71] Lobley A, Sadowski MI, Jones DT. pGentHREADER and pDomHREADER: new methods for improved protein fold recognition and superfamily discrimination. *Bioinformatics* 2009;25:1761–7. <https://doi.org/10.1093/bioinformatics/btp302>.
- [72] Mak L, Marcus D, Howlett A, Yarova G, Duchateau G, Klaffke W, et al. Metrabase: a cheminformatics and bioinformatics database for small molecule transporter data analysis and (Q)SAR modelling. *J Chem* 2015;7:31. <https://doi.org/10.1186/s13321-015-0083-5>.
- [73] Mandery K, Sticht H, Bujok K, Schmidt I, Fahrmayr C, Balk B, et al. Functional and structural relevance of conserved positively charged lysine residues in organic anion transporting polypeptide 1B3. *Mol Pharmacol* 2011;80:400–6. <https://doi.org/10.1124/mol.111.071282>.
- [74] Martínez-Guerrero LJ, Wright SH. Substrate-dependent inhibition of human MATE1 by cationic ionic liquids. *J Pharmacol Exp Ther* 2013;346:495–503. <https://doi.org/10.1124/jpet.113.204206>.
- [75] Martí-Renom MA, Stuart AC, Fiser A, Sánchez R, Melo F, Šali A. Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct* 2000;29:291–325. <https://doi.org/10.1146/annurev.biophys.29.1.291>.
- [76] Masuda S, Terada T, Yonezawa A, Tanihara Y, Kishimoto K, Katsura T, et al. Identification and functional characterization of a new human kidney-specific H⁺/organic cation antiporter, kidney-specific multidrug and toxin extrusion 2. *J Am Soc Nephrol* 2006;17:2127–35. <https://doi.org/10.1681/ASN.2006030205>.
- [77] Meier-Abt F, Mokrab Y, Mizuguchi K. Organic anion transporting polypeptides of the OATP/SLCO superfamily: identification of new members in nonmammalian species, comparative modeling and a potential transport mode. *J Membr Biol* 2006;208:213–27. <https://doi.org/10.1007/s00232-005-7004-x>.
- [78] Meyer EF, Swanson SM, Williams JA. Molecular modelling and drug design. *Pharmacol Ther* 2000;85:113–21.
- [79] Moaddel R, Ravichandran S, Bigli F, Yamaguchi R, Wainer IW. Pharmacophore modelling of stereoselective binding to the human organic cation transporter (hOCT1). *Br J Pharmacol* 2007;151:1305–14. <https://doi.org/10.1038/sj.bjp.0707341>.
- [80] Montanari F, Ecker GF. BCRP inhibition: from data collection to ligand-based modeling. *Mol Inf* 2014;33:322–31. <https://doi.org/10.1002/minf.201400012>.
- [81] Montanari F, Zdrážil B, Digles D, Ecker GF. Selectivity profiling of BCRP versus P-gp inhibition: from automated collection of polypharmacology data to multi-label learning. *J Chem* 2016;8. <https://doi.org/10.1186/s13321-016-0121-y>.
- [82] Mori T, Miyashita N, Im W, Feig M, Sugita Y. Molecular dynamics simulations of biological membranes and membrane proteins using enhanced conformational sampling algorithms. *Biochim Biophys Acta (BBA) - Biomembr* 2016;1858:1635–51. <https://doi.org/10.1016/j.bbamem.2015.12.032>.
- [83] Morrissey KM, Wen CC, Johns SJ, Zhang L, Huang S-M, Giacomini KM. The UCSF-FDA TransPortal: a public drug transporter database. *Clin Pharmacol Ther* 2012;92:545–6. <https://doi.org/10.1038/clpt.2012.44>.
- [84] Motohashi H, Inui K. Organic cation transporter OCTs (SLC22) and MATEs (SLC47) in the human kidney. *AAPS J* 2013;15:581–8. <https://doi.org/10.1208/s12248-013-9465-7>.
- [85] Muegge I. PMF scoring revisited. *J Med Chem* 2006;49:5895–902. <https://doi.org/10.1021/jm050038s>.
- [86] Neuvonen PJ, Niemi M, Backman JT. Drug interactions with lipid-lowering drugs: mechanisms and clinical relevance. *Clin Pharmacol Ther* 2006;80:565–81. <https://doi.org/10.1016/j.clpt.2006.09.003>.
- [87] Newstead S, Ferrandon S, Iwata S. Rationalizing alpha-helical membrane protein crystallization. *Protein Sci* 2008;17:466–72. <https://doi.org/10.1110/ps.073263108>.
- [88] Nozaki Y, Kusuhashi H, Kondo T, Iwaki M, Shiroyanagi Y, Nakayama H, et al. Species difference in the inhibitory effect of nonsteroidal anti-inflammatory drugs on the uptake of methotrexate by human kidney slices. *J Pharmacol Exp Ther* 2007;322:1162–70. <https://doi.org/10.1124/jpet.107.121491>.
- [89] Parker JL, Newstead S. Membrane protein crystallisation: current trends and future perspectives. *Adv Exp Med Biol* 2016;922:61–72. https://doi.org/10.1007/978-3-319-35072-1_5.
- [90] Pawson AJ, Sharman JL, Benson HE, Facenda E, Alexander SPH, Buneman OP, et al. The IUPHAR/BPS guide to pharmacology: an expert-driven knowledgebase of drug targets and their ligands. *Nucleic Acids Res* 2014;42:D1098–106. <https://doi.org/10.1093/nar/gkt1143>.
- [91] Pendse PY, Brooks BR, Klauda JB. Probing the periplasmic-open state of lactose permease in response to sugar binding and proton translocation. *J Mol Biol* 2010;404:506–21. <https://doi.org/10.1016/j.jmb.2010.09.045>.
- [92] Perländ E, Fredriksson R. Classification Systems of Secondary Active Transporters. *Trends Pharmacol Sci* 2017;38:305–15. <https://doi.org/10.1016/j.tips.2016.11.008>.
- [93] Perry JL, Dembla-Rajpal N, Hall LA, Pritchard JB. A three-dimensional model of human organic anion transporter 1: aromatic amino acids required for substrate transport. *J Biol Chem* 2006;281:38071–9. <https://doi.org/10.1074/jbc.M608834200>.
- [94] Pochini L, Scalise M, Galluccio M, Indiveri C. OCTN cation transporters in health and disease: role as drug targets and assay development. *J Biomol Screen* 2013;18:851–67. <https://doi.org/10.1177/1087057113493006>.
- [95] Ray A, Lindahl E, Wallner B. Model quality assessment for membrane proteins. *Bioinformatics* 2010;26:3067–74. <https://doi.org/10.1093/bioinformatics/btq581>.
- [96] Rodgers TL, Townsend PD, Burnell D, Jones ML, Richards SA, McLeish TCB, et al. Modulation of global low-frequency motions underlies allosteric regulation: demonstration in CRP/FNR family transcription factors. *PLoS Biol* 2013;11:e1001651. <https://doi.org/10.1371/journal.pbio.1001651>.
- [97] Roth M, Obaidat A, Hagenbuch B. OATPs, OATs and OCTs: the organic anion and cation transporters of the SLCO and SLC22A gene superfamilies: OATPs, OATs and OCTs. *Br J Pharmacol* 2012;165:1260–87. <https://doi.org/10.1111/j.1476-5381.2011.01724.x>.
- [98] Rouck J, Krapf J, Roy J, Huff H, Das A. Recent advances in nanodisc technology for membrane proteins studies (2012–2017). *FEBS Lett* 2017;591:2057–88. <https://doi.org/10.1002/1873-3468.12706>.
- [99] Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* 1993;234:779–815. <https://doi.org/10.1006/jmbi.1993.1626>.
- [100] Schafferhans A, Klebe G. Docking ligands onto binding site representations derived from proteins built by homology modelling 11 Edited by J. Thornton. *J Mol Biol* 2001;307:407–27. <https://doi.org/10.1006/jmbi.2000.4453>.
- [101] Schlegel S, Klepsch M, Gialama D, Wickström D, Slotboom DJ, De Gier J. Revolutionizing membrane protein overexpression in bacteria. *J Microbiol Biotechnol* 2010;3:403–11. <https://doi.org/10.1111/j.1751-7915.2009.00148.x>.
- [102] Schlessinger A, Khuri N, Giacomini KM, Sali A. Molecular modeling and ligand docking for solute carrier (SLC) transporters. *Curr Top Med Chem* 2013;13:843–56.
- [103] Schlessinger A, Welch MA, van Vlijmen H, Korzekwa K, Swaan PW, Matsson P. Molecular modeling of drug-transporter interactions—an international transporter consortium perspective. *Clin Pharmacol Ther* 2018;104:818–35. <https://doi.org/10.1002/cpt.1174>.
- [104] Shaikh N, Sharma M, Garg P. Selective fusion of heterogeneous classifiers for predicting substrates of membrane transporters. *J Chem Inf Model* 2017;57:594–607. <https://doi.org/10.1021/acs.jcim.6b00508>.
- [105] Soars MG, Barton P, Ismail M, Jupp R, Riley RJ. The development, characterization, and application of an OATP1B1 inhibition assay in drug discovery. *Drug Metab Dispos* 2012;40:1641–8. <https://doi.org/10.1124/dmd.111.042382>.
- [106] Sonoda Y, Newstead S, Hu N-J, Alguet Y, Nji E, Beis K, et al. Benchmarking membrane protein detergent stability for improving throughput of high-resolution X-ray structures. *Structure* 2011;19:17–25. <https://doi.org/10.1016/j.str.2010.12.001>.
- [107] Steindl TM, Schuster D, Wolber G, Lagner C, Langer T. High-throughput structure-based pharmacophore modelling as a basis for successful parallel virtual screening. *J Comput Aided Mol Des* 2006;20:703–15. <https://doi.org/10.1007/s10822-006-9066-y>.
- [108] Stevenson JM, Mulready PD. Pipeline pilot 2.1 by Scitegic, 9665 Chesapeake drive, suite 401, San Diego, CA 92123-1365. www.scitegic.com. See web site for pricing information *J Am Chem Soc* 2003;125:1437–8. <https://doi.org/10.1021/ja025304v>.
- [109] Suhre WM, Ekins S, Chang C, Swaan PW, Wright SH. Molecular determinants of substrate/inhibitor binding to the human and rabbit renal organic cation transporters hOCT2 and rOCT2. *Mol Pharmacol* 2005;67:1067–77. <https://doi.org/10.1124/mol.104.004713>.
- [110] Tate CG, Schertler GFX. Engineering G protein-coupled receptors to facilitate their structure determination. *Curr Opin Struct Biol* 2009;19:386–95. <https://doi.org/10.1016/j.sbi.2009.07.004>.
- [111] The International Transporter Consortium, Giacomini KM, Huang S-M, Tweedie DJ, Benet LZ, Brouwer KLR, et al. Membrane transporters in drug development. *Nat Rev Drug Discov* 2010;9:215–36. <https://doi.org/10.1038/nrd3028>.
- [112] Tsigelny IF, Kovalsky D, Kouznetsova VL, Balinsky O, Sharikov Y, Bhatnagar V, et al. Conformational changes of the multispecific transporter organic anion transporter 1 (OAT1/SLC22A6) suggests a molecular mechanism for initial stages of drug and metabolite transport. *Cell Biochem Biophys* 2011;61:251–9. <https://doi.org/10.1007/s12013-011-9191-7>.
- [113] Türková A, Jain S, Zdrážil B. Integrative data mining, scaffold analysis, and sequential binary classification models for exploring ligand profiles of hepatic organic anion transporting polypeptides. *J Chem Inf Model* 2018. <https://doi.org/10.1021/acs.jcim.8b00466>.
- [114] van de Steeg E, Venhorst J, Jansen HT, Noolen IHG, DeGroot J, Wortelboer HM, et al. Generation of Bayesian prediction models for OATP-mediated drug–drug interactions based on inhibition screen of OATP1B1, OATP1B1+15 and OATP1B3. *Eur J Pharm Sci* 2015;70:29–36. <https://doi.org/10.1016/j.ejps.2015.01.004>.
- [115] van Westen GJP, Wegner JK, Ijzerman AP, van Vlijmen HWT, Bender A. Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets. *Med Chem Commun* 2011;2:16–30. <https://doi.org/10.1039/C0MD00165A>.
- [116] Verma J, Khedkar VM, Coutinho EC. 3D-QSAR in drug design—a review. *Curr Top Med Chem* 2010;10:95–115.
- [117] Wei BQ, Weaver LH, Ferrari AM, Matthews BW, Shoichet BK. Testing a flexible-receptor docking algorithm in a model binding site. *J Mol Biol* 2004;337:1161–82. <https://doi.org/10.1016/j.jmb.2004.02.015>.
- [118] Williams-Noonan BJ, Yuriev E, Chalmers DK. Free energy methods in drug design: prospects of “alchemical perturbation” in medicinal chemistry. *J Med Chem* 2018;61:638–49. <https://doi.org/10.1021/acs.jmedchem.7b00681>.
- [119] Wilson C, Gregoret LM, Agard DA. Modeling side-chain conformation for homologous proteins using an energy-based rotamer search. *J Mol Biol* 1993;229:996–1006. <https://doi.org/10.1006/jmbi.1993.1100>.

- [120] Wisedchaisri G, Park M-S, Iadanza MG, Zheng H, Gonen T. Proton-coupled sugar transport in the prototypical major facilitator superfamily protein Xyle. *Nat Commun* 2014;5:4521. <https://doi.org/10.1038/ncomms5521>.
- [121] Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* 2018;46:D1074–82. <https://doi.org/10.1093/nar/gkx1037>.
- [122] Wittwer MB, Zur AA, Khuri N, Kido Y, Kosaka A, Zhang X, et al. Discovery of potent, selective multidrug and toxin extrusion transporter 1 (MATE1, SLC47A1) inhibitors through prescription drug profiling and computational modeling. *J Med Chem* 2013;56:781–95. <https://doi.org/10.1021/jm301302s>.
- [123] Wong CF. Flexible receptor docking for drug discovery. *Expert Opin Drug Discovery* 2015;10:1189–200. <https://doi.org/10.1517/17460441.2015.1078308>.
- [124] Wu X, Prasad PD, Leibach FH, Ganapathy V. cDNA sequence, transport function, and genomic organization of human OCTN2, a new member of the organic cation transporter family. *Biochem Biophys Res Commun* 1998;246:589–95. <https://doi.org/10.1006/bbrc.1998.8669>.
- [125] Xu Y, Liu X, Li S, Zhou N, Gong L, Luo C, et al. Combinatorial pharmacophore modeling of organic cation transporter 2 (OCT2) inhibitors: insights into multiple inhibitory mechanisms. *Mol Pharm* 2013;10:4611–9. <https://doi.org/10.1021/mp400423g>.
- [126] Xu Y, Liu X, Wang Y, Zhou N, Peng J, Gong L, et al. Combinatorial pharmacophore modeling of multidrug and toxin extrusion transporter 1 inhibitors: a theoretical perspective for understanding multiple inhibitory mechanisms. *Sci Rep* 2015;5:13684. <https://doi.org/10.1038/srep13684>.
- [127] Yang J, Zhang Y. Protein structure and function prediction using I-TASSER. *Curr Protoc Bioinformatics* 2015;52:5.8.1–15. <https://doi.org/10.1002/0471250953.bi050852>.
- [128] Yang S-Y. Pharmacophore modeling and applications in drug discovery: challenges and recent advances. *Drug Discov Today* 2010;15:444–50. <https://doi.org/10.1016/j.drudis.2010.03.013>.
- [129] Young D. Computational chemistry: A practical guide for applying techniques to real world problems. John Wiley & Sons; 2004.
- [130] Zhang X, He X, Baker J, Tama F, Chang G, Wright SH. Twelve transmembrane helices form the functional core of mammalian multidrug and toxin extruder 1 (MATE1). *J Biol Chem* 2012. <https://doi.org/10.1074/jbc.M112.386979>.
- [131] Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. *Proteins* 2004;57:702–10. <https://doi.org/10.1002/prot.20264>.
- [132] Tanihara Y, Masuda S, Sato T, Katsura T, Ogawa O, Inui K-I. Substrate specificity of MATE1 and MATE2-K, human multidrug and toxin extrusions/H(+) -organic cation antiporters. *Biochem Pharmacol* 2007;74:359–71. <https://doi.org/10.1016/j.bcp.2007.04.010>.
- [133] Truong DM, Kaler G, Khandelwal A, Swaan PW, Nigam SK. Multi-level Analysis of Organic Anion Transporters 1, 3, and 6 Reveals Major Differences in Structural Determinants of Antiviral Discrimination. *J Biol Chem* 2008;283:8654–63. <https://doi.org/10.1074/jbc.M708615200>.
- [134] Soars MG, Barton P, Elkin LL, Mosure KW, Sproston JL, Riley RJ. Application of an in vitro OAT assay in drug design and optimization of renal clearance. *Xenobiotica* 2014;44:657–65. <https://doi.org/10.3109/00498254.2013.879625>.
- [135] Bednarczyk D, Ekins S, Wikel JH, Wright SH. Influence of molecular structure on substrate binding to the human organic cation transporter, hOCT1. *Mol Pharmacol* 2003;63:489–98.
- [136] Suhre WM, Ekins S, Chang C, Swaan PW, Wright SH. Molecular determinants of substrate/inhibitor binding to the human and rabbit renal organic cation transporters hOCT2 and rOCT2. *Mol Pharmacol* 2005;67:1067–77. <https://doi.org/10.1124/mol.104.004713>.
- [137] Hong W, Wu Z, Fang Z, Huang J, Huang H, Hong M. Amino Acid Residues in the Putative Transmembrane Domain 11 of Human Organic Anion Transporting Polypeptide 1B1 Dictate Transporter Substrate Binding, Stability, and Trafficking. *Molecular Pharmacology* 2015;12:4270–6. <https://doi.org/10.1021/acs.molpharmaceut.5b00466>.
- [138] van Westen GJP, Hendriks A, Wegner JK, Ijzerman AP, van Vlijmen HWT, Bender A. Significantly improved HIV inhibitor efficacy prediction employing proteochemometric models generated from antivirogram data. *PLoS Comput Biol* 2013;9. <https://doi.org/10.1371/journal.pcbi.1002899>.
- [139] Kalliokoski A, Niemi M. Impact of OATP transporters on pharmacokinetics. *Br J Pharmacol* 2009;158:693–705. <https://doi.org/10.1111/j.1476-5381.2009.00430.x>.
- [140] Stieger B, Hagenbuch B. Organic Anion Transporting Polypeptides. *Curr Top Membr* 2014;73:205–32. <https://doi.org/10.1016/B978-0-12-800223-0.00005-0>.
- [141] König J, Zolk O, Singer K, Hoffmann C, Fromm MF. Double-transfected MDCK cells expressing human OCT1/MATE1 or OCT2/MATE1: determinants of uptake and transcellular translocation of organic cations. *Br J Pharmacol* 2011;163:546–55. <https://doi.org/10.1111/j.1476-5381.2010.01052.x>.
- [142] Bahar I. On the functional significance of soft modes predicted by coarse-grained models for membrane proteins. *Journal of General Physiology* 2010;135:563–73. <https://doi.org/10.1085/jgp.200910368>.
- [143] Isin B, Tirupula KC, Oltvai ZN, Klein-Seetharaman J, Bahar I. Identification of Motions in Membrane Proteins by Elastic Network Models and Their Experimental Validation. *Methods Mol Biol* 2012;914:285–317. https://doi.org/10.1007/978-1-62703-023-6_17.
- [144] Dakal TC, Kumar R, Ramotar D. Structural modeling of human organic cation transporters. *Comput Biol Chem* 2017;68:153–63. <https://doi.org/10.1016/j.compbiolchem.2017.03.007>.
- [145] Zhang Y. I-TASSER server for protein 3D structure prediction. *BMC Bioinform* 2008;9:40.
- [146] Türková A, Jain S, Zdrazil B. Integrative Data Mining, Scaffold Analysis, and Sequential Binary Classification Models for Exploring Ligand Profiles of Hepatic Organic Anion Transporting Polypeptides. *J Chem Inf Model* 2018. <https://doi.org/10.1021/acs.jcim.8b00466>.

Part III

RESULTS

Chapter 3

Synopsis of Results

The following chapter represents an outcome of the original research performed in the course of this thesis. In Study 1-3 an extensive mining, curation, and analysis of ligand-based data (bioactivity measurements, compound substructures, molecular features) was performed. With the aim to elucidate structural aspects of ligand recognition for the selected hepatic transporters (here: OCT1, OATP1B1, OATP1B3, OATP2B1), three different structure-based modeling studies (Study 4-6) were conducted.

Study 1 (Section 3.1) was performed to integrate OATP bioactivity data from the open databases (ChEMBL, DrugBank, MetraBase, UCSF-FDA, and IUPHAR/Guide-to-Pharmacology). The merged substrate and inhibitor datasets were analyzed with respect to the data coverage, distribution of multiple bioactivity measurements per compound, and enriched substructures with pronounced selectivity profiles. Further, binary classification models were trained to extract important molecular features which would help explain OATP activity and selectivity.

In Study 2 (Section 3.2) we expand on the R-group decomposition and apply the methodology for studying congeneric series of 13-epiestrones, possessing subtle variations at the R-2,R-3, and R-4 positions. Presence of halogenated substituents at the R-2 position has been identified as an important molecular determinant of OATP2B1 inhibition. The conclusions drawn here suggest an existence of halogen bonds in OATP2B1-ligand complexes.

In Study 3 (Section 3.3) we further explore KNIME possibilities and implement another workflow-driven approach - a ligand-based drug repurposing. The usefulness of the workflow is demonstrated on the two cases (GLUT1-deficiency syndrom and COVID-19). The study is primarily conceived as a methodological paper which can be leveraged as teaching material.

In Study 4 (Section 3.4) differences in the uptake of clinical substrates (metformin and thiamine) between human and mouse hepatic OCT1 transporters were investigated. Human-mouse chimeric OCT1 construct have shown a simultaneous effect of TMH2 and TMH3 on OCT1 uptake activity. Computational modeling helped identify distinct tertiary interactions in human (between ILE35 at TMH1 and LEU155 at TMH2) and mouse (between ILE35 at TMH1 and VAL156 at TMH2) transporters. Replacement of LEU155 (human OCT1) to VAL156 (mouse OCT1) was experimentally confirmed. We conclude that difference in the uptake characteristics between the two species is associated to different ability to adopt hydrophobic packing interactions in between TMH1 and TMH2.

The last two sections contain so-far unpublished data from the structure-based modeling studies on hepatic OATPs. Study 5 (Section 3.6) was conducted to investigate signature dynamics of MFS proteins, to generate structural models for OATP1B1, OATP1B3, and OATP2B1 using ensemble docking, and to establish a binding mode hypothesis for compounds with steroidal scaffold. Differences in steroids binding across the three transporters are attributed to different electrostatics and shape complementarity. Especially, a single non-conserved residue at TMH1 (position 45 in OATP1B1/OATP1B3, corresponding position 66 in OATP2B1) was suggested to play key role as OATP selective switch given its both chemical specificity (i.e., ability to form hydrogen bonds) and regiospecificity (i.e., differences in pocket volume).

Study 6 (Section ??) presents novel OATP inhibitors identified upon a combination of different computational approaches (structure-based virtual screening, conformational prediction, proteochemometric and deep learning models, respectively), which were subsequently validated by the transporter inhibition assay. Initial screens identified 32% OATP1B1, 32% OATP1B3, and 70.5% OATP2B1 inhibitors (activity threshold ≤ 10 μ M). By subsequent full-dose response measurements IC_{50} values for eight selected compounds were determined. Four out of eight inhibitors possessed a high activity against OATP2B1 ($IC_{50} \leq 2,5$ μ M). Remarkably, one OATP2B1 inhibitor exhibits inhibitory activity at nanomolar range ($IC_{50} = 40$ nM), which outperforms the most potent OATP2B1 inhibitors known from other screening studies. By investigating binding modes of newly measured inhibitors we conclude that different localization of aromatic residues in OATP1B1/OATP1B3 versus OATP2B1 contributed to the identification of preferentially OATP2B1-active compounds.

The Supplementary Information for the individual studies can be found in Part V.

3.1 Integrative Data Mining, Scaffold Analysis, and Sequential Binary Classification Models for Exploring Ligand Profiles of Hepatic Organic Anion Transporting Polypeptides

TÜRKOVÁ, Alžběta; JAIN, Sankalp; ZDRAZIL, Barbara. *Journal of Chemical Information and Modeling*, **2018**, 59.5: 1811-1825.

* Corresponding author: barbara.zdrazil@univie.ac.at

A. Tuerkova generated the KNIME workflows, performed data fusion, curation, data analysis, data interpretation, and drafted the article. S. Jain developed binary classification models and helped interpreting the models. A. Tuerkova and B. Zdrazil analyzed molecular features from binary classification models. B. Zdrazil conceived the study design, analyzed and interpreted data, and critically revised the paper draft. The manuscript was written by the contribution of all authors. All authors approved the final version of the manuscript to be published.

The Supplementary Information can be found in Part V.

This article is reprinted with permission from:

Türkova, Alzbeta; Jain, Sankalp; Zdrazil, Barbara. Integrative data mining, scaffold analysis, and sequential binary classification models for exploring ligand profiles of hepatic organic anion transporting polypeptides. *Journal of Chemical Information and Modeling*, 2018, 59.5: 1811-1825. Copyright 2020 American Chemical Society.

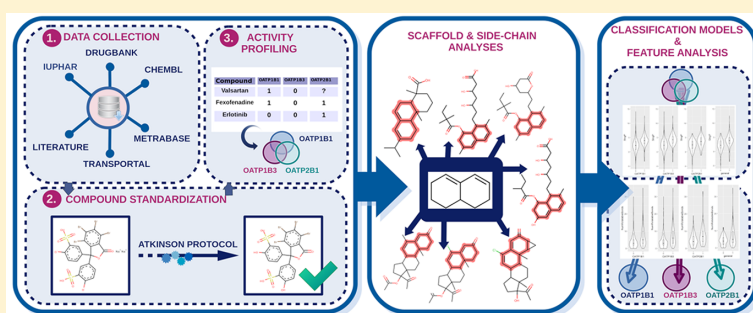


Integrative Data Mining, Scaffold Analysis, and Sequential Binary Classification Models for Exploring Ligand Profiles of Hepatic Organic Anion Transporting Polypeptides

Alžběta Türková, Sankalp Jain, and Barbara Zdrazil*[✉]

Department of Pharmaceutical Chemistry, Division of Drug Design and Medicinal Chemistry, University of Vienna, Althanstraße 14, A-1090 Vienna, Austria

Supporting Information



ABSTRACT: Hepatocellular organic anion transporting polypeptides (OATP1B1, OATP1B3, and OATP2B1) are important for proper liver function and the regulation of the drug elimination process. Understanding their roles in different conditions of liver toxicity and cancer requires an in-depth investigation of hepatic OATP–ligand interactions and selectivity. However, such studies are impeded by the lack of crystal structures, the promiscuous nature of these transporters, and the limited availability of reliable bioactivity data, which are spread over different data sources in the open domain. To this end, we integrated ligand bioactivity data for hepatic OATPs from five open data sources (ChEMBL, the UCSF–FDA TransPortal database, DrugBank, Metabase, and IUPHAR) in a semiautomatic KNIME workflow. Highly curated data sets were analyzed with respect to enriched scaffolds, and their activity profiles and interesting scaffold series providing indication for selective, dual-, or pan-inhibitory activity toward hepatic OATPs could be extracted. In addition, a sequential binary modeling approach revealed common and distinctive ligand features for inhibitory activity toward the individual transporters. The workflows designed for integrating data from open sources, data curation, and subsequent substructure analyses are freely available and fully adaptable. The new data sets for inhibitors and substrates of hepatic OATPs as well as the insights provided by the feature and substructure analyses will guide future structure-based studies on hepatic OATP–ligand interactions and selectivity.

INTRODUCTION

Organic anion transporting polypeptides (OATPs) belong to the SLCO (SLC21) superfamily of the solute carrier (SLC) group of membrane transport proteins, which mediate the transport of natural substrates as well as nutrients, clinically relevant drugs, and other xenobiotics across cellular membranes.¹ Here we focus on OATP1B1, OATP1B3, and OATP2B1 (encoded by the genes SLCO1B1, SLCO1B3, and SLCO2B1, respectively), all of which are expressed at the basolateral membrane of hepatocytes mediating the uptake of endogenous compounds like bile salts and bilirubin into liver cells. Therefore, hepatocellular OATPs are important for proper liver function and physiological processes like the enterohepatic circulation of bile salts² and bilirubin metabolism.³

Apart from the endogenous substrates (bile acids, steroid conjugates, hormones, and linear and cyclic peptides), hepatic

OATPs accept a broad spectrum of structurally unrelated pharmaceuticals, including antibiotics (e.g., rifampicin, benzylpenicillin, azithromycin, clarithromycin, and erythromycin⁴), antivirals (e.g., telaprevir⁵), anticancer drugs (e.g., rapamycin, SN-38, paclitaxel, docetaxel, and imatinib⁶), antifungals (e.g., caspofungin⁷), statins (e.g., pravastatin, rosuvastatin, and cerivastatin⁸), antihistamines (e.g., fexofenadine⁹), antidiabetics (e.g., repaglinide and rosiglitazone¹⁰), cardiac glycosides (e.g., digoxin¹¹), and anti-inflammatory drugs (e.g., diclofenac, ibuprofen, and lumiracoxib¹²). Importantly, impairment of the hepatic OATPs has been found to alter the pharmacokinetic profiles of various compounds and drugs, which can lead to

Special Issue: Women in Computational Chemistry

Received: July 13, 2018

Published: October 29, 2018



drug–drug interactions and consequently adverse drug reactions and liver toxicity.¹³

The substrate and inhibitor profiles of the three hepatic OATPs are partly overlapping, and some selective substrates and inhibitors are known (e.g., pravastatin for OATP1B1 and erlotinib for OATP2B1). Whereas hepatocytes are the exclusive location for the expression of OATP1B1 and OATP1B3, OATP2B1 is additionally expressed, e.g., in the intestine, the mammary gland, and the placenta and at the blood–brain barrier.¹⁴ Also, by sequence OATP2B1 is less related to the hepatic members of the OATP1 family (approximately 30%), and knowledge about this transporter is the least among the three in terms of available ligand data and biochemical studies. As our knowledge about all three hepatic OATPs is increasing, we will learn more about their interplay with respect to the delivery and disposition of endogenous substances and drugs. These efforts are impeded by the lack of crystal or NMR structures of any member of the OATP family to be used as templates for structure-based modeling as well as the limited availability of high-quality bioactivity data, which are spread over different data sources in the public domain. Furthermore, the promiscuous nature of hepatic OATPs turns modeling efforts into even more challenging tasks.

Several ligand-based computational studies have been performed to predict hepatocellular OATP–ligand interactions, with a predominance of studies focusing on inhibitors of the structurally more closely related transporters OATP1B1 and OATP1B3 (approximately 80% sequence identity). For example, de Bruyn et al.¹⁵ carried out *in vitro* high-throughput screening of almost 2000 potential molecules against OATP1B1 and OATP1B3, which identified 212 inhibitors for OATP1B1 and 139 inhibitors for OATP1B3. Subsequently, proteochemometric modeling for predicting OATP1B1/1B3 inhibitors was applied. In other studies, Bayesian models for OATP1B1 and its mutated form OATP1B1*15 were employed for inhibitor prediction,¹⁶ and Kotsampasakou et al.¹⁷ used six *in silico* consensus classification models to predict OATP1B1 and OATP1B3 inhibition. With respect to OATP2B1, only very few computational studies are available to date, likely because of the shortage of available data for this member of the hepatic OATPs. Just recently, Giacomini and co-workers addressed this shortcoming by combining biochemical studies with *in silico* ligand-based and structure-based approaches for the identification of novel OATP2B1 inhibitors.¹⁸

To the best of our knowledge, only one study is available comparing the inhibitory activity profiles of 225 compounds on these three hepatocellular OATPs. In that study, 27, 9, and 3 specific inhibitors of OATP1B1 (e.g., amprenavir, indomethacin, rosiglitazone, and spironolactone), OATP2B1 (e.g., erlotinib, astemizole, piroxicam, and valproic acid), and OATP1B3 (Hoechst 33342, mitoxantrone, and vincristine), respectively, were identified.¹⁹

In the present work, we expanded on the investigations by Karlgren et al.,¹⁹ including in our study different aspects related to the chemical structures of the ligands contributing to hepatic OATP–ligand interactions or selectivity. Since the major aim of this study was to perform an in-depth investigation of ligand availability, ligand profiles, and ligand properties across the three related transporters, we started our analysis with an extensive data curation exercise by integrating ligand data from various open data sources via semiautomatic

KNIME²⁰ workflows. By fusing ligand bioactivity data from five different databases (ChEMBL,²¹ the UCSF–FDA TransPortal database,²² DrugBank,²³ Metrabase,²⁴ and IUPHAR²⁵), we could increase the size of the data sets, their coverage of chemical space, and the confidence in the data quality by considering data from multiple independent bioactivity measurements. In order to retrieve reliable annotations for activity and selectivity, we filtered out ambiguous compounds from multiple independent measurements. In order to be able to systematically annotate a compound as either an inhibitor or noninhibitor or as a substrate or nonsubstrate, we considered the different bioactivity end points as well as different activity annotations or activity comments available in the respective databases. As a result, a total of six high-quality data sets including selective, dual-selective, and pan-interacting ligands for OATP1B1, OATP1B3, and OATP2B1 were retrieved, treating inhibitors and substrates separately.

As we were interested in the structural determinants of ligand selectivity, scaffold decomposition was applied, and frequently occurring scaffolds per transporter were inspected further. Here the focus was on the extraction of frameworks with a higher prevalence for just one or two of the three transporters. Scaffold series of this kind will be important candidates for future detailed structure–activity relationship (SAR) studies (including, e.g., molecular docking). We also looked for pan-interacting scaffolds (e.g., the steroidal scaffold and its conjugates derived from natural substrates). These interesting cases can provide information on the influence of side chains in conferring selectivity switches.

Finally, binary classification modeling by using hierarchical levels for compound classification (sequential binary classification models) revealed important descriptors that might trigger ligand activity or selectivity.

Here, we present an integrative, semiautomatic data mining approach that combines data from various open data sources, preprocesses and curates the data, and analyzes the chemical compounds with respect to chemical features related to transporter selectivity.

The novel high-quality data sets for OATP1B1, OATP1B3, and OATP2B1 for (non)inhibitors and (non)substrates are provided in the [Supporting Information](#), and the [data mining workflows](#) (which can be reused for ligand profiling on other related targets of interest) are described. Insights provided by the scaffold and substructure analyses as well as the binary classification modeling will be helpful for subsequent ligand- and structure-based *in silico* and *in vitro* studies investigating novel tool compounds for hepatic OATPs.

■ MATERIALS AND METHODS

Fetching Data from Different Sources. KNIME Analytics Platform²⁰ (version 3.4) is an open-source solution for the automatization of data integration and analysis that is extensively used in the field of chemoinformatics. Here we created (semi)automatic KNIME workflows for integrative data mining from the open domain.

Bioactivity measurements and/or annotations (substrate, nonsubstrate, inhibitor, noninhibitor) were fetched from five different sources: ChEMBL,²¹ the UCSF–FDA TransPortal database,²² DrugBank,²³ Metrabase,²⁴ and IUPHAR.²⁵ In addition, three novel OATP2B1 (non)inhibitors from Khuri et al.¹⁸ as well as ten novel OATP1B1 and OATP1B3 (non)inhibitors from Kotsampasakou et al.¹⁷ were manually added to the data set.

Ligands from ChEMBL23 were collected via RESTful web services by providing UniProt protein accession numbers for OATP1B1 [Q9Y6L6], OATP1B3 [Q9NPD5], and OATP2B1 [O94956] to the “ChEMBLdb Connector” node. Data sets retrieved from the UCSF–FDA TransPortal do not contain any type of structural format. Therefore, an automated “name-to-structure” mapping workflow was created to retrieve InChIKeys according to generic names using PubChem’s (<https://pubchem.ncbi.nlm.nih.gov>) PUG REST services. URL links for retrieving compound identifiers (CIDs) from PubChem were created by inserting the compound names as variables. Records with CIDs were downloaded in XML file format by the “GET Request” node, and the CIDs were extracted (“XPath”). In the case of multiple CIDs for a single entity, only the first one was retained. Unmapped compounds were curated manually. Furthermore, InChIKeys for the respective CIDs were retrieved (“GET Request” node) in XML format and further extracted via an “XPath” query. The quality of the bioactivity measurements from ChEMBL was also assessed by the confidence score. This parameter is included in all ChEMBL entries and evaluates the assay-to-target relationships, ranging from 0 (i.e., so-far uncurated entries) to 9 (i.e., high confidence level of the data). The curated ChEMBL data in our data set have high confidence scores of 9 (898 bioactivities) or 8 (2487 bioactivities), which is a positive indicator of the quality of our curated data sets.

Data from DrugBank and IUPHAR were fetched from the UniProt webpage by downloading the respective XML (DrugBank) and JSON (IUPHAR) files for human OATP1B1, OATP1B3, and OATP2B1. Compound identifiers, compound names, and standard InChIKeys were further extracted via the “XPath” or “JSON Path” node. Metabase data were fetched from its website using the “HttpRetriever” and “HtmlParser” nodes. The HTML document was processed via an “XPath” query to retrieve the compound names and the associated activity values. InChIKeys for Metabase compounds were retrieved from PubChem using the same procedure as for UCSF–FDA TransPortal data.

Data Preprocessing and Curation and Assignment of Binary Activity Labels. For each data source, the ligand data were split into two different tables to treat the substrates and inhibitors separately. First, assignment was done on the basis of the “Activity annotation” (substrate, nonsubstrate, inhibitor, or noninhibitor), if available. If the manual activity annotation was not available, the “bioactivity_type” was used as a criterion for classification as either a substrate or inhibitor. For substrates, data entries with either K_m or EC_{50} end points were considered. For inhibitors, data entries with K_i , IC_{50} , and/or percentage inhibition were considered. Potential data errors (activity values greater than 10^8) were removed, as were data points with missing activity values.

For all end points except percentage inhibition, activity units other than nanomolar (e.g., micromolar) were converted into nanomolar units and further into their negative logarithmic molar values ($-\log \text{Activity}$ [molar]). The distribution of bioactivity measurements for each transporter was analyzed systematically in order to be able to rationally select a good cutoff for the separation of actives from inactives. A compound was defined as active if the bioactivity was $<10 \mu\text{M}$ and inactive if the bioactivity was greater than or equal to $10 \mu\text{M}$. Data with percentage inhibition values were inspected further since we noted that some of them were rather measurements of uptake stimulation. Data with such inverse expression of the inhibitory

effect (i.e., “% of control”) were converted into direct inhibition values ($100 - [\% \text{ of control}]$). Values greater than 100% were interpreted as 100%.

Classification of percentage inhibition data into actives and inactives was done on the basis of recommended thresholds that were manually extracted from primary literature sources (detailed information is available in [Tables S1 and S2](#)). If no threshold was recommended but in one of the other sources the same compound concentration was used, the threshold was adopted accordingly. If such information was not available, the data point was removed from the data set.

Percentage inhibition data with negative values (interpreted as “stimulators of uptake”) were filtered out of the data set. Retrieved chemical compounds were further standardized via the Atkinson standardization protocol (available at <https://www.ebi.ac.uk/chembl/extra/francis/standardiser/>). This procedure includes breakage of covalent bonds between oxygen/nitrogen atoms and metal atoms, charge neutralization, application of structure normalization rules (e.g., proton shift between heteroatoms, protonation of bicyclic heterocycles, or correction of charge conjugation), and removal of salt/solvent. All of the incorrectly standardized compounds were filtered out (24 compounds). Compounds from various data sets were subsequently grouped by their standardized InChIKeys. If multiple measurements for a single compound/target pair were available, the median activity label was retained. Compounds with conflicting activity labels [median activity label (mean of middle values) = 0.5] were sorted out. All of the compounds with contradictory activity labels are listed in the [Supporting Information](#) [[Tables S3–S5](#) for (non)substrates and [Tables S6–S8](#) for (non)inhibitors]. A pivot table was generated by grouping the data by compounds (standardized InChIKeys) and targets. The applied data mining procedure is visually depicted in [Figure S1](#).

Scaffold Generation and Clustering. The three hepatic OATPs were analyzed with respect to privileged scaffolds. Murcko scaffolds²⁶ were extracted via the “RDKit Find Murcko Scaffolds” node in a targetwise manner. The obtained scaffolds were used as queries for substructure mining against the sparse data set for the respective target for the sake of enrichment of existing clusters by additional molecules with analogous scaffolds (since the addition of (a) ring(s) leads to a novel Murcko scaffold). The relative occurrences of scaffolds in the “active” and “inactive” activity classes were subsequently calculated, and only scaffolds with higher prevalence in the “active” class were kept. Generic scaffolds (i.e., those composed of only one aromatic ring with zero or one heteroatom) were filtered out. The Fisher exact test was applied to keep only statistically significant scaffolds ($p < 0.05$, unless otherwise stated). Hierarchical scaffold clustering [“Hierarchical Clustering (DistMatrix)” node] was applied for scaffolds that appeared in multiple data sets (for different OATPs) by calculation of their maximum common substructure as a measure of similarity. Scaffolds were assigned to discrete clusters on the basis of their distance threshold (set to 0.7). Retrieved compounds belonging to a particular cluster were selected in cases where they exerted the same pharmacological profile as the parent scaffold. All inadequate compounds were reassigned to a corresponding scaffold cluster.

The same analysis was repeated with the dense data set (compounds with measurements for all three hepatic OATPs) in order to retrieve enriched scaffolds with a full pharmaco-

logical profile. We also repeated the analysis with full dose–response curve data only (excluding percentage inhibition data) in order to be able to see whether major trends in enriched scaffolds persist with data of higher confidence.

Side-Chain Analysis. The SMARTS pattern for steroidal scaffolds was generated as a query for substructure mining with the aim of detecting all steroid-associated compounds in the sparse data set. The “A” ring (according to IUPAC nomenclature) was defined to be less structurally restricted in order to search for both sp^3 - and sp^2 -hybridized carbocycles (estrone-like and cholate-like).

The “RDKit R Group Decomposition” node was used to identify all distinct side chains across the given steroidal scaffold of retrieved compounds. The frequencies of side-chain attachment to different positions of steroidal scaffolds for the different hepatic OATPs were subsequently calculated.

Semiautomatic KNIME Workflows. Workflows for fetching data from different sources, scaffold clustering and analysis, and side-chain analysis are available from myExperiment (<https://www.myexperiment.org/workflows/5097.html>; <https://www.myexperiment.org/workflows/5098.html>).

Data Sets for Binary Classification Models: Training and Test Set Selection. Predictive binary classification models were generated in KNIME in order to identify driving factors for inhibitory activity (and eventually selectivity) in terms of molecular features. Only data on transport inhibition were considered, representing data sets more comprehensive than that for substrates/nonsubstrates. Seventy percent of each class was randomly selected to be used as the training set; the remaining compounds were considered as the test set. The compositions of the resulting data sets are shown in Table 1.

Table 1. Compositions of the Data Sets Used in the Sequential Binary Classification Modeling

transport inhibition data	total	inhibitor	noninhibitor
all inhibitors + general noninhibitors (training set)	324	262	62
all inhibitors + general noninhibitors (test set)	139	113	26
OATP1B1 training set	937	232	705
OATP1B1 test set	403	100	303
OATP1B3 training set	875	139	736
OATP1B3 test set	375	59	316
OATP2B1 training set	161	43	118
OATP1B1 test set	69	19	50

Descriptor Calculation and Feature Selection.

Twenty-six two-dimensional descriptors representing interpretable physicochemical properties were calculated using the “RDKit Descriptor Calculation node” in KNIME. The most relevant descriptors for the respective data set were selected using the “CfsSubsetEval” algorithm implemented in Weka²⁷ with the “BestFirst” search method. Weka is an open-source tool comprising different machine learning algorithms. The exact list of descriptors is given in Tables S13–S16.

Machine Learning Models. Weka²⁷ nodes implemented in KNIME²⁸ were used to train binary classification models for inhibitors of OATP1B1, OATP1B3, and OATP2B1. “Random tree”^{29,30} (with default parameters) was used as the base classifier. In order to overcome the problem of data imbalance, two different meta-classifiers were used: a cost-sensitive classifier³¹ and stratified bagging.^{32,33} In case of the cost-

sensitive classifier, the misclassification was applied in accordance with the imbalance ratio. For stratified bagging, the number of bags was adjusted to 64, as a previous study^{33,34} suggested that generation of 64 models provides satisfactory results without exponentially increasing the computational cost.

Evaluation Method. All of the models were validated by 10-fold cross-validation and by their performances on the external test sets. In both validation schemes, the confusion matrix, sensitivity, specificity, balanced accuracy, and Matthews correlation coefficient (MCC) are reported as measures of the predictive power of the models.

Analyzing Important Molecular Features for OATP Inhibition. The features appearing as most relevant for hepatic OATP inhibition (as selected by the feature selection methodology) were further analyzed by plotting the distribution of their values for inhibitors versus noninhibitors for the three hepatic OATPs and the level 1 (general inhibitors) data set. These analyses as well as the calculations of the statistical significance of the pairwise comparisons of the distributions using the Wilcoxon test were done in R version 1.0.143. The R Project is a software for statistical analysis and data visualization and is freely available at <https://www.r-project.org/>.

RESULTS AND DISCUSSION

Semiautomatic Integration of Pharmacological Data from Different Sources. Compound bioactivity data on human OATP1B1, OATP1B3, and OATP2B1 were collected, mapped, and integrated from five different data sources openly available in the public domain: ChEMBL,²¹ Metrabase,²⁴ DrugBank,²³ the UCSF–FDA TransPortal database,²² and IUPHAR/Guide to Pharmacology.²⁵ The motivation for curating data sets from such a large number of different sources was the wish to enhance the particular data sets not only in terms of their unique enumerated compounds but also in terms of chemical space. Since the different data sources focus on different aspects of bioactivity data (e.g., ChEMBL contains literature data from primarily SAR series, Metrabase has a focus on transporter substrates, and DrugBank contains a collection of marketed or withdrawn drugs), it can be expected that a greater variety in some molecular properties of pharmaceutical interest (e.g., lipophilicity, molecular weight, topological polar surface area, and the number of rotatable bonds) would be introduced by integrating these various sources. As shown in Figure S2, all four features are significantly different in the other databases (DrugBank, Metrabase, IUPHAR, TransPortal) compared with ChEMBL (the Wilcoxon test revealed $p < 0.05$ in all pairwise comparisons; data not shown), which illustrates the different constitution of the five considered data sources.

A major goal in this study was the generation of the most comprehensive data sets for hepatic OATPs available from the open domain. These data sets should reflect both the state of the art of available inhibitor and substrate compound spaces, and there was a particular attempt to separate the two sets. This objective was achieved by classifying compounds according to different types of activity end points (K_m and EC_{50} for substrates; IC_{50} , K_i , and percentage inhibition for inhibitors) and activity annotations (substrate, nonsubstrate, inhibitor, or noninhibitor). Interestingly, in terms of the increase in the size of the data sets achieved by integrating data from different sources, the situation looks strikingly different

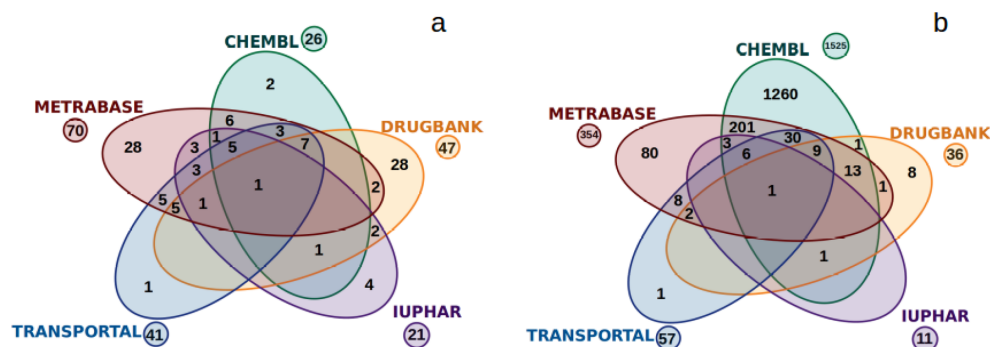


Figure 1. Venn diagrams showing the contributions from the different data sources (in terms of the numbers of unique compounds extracted and curated from them) to the final data sets for (a) (non)substrates and (b) (non)inhibitors.

for inhibitor data sets versus substrate data sets for hepatic OATPs (see Figure 1).

Whereas ChEMBL accounts for the largest collection of compounds contributing to the inhibitor data set (1525 unique compounds; 94% of all unique inhibitors/noninhibitors), for substrates, Metrabase (70 unique compounds; 69% of all unique substrates/nonsubstrates) and DrugBank (47 unique compounds; 46%) were identified as the most useful resources. Interestingly, just 25% (26 unique compounds) of all substrates/nonsubstrates could be retrieved from ChEMBL, which indeed justifies the integration of data from various sources, especially when it comes to investigations on transporter substrates.

Metrabase²⁴ was originally created to serve as a large open source for transporter ligand data with a special focus on substrates. In total, 631 substrates, 183 nonsubstrates, 1256 inhibitors, and 370 noninhibitors of hepatic OATPs are currently reported in Metrabase. Nevertheless, only a minority of the data entries in Metrabase also report distinct bioactivity values; instead, mostly the data are presented with activity annotations only (e.g., substrate, nonsubstrate, inhibitor, or noninhibitor). However, it is unclear how the data curators decided upon the particular annotations in certain cases. To give an example, primovist was defined as an OATP1B3 substrate, having $K_m = 4.1$ mM.³⁵ On the other hand, clarithromycin was classified as an OATP1B3 nonsubstrate on the basis of its reported K_m value of 1 μ M.³⁶ In order to further assess the confidence of Metrabase entries, activity annotations from Metrabase were compared with annotations that were assigned to bioactivity measurements from ChEMBL (for the chosen cutoff for classifying actives/inactives, see below). Strikingly, we found conflicting annotations for up to 74% of the compounds retrieved from Metrabase (see Table S9). Thus, only Metrabase entries including numerical bioactivity values were included in our final data sets. Consequently, only 60 substrates/nonsubstrates (7% of the available substrates in Metrabase) and 350 inhibitors/noninhibitors (22% of the available inhibitors in Metrabase) from Metrabase are part of our final data sets for hepatic OATPs.

DrugBank is a comprehensive repository comprising detailed descriptions of small-molecule drugs and their associated targets. Drug activity linked to a respective target is expressed in the form of activity annotations (e.g., substrate, inhibitor, unknown, stimulator, activator, or reducer). Interestingly, DrugBank provided quite a balanced number of both (non)substrates (47 unique compounds) and (non)inhibitors

(36 unique compounds) for our final data sets. A similar number of total compounds was included from the UCSF–FDA TransPortal database, but with a predominance of (non)inhibitors (57 unique compounds) over (non)substrates (27 unique compounds). Providing data about FDA-approved drugs linked to pharmaceutically relevant targets, UCSF–FDA TransPortal comprises numerical bioactivity measurements (e.g., K_m , IC_{50} , K_i) for hepatic OATPs. The source with the lowest number of compounds for hepatic OATPs [21 unique (non)substrates, 11 unique (non)inhibitors] turned out to be IUPHAR, which provides both real activity measurements and/or annotations for all licensed drugs and other ligands of biologically relevant targets, including transporters. It mainly provided additional information about the hepatic OATP natural substrates. Finally, three novel OATP2B1 inhibitors/noninhibitors recently reported by Giacomini and co-workers¹⁸ and 10 novel OATP1B1 and OATP1B3 inhibitors/noninhibitors reported by the group of Ecker¹⁷ (just one compound, sirolimus, has been annotated to be a OATP1B1 inhibitor in DrugBank before) were also manually added to the data sets.

In addition to enrichment in terms of chemical space and data set size, we sought to increase the confidence in the final data annotations (as actives or inactives) by collecting multiple independently measured bioactivities or activity annotations for compound/target pairs. Box plots showing the distributions of the number of bioactivities/annotations per single compound and transporter are shown in Figure S3.

For the sake of establishing quantitative SAR (QSAR) models, it is not advisable to mix data from different bioactivity end points or different assay setups.^{37,38} When it comes to binary classification (e.g., into actives and inactives), however, the final label (e.g., inhibitor or noninhibitor) should be independent of the specific experimental protocol.³⁹ Combining data from different activity end points can thus provide a more accurate perception of the OATP pharmacological profiles since measurement errors will be detected and sorted out to a higher extent.

Data Curation. Once the small-molecule bioactivity data had been successfully fetched, the compound data had to be mapped across the various sources in order to identify all assays/bioactivity measurements for a particular compound against one particular target but also across the three different transporters. Hereby, the availability of encoded chemical structures (in the form of InChIKeys, InChIs, or SMILES) was a great advantage. However, this information is not implicitly

included in all of the databases used herein (e.g., the UCSF–FDA TransPortal provides only generic names for the compounds). In such cases, the Chemical Identifier Resolver (CIR) web service provided by NIH (available at <https://cactus.nci.nih.gov/chemical/structure>) can be used in order to assign chemical structural information (SMILES, InChI, InChIKey, etc.) to a compound's generic name.⁴⁰ Since for our data sets this procedure failed for 132 compounds, we generated in house a fit-for-purpose “name-to-structure” conversion workflow that retrieves standard InChIKeys from the PubChem database. The majority of these compounds could be mapped by this procedure (68%); however, for 41 compounds the mapping failed because of the wide range of compound expressions and associated synonyms. InChIKeys were manually added in these cases.

All of the precurated entries were subjected to Atkinson's standardization procedure. To account for consistency during mapping of data from different sources, unified standard InChIKeys were calculated from standardized compounds.

The selected cutoff for separating actives from inactives at 10 μ M appears as a good choice upon inspection of the distribution of the median bioactivities for each target since we can observe a certain plateau when looking at the density plots (see Figure S4).

Setting the cutoff for percentage inhibition values resulted in a more complicated procedure. As can be seen from Table 2,

Table 2. Numbers of Unique Compounds for Different Activity End Points

	K_i	IC_{50}	K_m	EC_{50}	% inhibition	manual annotation
(non) substrates	—	—	74	4	—	63
(non) inhibitors	170	236	—	—	1526	45

percentage inhibition values account for approximately 77% of entries from the overall inhibitor data set. Interestingly, the interpretation of percentage inhibition values is highly inconsistent in different data sets originating from different articles. In the case of ChEMBL entries, three out of 11 integrated data sets reported percentage inhibition values in the form of the inhibitory effect, i.e., the higher the value, the stronger the inhibitor. However, the remaining eight data sets present inhibition as a percentage of control (also expressed as “residual activity”), i.e., the lower the value, the stronger the inhibitor. Interpretation of ChEMBL data gets even more complicated, as some of the data (e.g., the data set reported by Nozawa et al.⁴¹) were converted to the opposite form of percentage inhibition values prior to being uploaded to ChEMBL. Since a strict removal of entries with percentage inhibition values would have resulted in a tremendous reduction in the compound numbers of the inhibitor data sets, we manually curated these data sets and transformed the data into a uniform representation of the activity end point “percentage inhibition”. For the ~150 data sets with percentage inhibition data provided by Metrabase, this curation exercise was alleviated by the availability of activity comments [“Uptake/Inhibition (% of control)” or “Inhibition”]. Cutoffs for separating inhibitors and noninhibitors were set individually on the basis of recommendations given in the primary literature (Tables S1 and S2). The assignment of activity labels was done prior to the creation of a

pharmacological overlap matrix. Consequently, compounds with conflicting activity measurements (i.e., equivalent frequencies of the active and inactive binary labels) could be sorted out during this important step of mapping standard InChIKeys in order to represent the whole data set together with their activity labels toward the three transporters. Activity labels for more than 65% of the compounds of the final data set were assessed on basis of more than a single bioactivity measurement. To give an example, we retrieved 59 independent data points (measured bioactivities and/or pure annotations) for cyclosporine from all of the integrated databases, including 19 values from ChEMBL (14 K_i/IC_{50} and five percentage inhibition values), 26 values from Metrabase (22 K_i/IC_{50} and four percentage inhibition values), 12 K_i/IC_{50} values from the UCSF–FDA TransPortal, and two IC_{50} values from IUPHAR.

For the subsequent analyses on chemical fragments and features, two different data sets were generated. The “sparse hepatic OATP data set” comprises the whole data matrix (including missing annotations for one or two of the transporters) and is made up of 102 unique substrates/nonsubstrates and 1630 unique inhibitors/noninhibitors (see Table 3 for the respective data subset compositions). The

Table 3. Constitution of the “Sparse Hepatic OATP Data Set”: Numbers of Compounds Per Annotation and Transporter Are Shown (Compounds Might Appear Annotated to More than One Target)

activity	OATP1B1	OATP1B3	OATP2B1
substrates	53	45	26
nonsubstrates	19	16	6
inhibitors	332	198	62
noninhibitors	1008	1052	168

“dense hepatic OATP data set”, however, comprises only 13 substrates and 163 inhibitors whose bioactivities have been measured against all three hepatic OATPs [see Table S10 for (non)substrates and Table S11 for (non)inhibitors]. Data from the latter data set provide information about general (i.e., completely overlapping), partially overlapping, and selective substrates/inhibitors. Both data sets are useful sources for studying features that are potentially important for hepatic OATP ligand activity or selectivity.

Scaffold Clustering and Analysis. First, the analysis on structural determinants for ligand interaction and selectivity among hepatic OATPs was conducted at the scaffold level. As demonstrated previously by looking at the distributions of certain chemical features in the different data sources (Figure S2), adding data sources led to an increase in chemical space. In terms of new scaffolds, the addition of data from the UCSF–FDA TransPortal database, DrugBank, Metrabase, IUPHAR, and the literature to the data from ChEMBL also led to a gain in terms of new chemical scaffolds (as demonstrated for OATP1B1 inhibitors in Figure 2). Visualizations of new chemical scaffolds for OATP1B3 and OATP2B1 inhibitors are included in Figures S5 and S6, respectively.

In order to analyze the frequencies of scaffolds across the different transporters, compounds were grouped by their Murcko scaffolds²⁶ for each transporter. We have to point out that although these analyses were carried out for inhibitors and substrates separately, the majority of the results discussed here

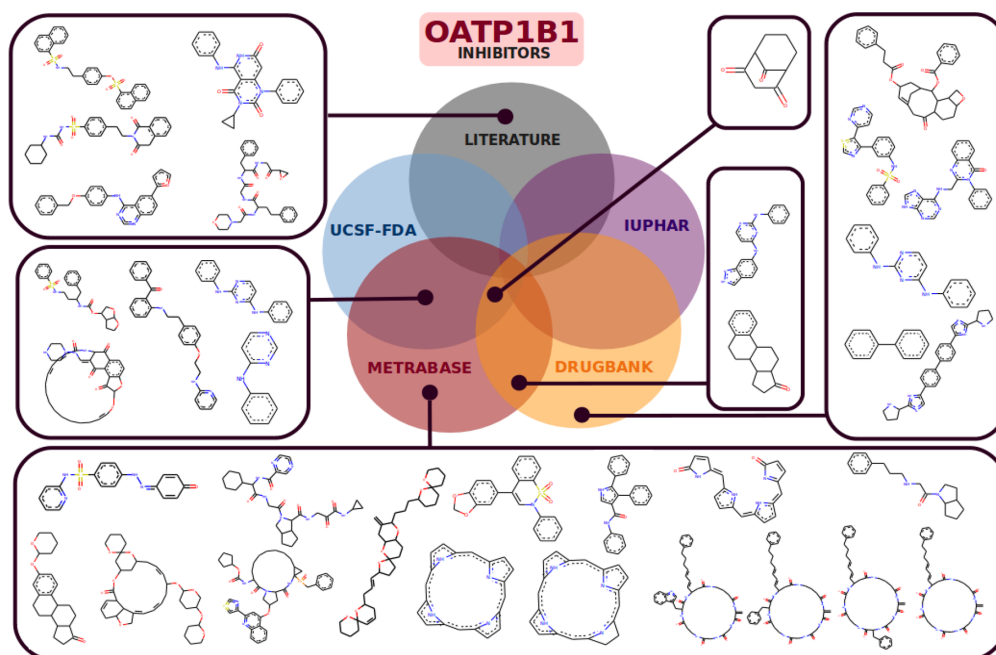


Figure 2. Murcko Scaffolds for OATP1B1 inhibitors retrieved from databases other than ChEMBL.

were derived from inhibitor data because of data sparseness for substrates in that domain.

The large number of different scaffolds (reflected by the scaffold-to-compound ratio; Table 4)⁴² strongly indicates that

Table 4. Numbers of Unique Scaffolds in Substrate and Inhibitor Data Sets and (in Parentheses) Their Scaffold-to-Compound Ratios

	OATP1B1	OATP1B3	OATP2B1
substrates	43 (0.86)	39 (0.86)	23 (0.88)
inhibitors	250 (0.75)	155 (0.78)	54 (0.87)

OATP ligands are structurally highly diverse compounds. However, a few scaffolds (23 for inhibitors) were significantly enriched in actives versus inactives (Fisher's exact test, $p < 0.05$; see Figure 3).

One limitation of the scaffold algorithm of Bemis and Murcko²⁶ is the fact that adding (an) additional ring(s) leads to a new Murcko scaffold. Therefore, for detecting congeneric SAR series of compounds sharing a common scaffold within a data set, the grouping by scaffolds should be combined with additional substructure searches.⁴³ In our case, this strategy has proven useful, e.g., in order to find additional structural analogues of pravastatin-like compounds in the inhibitor data set. In the first instance, only three compounds sharing a hexahydronaphthalene scaffold were detected in the 1B1 inhibitor data set, with pravastatin being a selective inhibitor for OATP1B1 (lovastatin acid and tenivastatin are OATP1B1 inhibitors but have unknown activity toward the other two transporters). By the subsequent substructure search, we could retrieve seven additional compounds with a hexahydronaphthalene substructure but with some variation in their activity profiles (see Table S12). While six compounds show activity

against OATP1B1, some do possess additional activity against one of the other two transporters. A closer look at their structures revealed that potentially the addition of more rings, leading to three- or four-ring systems, is responsible for the shift in activity, turning them into unselective hepatic OATP inhibitors (also see the discussion on steroidal scaffolds below).

After enrichment of the scaffold series with additional compounds (by substructure searches), their pharmacological profiles were inspected in order to identify scaffolds with a pronounced activity for only one OATP, for two OATPs (dual inhibitors), or for all three OATPs (pan inhibitors). Furthermore, hierarchical scaffold clustering was applied in order to group structurally similar scaffolds with the same selectivity profile. Within the inhibitor data set, this procedure led to seven enriched scaffold clusters for OATP1B1 (eight scaffolds) and 11 enriched scaffold clusters for both OATP1B1 and OATP1B3 (15 scaffolds) (see Figure 3). Of course, this analysis is influenced by data availability/sparseness and by no means reflects a complete picture of the pharmacological profiles (which especially accounts for the less investigated target OATP2B1).

In order to be able to sort out scaffolds where a real selectivity claim can be made (compared with just enriched scaffolds without a complete pharmacological profile for hepatic OATPs) we applied the scaffold frequency analysis to the dense data set as well. This analysis delivered two scaffolds with indications for OATP1B1 selectivity (pravastatin-like scaffold, estrone-like) and one scaffold with an indication for OATP1B subfamily selectivity (cyclosporin-like scaffold) (Figure S7). In these cases, the available full pharmacological profiles indicate inactivity toward the other targets.

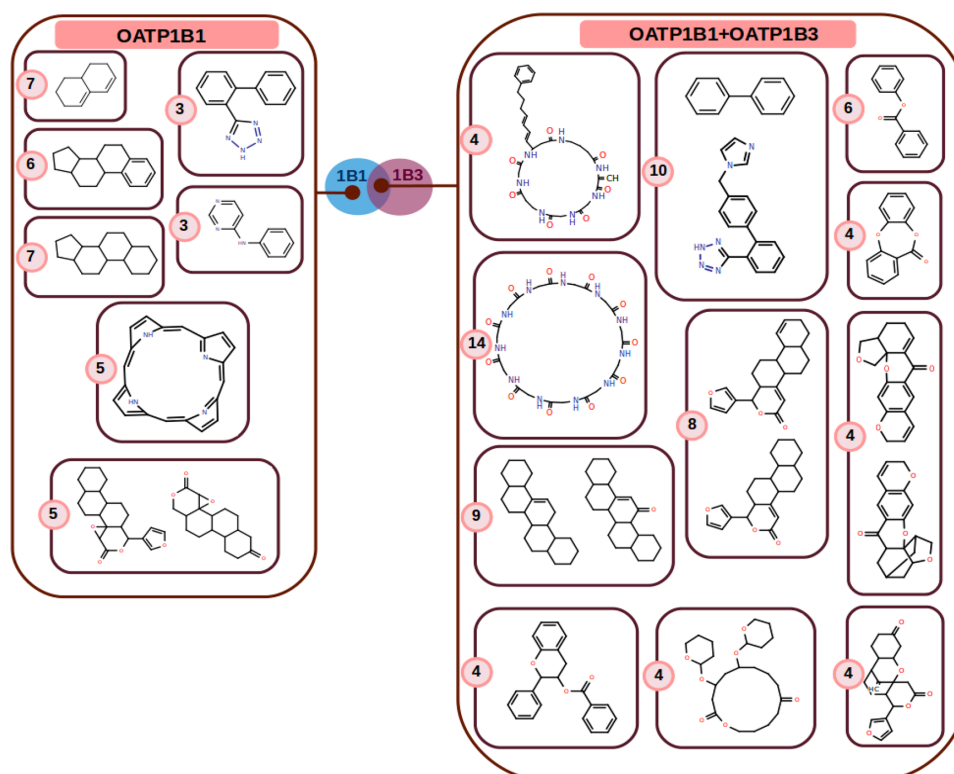


Figure 3. Enriched scaffolds ($p < 0.05$) for hepatic OATP inhibitors grouped by their pharmacological profiles with respect to hepatic OATPs. Numbers in pink circles are the numbers of associated compounds for the respective scaffold clusters.

We were also interested in whether some of the trends in enriched scaffolds would remain if the analysis were repeated with full dose–response curve data only. As can be seen from Figure S8, upon exclusion of percentage inhibition data points, most of the enriched scaffolds persisted (20 scaffolds out of 23).

Enriched Scaffolds for OATP1B1 Inhibitors. As shown in Figure 3, frequently occurring scaffolds among the OATP1B1 inhibitors (eight scaffolds) can be grouped into seven different clusters with the available data. Some of the most populated clusters are those comprising steroid derivatives (estrone derivatives and cholate derivatives), with 13 associated compounds in total (six and seven compounds, respectively). The scaffold made up of pravastatin-like compounds, as already discussed above, is also among the most frequent ones for OATP1B1. The seven member compounds have been detected as either OATP1B1-selective inhibitors (pravastatin, simvastatin, and mevinolin) or as OATP1B1 inhibitors (e.g., cyproterone and lovastatin acid; no measurements against OATP1B3 and OATP2B1) in our data sets. Another cluster is derived from porphyrin (five associated compounds). This scaffold has been suggested for the design of new tool compounds for therapeutic applications, mainly because of its photodynamic effects against ovarian cancer. Current findings show that porphyrin and its derivatives exert inhibitory activity against OATP1B1.⁴⁴ There is also evidence from activity measurements for OATP1B3, suggesting that protoporphyrin acts as a noninhibitor against OATP1B3.¹⁵ However, measurements for all porphyrin-associated compounds are needed to

confirm the selectivity of this scaffold toward OATP1B1. The remaining three scaffold clusters represent gedunin- and khivorin-associated scaffolds (five associated compounds), *N*-phenylpyrimidin-4-amine (three associated compounds), and the valsartan-like scaffold (three compounds).

Enriched Scaffolds for Dual OATP1B1/OATP1B3 Inhibitors. In contrast to OATP1B1, inhibitors for OATP1B3 and OATP2B1 do not constitute enriched scaffolds that are specific for these transporters, since the number of respective enumerated compounds does not exceed two in these cases (data not shown).

Interestingly, the group of compounds and scaffolds with the highest occupied clusters belong to the class of compounds showing a pronounced activity against both OATP1B1 and OATP1B3 (dual inhibitors) (15 scaffolds and 11 scaffold clusters; depicted in Figure 3). This can be rationalized by the high sequence similarity between these two targets (~80%). The largest scaffold cluster with this activity annotation (14 compounds) is derived from cyclosporine and other associated macrocyclic compounds. There are two more clusters possessing macrocyclic scaffolds (four associated compounds each). Macrocyclic compounds in many cases show peptidomimetic properties and will be interesting candidates for future structure-based *in silico* studies, since it is likely that they accommodate different binding pockets than the smaller molecules.

Enriched Scaffolds for Pan Inhibitors of Hepatic OATPs. As a result of the scaffold frequency analysis undertaken for hepatic OATP inhibitors, no enriched scaffolds for pan

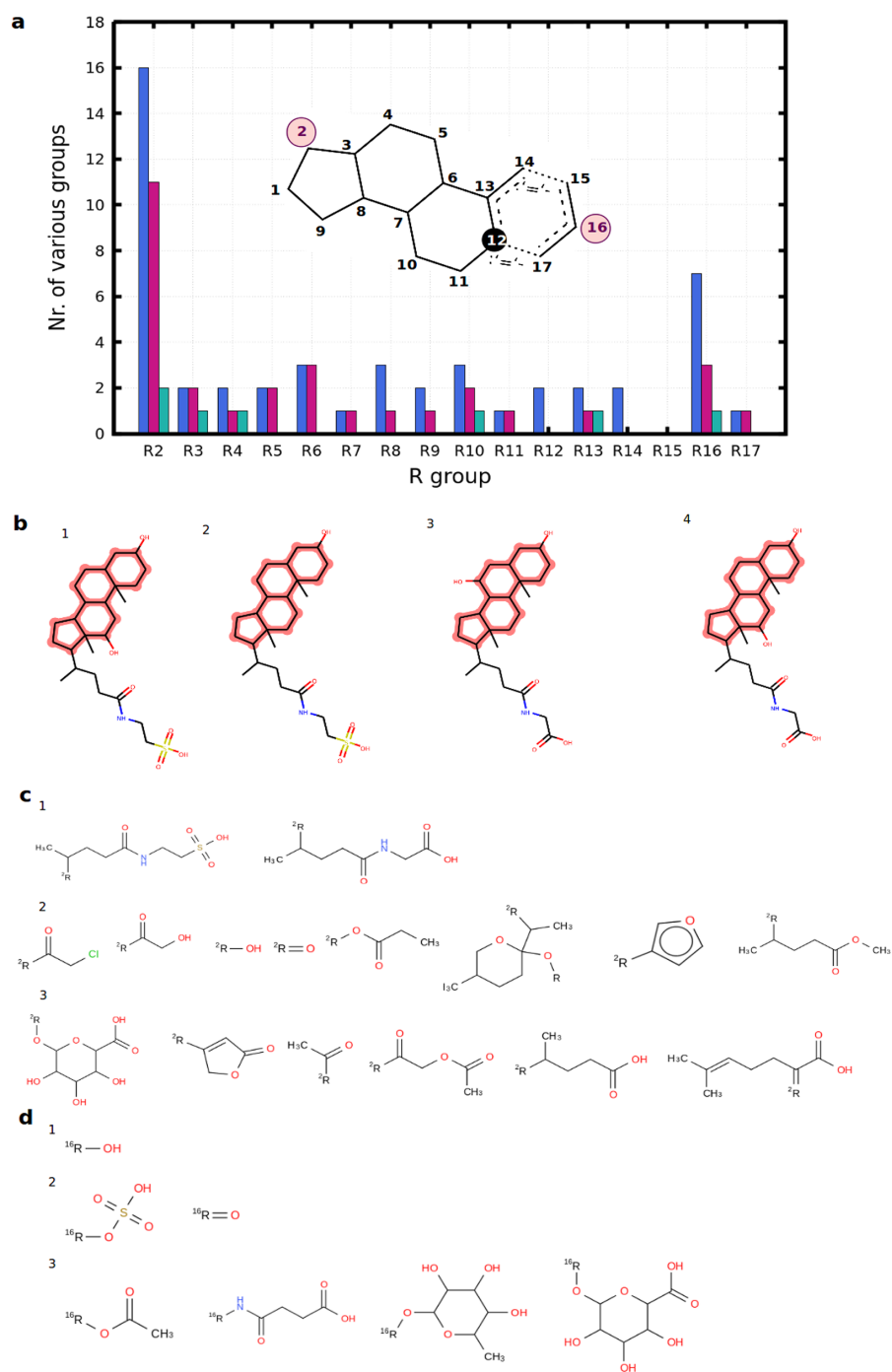


Figure 4. R-group decomposition of steroidal inhibitors. (a) Stacked bar plot showing the distribution of the number of various functional groups at certain R-group positions (blue bar plots, OATP1B1 inhibitors; purple bar plots, OATP1B3 inhibitors; green bar plots, OATP2B1 inhibitors). The maximum common substructure of all of the steroidal inhibitors is shown to highlight the R-group positions. (b) Steroidal ligands with proven pan-inhibitory effect: (1) taurodeoxycholic acid; (2) lithocholytaurine; (3) glyoursodeoxycholic acid; (4) glycodeoxycholic acid. (c) Functional groups identified at position 2 for (1) pan inhibitors, (2) dual OATP1B inhibitors, and (3) OATP1B1 inhibitors. (d) Functional groups identified at position 16 for (1) pan inhibitors, (2) dual OATP1B inhibitors, and (3) OATP1B1 inhibitors.

Table 5. Results on Level 1 (All Inhibitors + General Noninhibitors) and Level 2 (OATP1B1, OATP1B3, and OATP2B1 Inhibition Models) in Stratified Bagging for All Calculated Statistical Metrics: Sensitivity, Specificity, Balanced Accuracy, and MCC (The Performance Is Given for Both 10-Fold Cross-Validation and on the External Test Set)

	validation	sensitivity	specificity	balanced accuracy	MCC
level 1	training set	0.760	0.790	0.775	0.455
level 1	test set	0.796	0.769	0.783	0.477
level 2—OATP1B1	training set	0.703	0.799	0.751	0.462
level 2—OATP1B1	test set	0.730	0.809	0.769	0.497
level 2—OATP1B3	training set	0.748	0.834	0.791	0.486
level 2—OATP1B3	test set	0.746	0.829	0.787	0.476
level 2—OATP2B1	training set	0.698	0.771	0.734	0.434
level 2—OATP2B1	test set	0.632	0.840	0.736	0.464

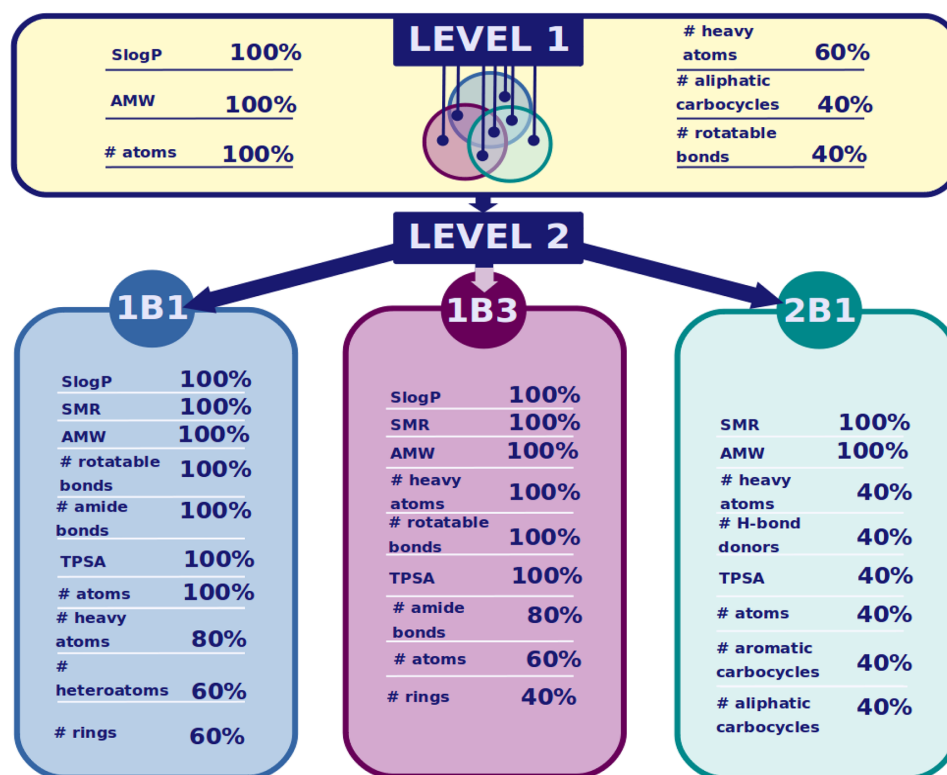


Figure 5. List of relevant features extracted from four different binary classification models with percentage of descriptor importance: level 1 model (any inhibitor vs general noninhibitors); level 2 models (separate models for OATP1B1 inhibition, OATP1B3 inhibition, and OATP2B1 inhibition).

inhibitors were detected as significantly enriched at $p < 0.05$. However, when the analysis was repeated at a bit weaker significance level ($p < 0.1$), we found the cholate-like steroidal scaffold to be enriched for all three hepatic OATPs (13 compounds in the sparse data set, four compounds in the dense data set; Figure S9). This is not surprising since the steroidal scaffold also occurs in natural substrates (e.g., cholate and taurocholate) and was already found to be enriched in the OATP1B1 inhibitor set. We applied an R-group decomposition procedure and analyzed the frequency of various R groups at certain positions in a targetwise manner. Positions 2 and 16 show the largest variety in terms of the numbers of functional groups. For substitutions at position 2, hydrophilic flexible side chains (e.g., *N*-sulfethylpropionamide-4-yl) occur

in ligands for all three hepatic OATPs, while, e.g., dihydrofuran or tetrahydropyran groups were detected only among OATP1B1 inhibitors at position 2 (Figure 4). At position 16, substitutions in general appear to be of hydrophilic nature, with tetrahydropyran rings with hydroxyl groups attached to the ring occurring only among OATP1B1 ligands (Figure 4). Looking at compounds with a proven pan-inhibitory effect for hepatic OATPs (four compounds from the dense data set; Figure 4b), we can see that the trends that we found among the sparse data set are verified for pan-inhibitory activity. In order to be able to make real selectivity claims here, more data with measurements on all three transporters will need to be investigated in the future.

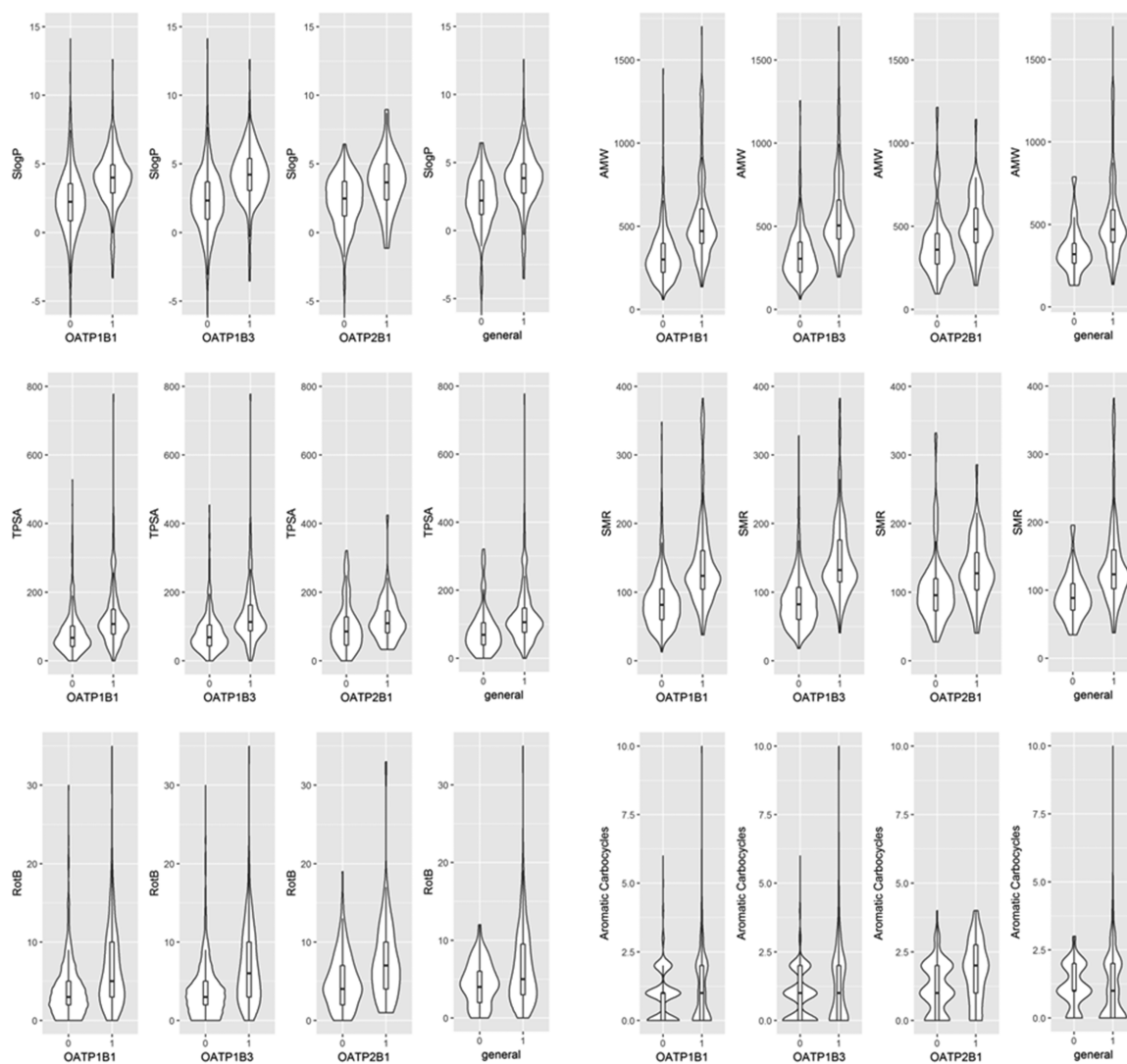


Figure 6. Violin and box plots showing the distributions of different molecular descriptors, namely, lipophilicity (SlogP), average molecular weight (AMW), topological polar surface area (TPSA), molecular refractivity (SMR), the number of rotatable bonds (RotB), and the number of aromatic carbocycles (Aromatic Carbocycles), for inhibitors vs noninhibitors within four different data sets. Labeling on abscissae: 0, inactive; 1, active.

OATP Substrates. An analogous analysis of scaffold frequency was also performed for OATP substrates. Because of the considerably lower number of known substrates for hepatic OATPs compared with inhibitors (see Table 3), this analysis could not retrieve any statistically significantly enriched scaffolds. It will be interesting to repeat this analysis when more data become available for hepatic OATP substrates.

In terms of side chains of steroid-associated substrates, we observed consistent trends, as positions 2 and 16 also show the largest variety of different side chains (data not shown).

Important Molecular Features for Inhibitory Activity. After the investigation of molecular determinants for ligand profiles at the scaffold level, it appeared interesting to look at a more abstract representation of structural features: molecular features/descriptors. Such representations might capture

commonalities among ligand sets of different hepatic OATPs that would not at first sight appear obvious at the level of scaffolds. The implemented strategy for retrieving important molecular features for the different data sets included the generation of binary classification models for hepatic OATP inhibitors. In more detail, we followed a sequential binary classification approach in which the first level comprised a machine learning model for general noninhibitors (compounds with annotations as “noninhibitors” for all three transporters) versus all inhibitors (OATP1B1 or/and OATP1B3 or/and OATP2B1 inhibitors). At the second level, three models for OATP inhibition (separately for OATP1B1, OATP1B3, and OATP2B1) were generated. It has to be pointed out that the major aim of this modeling approach was the extraction of relevant molecular descriptors and their careful analysis with respect to the transporters and already existing knowledge in

that domain. The use of these models for screening purposes and the subsequent identification of novel compounds/scaffolds (potentially active on hepatic OATPs) is not the focus of this investigation but will be conducted in follow-up studies.

A similar approach was used by Karlgren et al.¹⁹ in order to describe hepatic OATP inhibitors in terms of chemical features. One of the motivations to repeat this analysis was our curiosity to check whether our models built on basis of the chemically enhanced data sets would still prioritize the same chemical features or if we could retrieve other or additional features that likely better describe the data added since then.

We performed attribute selection ("CfsSubsetEval"⁴⁵) as implemented in the "BestFirst" search method in Weka²⁷ before model building. For each inhibitor data set, significant molecular features that would aid in distinguishing between inhibitors and noninhibitors could thus be retrieved. On basis of these "relevant" features, classification models were built assuring that highly correlated features were eliminated in order to get rid of redundant information. To account for difficulties due to imbalanced data sets (imbalance ratios between 1:2.5 and 1:4.5 for the different models), which usually affect model accuracies, two different meta-classifiers were used on top of "random tree" as the base classifier: a cost-sensitive classifier³¹ and stratified bagging.^{32,33} In a recent study by Jain et al.,³⁴ these two meta-classifiers were found to be the best-performing ones when dealing with imbalanced data sets. Assessing the performances of the final models, stratified bagging outperformed the cost-sensitive classifier. The balanced accuracies of the final models were in the range of 0.73 to 0.79, and the MCC values were between 0.43 and 0.5 (Table 5; model accuracies of all models built are given in Tables S13–S16).

Figure 5 shows the list of important features for each level and category of our sequential modeling approach. Since some of the descriptors were correlated, the final models were constructed with only a selection of those features (available in Tables S13–S16). Upon inspection of the relevant features given in Figure 5 and comparison of them across level 1 and to the models from level 2, it becomes clear that the general inhibitor model (level 1) broadly reflects the important features from the three individual models at level 2. This is not unexpected but shows that our methodology can capture differences and commonalities in the data sets.

For all four models, average molecular weight (AMW) (100% descriptor importance), the number of atoms (100–40%), and the number of heavy atoms (100–40%) are among the most important features for separating hepatic OATP inhibitors from noninhibitors (Figure 5). Since these three features are highly correlated, for building the final models only AMW was considered.

Lipophilicity (SlogP) was found to be an important descriptor (100% descriptor importance) for all of the models except the OATP2B1 model (Figure 5). It was therefore not taken into account for building the OATP2B1 model. For topological polar surface area (TPSA), we observe that it plays a role for the individual models but not for the general level 1 model. In addition, it seems to be less important in the case of OATP2B1 (40% descriptor importance; Figure 5). Thus, TPSA was not considered for building the final level 1 and OATP2B1 models.

Upon examination of the distribution of those features within the individual data sets (Figure 6 and Table S17) it

becomes obvious that in general hepatic OATP inhibitors do possess a higher lipophilicity, molecular weight, and polarity than noninhibitors. These findings are in accordance with the findings of Karlgren et al.,¹⁹ but in addition, we were able to prioritize a few other important features, one of which is the molecular refractivity or polarizability (SMR), which reflects the charge distribution on a molecules' surface. Since in the case of OATPs an inwardly directed pH gradient likely drives the transport,⁴⁶ a generally higher polarizability in the case of inhibitors versus noninhibitors together with a higher polarity seems very plausible (Figure 6). Interestingly, SMR appears with 100% descriptor importance for all of the individual level 2 models but does not contribute to the general level 1 model.

Other important parameters that were not discussed before by Karlgren et al.¹⁹ include the influence of flexibility (expressed by the number of rotatable bonds) and counts of different ring systems (especially aromatic rings). The number of rotatable bonds has previously been described as a discriminating factor for OATP1B1 inhibitors versus noninhibitors by van de Steeg et al.¹⁶ Our analysis suggests an important role of this feature for all hepatic OATP inhibitors (Figure 6 and Table S17). The number of rings was previously described as a discriminative molecular property by van de Steeg et al.¹⁶ for OATP1B1 inhibitors. De Bruyn et al.¹⁵ correlated a number of rings < 4 with OATP1B inactivity, which could be confirmed by our analysis and was also observed here for OATP2B1 (see Table S17). We found the number of rings to be discriminative for OATP1B1 and OATP1B3 inhibitors versus the respective noninhibitors (60–40% descriptor importance). However, for OATP2B1 inhibitors, more specific descriptors—namely, the numbers of aliphatic and aromatic carbocycles—were among the list of selected features. Since aromaticity can be linked to molecular complexity or 3D-ness, we were interested in how the feature "number of aromatic carbocycles" was distributed among the four inhibitor data sets. From Figure 6 and Table S17 it becomes obvious that only for OATP2B1 inhibitor data there is a significant difference in the distribution of this feature for inhibitors versus noninhibitors (for OATP1B1/OATP1B3, $p > 0.05$ in the Wilcoxon test; for OATP2B1, $p = 0.0004$).

Although the feature "FractionCSP3" (Fsp3), i.e., the fraction of sp³-hybridized carbons, was not among the prioritized ones for any model, one would expect to observe a similar trend in the distribution of this feature across the different transporters. Indeed, it was observed that for all of the data sets except the OATP2B1 data set, the inhibitors show a significantly higher Fsp3 than the respective noninhibitors. For OATP2B1, it can be observed that inhibitors on average possess lower Fsp3 values than inhibitors from the OATP1 subfamilies, which correlates with higher aromaticity and therefore higher planarity (Figure S10). Here again, a lack of data might be the reason for a tendency of planar molecules to inhibit OATP2B1. As is also visible from Figure 3, inhibitors of the OATP1B family do include large, flexible ring systems (e.g., cyclosporine, antamanide, microcystin, caspofungin), which were mostly not tested against OATP2B1.

Finally, the number of amide bonds was highlighted in cases of OATP1B inhibition models but not for the OATP2B1 and the general inhibition model. This can again be explained by the availability of large ring systems containing up to 11 amide bonds (e.g., cyclosporin) in the OATP1B data sets preferentially.

SUMMARY, CONCLUSIONS, AND OUTLOOK

The main aim of this study was to investigate potential structural determinants responsible for ligand activity or selectivity among hepatic OATPs on the basis of data available from the open domain. In this first study, we focused merely on ligand information as a rich source of chemical structures and bioactivities (pharmacological data).

Emphasis was put on data integration and data curation during the course of this study, as well as on semiautomatic processing of the data. All of the workflows have been made openly available to the scientific community so that they can be reused for other case studies. In addition, since hepatic OATPs are transporters of emerging interest for the research field of hepatotoxicity⁴⁷ and also in relation to cancer⁴⁸ and drug resistance,^{49,50} the current knowledge in this domain is expected to constantly increase in the near future. Therefore, our data integration, curation, and substructure analysis workflows will especially prove useful when a substantial amount of new data become available since in that case the whole analyses can be repeated and refined efficiently and swiftly.

As a side effect of this study, we collected six high-quality curated data sets, for substrates and inhibitors of OATP1B1, OATP1B3, and OATP2B1. Although data sparseness does not always allow delivery of a full ligand profile for all three hepatic OATPs, this analysis exemplifies that nonetheless commonalities and differences among related transporters can be determined by using the methods of data mining, cheminformatics, and ligand-based modeling.

These data sets as well as the information gained on enriched scaffolds and ligand properties of individual and general hepatic OATP inhibitors will serve as a basis for future investigations on ligand interactions and selectivity of hepatic OATPs. Especially the scaffold analyses delivered interesting scaffold series that will be exploited further in terms of their selectivity profiles with the help of structure-based in silico studies exploring individual ligand–protein binding events at the molecular level.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.8b00466.

Data sets in (a) ChEMBL and (b) Metabase annotated with bioactivity end point “inhibition”; lists of removed substrates and inhibitors with conflicting annotations; percentages of conflicting compound activities based on comparison of the data from ChEMBL and Metabase; dense data sets for hepatic OATP substrates and inhibitors; 10 detected compounds with the hexahydronaphthalene-associated scaffold with pharmacological profiles included; results from level 1 models (all inhibitors + general noninhibitors) for all calculated statistical metrics; results from OATP1B1, OATP1B3, and OATP2B1 inhibition models (level 2) for all calculated statistical metrics; summary statistics for molecular descriptors calculated for inhibitors of OATP1B1, OATP1B3, and OATP2B1; schematic workflow for integrative data mining and curation; box-and-whisker plots showing the distribution of molecular properties for compounds measured against human OATP1B1, OATP1B3, and OATP2B1 originating from

five different data sources (ChEMBL, Metabase, DrugBank, IUPHAR, TransPortal); box plot with number of bioactivities/annotations per unique compound; histograms showing the distributions of median bioactivities for OATP1B1, OATP1B3, and OATP2B1; Murcko scaffolds for OATP1B3 and OATP2B1 inhibitors retrieved from databases other than ChEMBL; enriched scaffolds ($p < 0.05$) for hepatic OATP inhibitors considering the dense data set (with complete pharmacological profile); enriched scaffolds ($p < 0.05$) for hepatic OATP inhibitors, excluding percentage inhibition data; enriched scaffolds ($p < 0.1$) for hepatic OATP inhibitors; violin plots showing the distribution feature “FractionCSP3” (Fsp3) for inhibitors versus noninhibitors within four different data sets (PDF)

Supplementary data files with sparse substrate/non-substrate and inhibitor/noninhibitor data sets in CSV format (ZIP)

AUTHOR INFORMATION

Corresponding Author

*E-mail: barbara.zdrzil@univie.ac.at; phone: +43-1-4277-55113.

ORCID

Barbara Zdrzil: 0000-0001-9395-1515

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We are grateful for technical support by Jana Gurinova when fetching the data from different sources. Gratitude is further expressed to Jennifer Hemmerich for providing us with the Atkinson standardization protocol as a KNIME node. This work received funding from the Austrian Science Fund (FWF) (Grant P 29712).

ABBREVIATIONS

OATP, organic anion transporting polypeptide; SLC, solute carrier; KNIME, Konstanz Information Miner; WEKA, Waikato Environment for Knowledge Analysis; MCC, Matthews correlation coefficient; MW, molecular weight; SMR, molecular refractivity; AMW, average molecular weight; TPSA, topological polar surface area; RotB, number of rotatable bonds; Fsp3, fraction of sp^3 -hybridized carbons

REFERENCES

- (1) Lin, L.; Yee, S. W.; Kim, R. B.; Giacomini, K. M. SLC Transporters as Therapeutic Targets: Emerging Opportunities. *Nat. Rev. Drug Discovery* **2015**, *14* (8), 543–560.
- (2) Kullak-ublick, G. A.; Stieger, B.; Meier, P. J. Enterohepatic Bile Salt Transporters in Normal Physiology and Liver Disease. *Gastroenterology* **2004**, *126* (1), 322–342.
- (3) Keppler, D. The Roles of MRP2, MRP3, OATP1B1, and OATP1B3 in Conjugated Hyperbilirubinemia. *Drug Metab. Dispos.* **2014**, *42* (4), 561–565.
- (4) Seithel, A.; Eberl, S.; Singer, K.; Auge, D.; Heinkele, G.; Wolf, N. B.; Dörje, F.; Fromm, M. F.; König, J. The Influence of Macrolide Antibiotics on the Uptake of Organic Anions and Drugs Mediated by OATP1B1 and OATP1B3. *Drug Metab. Dispos.* **2007**, *35* (5), 779–786.

- (5) Kunze, A.; Huwyler, J.; Camenisch, G.; Gutmann, H. Interaction of the Antiviral Drug Telaprevir with Renal and Hepatic Drug Transporters. *Biochem. Pharmacol.* **2012**, *84* (8), 1096–1102.
- (6) Obaidat, A.; Roth, M.; Hagenbuch, B. The Expression and Function of Organic Anion Transporting Polypeptides in Normal Tissues and in Cancer. *Annu. Rev. Pharmacol. Toxicol.* **2012**, *52* (1), 135–151.
- (7) Sandhu, P.; Lee, W.; Xu, X.; Leake, B. F.; Yamazaki, M.; Stone, J. A.; Lin, J. H.; Pearson, P. G.; Kim, R. B. Hepatic Uptake of the Novel Antifungal Agent Caspofungin. *Drug Metab. Dispos.* **2005**, *33* (5), 676–682.
- (8) Kim, R. B. 3-Hydroxy-3-methylglutaryl-Coenzyme A Reductase Inhibitors (Statins) and Genetic Variability (Single Nucleotide Polymorphisms) in a Hepatic Drug Uptake Transporter: What's It All About? *Clin. Pharmacol. Ther.* **2004**, *75* (5), 381–385.
- (9) Cvetkovic, M.; Leake, B.; Fromm, M. F.; Wilkinson, G. R.; Kim, R. B. OATP and P-Glycoprotein Transporters Mediate the Cellular Uptake and Excretion of Fexofenadine. *Drug Metab. Dispos.* **1999**, *27* (8), 866–871.
- (10) Bachmakov, I.; Glaeser, H.; Fromm, M. F.; König, J. Interaction of Oral Antidiabetic Drugs With Hepatic Uptake Transporters: Focus on Organic Anion Transporting Polypeptides and Organic Cation Transporter 1. *Diabetes* **2008**, *57* (6), 1463–1469.
- (11) Mikkaichi, T.; Suzuki, T.; Tanemoto, M.; Ito, S.; Abe, T. The Organic Anion Transporter (OATP) Family. *Drug Metab. Pharmacokinet.* **2004**, *19* (3), 171–179.
- (12) Kindla, J.; Müller, F.; Mieth, M.; Fromm, M. F.; König, J. Influence of Non-Steroidal Anti-Inflammatory Drugs on Organic Anion Transporting Polypeptide (OATP) 1B1 and OATP1B3-Mediated Drug Transport. *Drug Metab. Dispos.* **2011**, *39* (6), 1047–1053.
- (13) Shitara, Y.; Maeda, K.; Ikejiri, K.; Yoshida, K.; Horie, T.; Sugiyama, Y. Clinical Significance of Organic Anion Transporting Polypeptides (OATPs) in Drug Disposition: Their Roles in Hepatic Clearance and Intestinal Absorption. *Biopharm. Drug Dispos.* **2013**, *34* (1), 45–78.
- (14) Hagenbuch, B.; Stieger, B. The SLCO (Former SLC21) Superfamily of Transporters. *Mol. Aspects Med.* **2013**, *34* (2–3), 396–412.
- (15) de Bruyn, T.; van Westen, G. J. P.; IJzerman, A. P.; Stieger, B.; de Witte, P.; Augustijns, P. F.; Annaert, P. P. Structure-Based Identification of OATP1B1/3 Inhibitors. *Mol. Pharmacol.* **2013**, *83* (6), 1257–1267.
- (16) van de Steeg, E.; Venhorst, J.; Jansen, H. T.; Nooijen, I. H. G.; DeGroot, J.; Wortelboer, H. M.; Vlamings, M. L. H. Generation of Bayesian Prediction Models for OATP-Mediated Drug–Drug Interactions Based on Inhibition Screen of OATP1B1, OATP1B1*15 and OATP1B3. *Eur. J. Pharm. Sci.* **2015**, *70*, 29–36.
- (17) Kotsampasakou, E.; Brenner, S.; Jäger, W.; Ecker, G. F. Identification of Novel Inhibitors of Organic Anion Transporting Polypeptides 1B1 and 1B3 (OATP1B1 and OATP1B3) Using a Consensus Vote of Six Classification Models. *Mol. Pharmaceutics* **2015**, *12* (12), 4395–4404.
- (18) Khuri, N.; Zur, A. A.; Wittwer, M. B.; Lin, L.; Yee, S. W.; Sali, A.; Giacomini, K. M. Computational Discovery and Experimental Validation of Inhibitors of the Human Intestinal Transporter OATP2B1. *J. Chem. Inf. Model.* **2017**, *57* (6), 1402–1413.
- (19) Karlgren, M.; Vildhede, A.; Norinder, U.; Wisniewski, J. R.; Kimoto, E.; Lai, Y.; Haglund, U.; Artursson, P. Classification of Inhibitors of Hepatic Organic Anion Transporting Polypeptides (OATPs): Influence of Protein Expression on Drug–Drug Interactions. *J. Med. Chem.* **2012**, *55* (10), 4740–4763.
- (20) Berthold, M. R.; Cebon, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinel, T.; Ohl, P.; Thiel, K.; Wiswedel, B. KNIME—the Konstanz Information Miner: Version 2.0 and Beyond. *SIGKDD Explor. Newsl.* **2009**, *11* (1), 26–31.
- (21) Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P. The ChEMBL Bioactivity Database: An Update. *Nucleic Acids Res.* **2014**, *42* (D1), D1083–D1090.
- (22) Morrissey, K. M.; Wen, C. C.; Johns, S. J.; Zhang, L.; Huang, S.-M.; Giacomini, K. M. The UCSF–FDA TransPortal: A Public Drug Transporter Database. *Clin. Pharmacol. Ther.* **2012**, *92* (5), S45–S46.
- (23) Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; Assempour, N.; Iynkkaran, I.; Liu, Y.; Maciejewski, A.; Gale, N.; Wilson, A.; Chin, L.; Cummings, R.; Le, D.; Pon, A.; Knox, C.; Wilson, M. DrugBank 5.0: A Major Update to the DrugBank Database for 2018. *Nucleic Acids Res.* **2018**, *46* (D1), D1074–D1082.
- (24) Mak, L.; Marcus, D.; Howlett, A.; Yarova, G.; Duchateau, G.; Klaffke, W.; Bender, A.; Glen, R. C. Metrabase: A Cheminformatics and Bioinformatics Database for Small Molecule Transporter Data Analysis and (Q)SAR Modeling. *J. Cheminf.* **2015**, *7*, 31.
- (25) Pawson, A. J.; Sharman, J. L.; Benson, H. E.; Faccenda, E.; Alexander, S. P. H.; Buneman, O. P.; Davenport, A. P.; McGrath, J. C.; Peters, J. A.; Southan, C.; Spedding, M.; Yu, W.; Harmar, A. J. The IUPHAR/BPS Guide to PHARMACOLOGY: An Expert-Driven Knowledgebase of Drug Targets and Their Ligands. *Nucleic Acids Res.* **2014**, *42* (D1), D1098–D1106.
- (26) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39* (15), 2887–2893.
- (27) Frank, E.; Hall, M.; Trigg, L.; Holmes, G.; Witten, I. H. Data Mining in Bioinformatics Using Weka. *Bioinformatics* **2004**, *20* (15), 2479–2481.
- (28) Beisken, S.; Meinel, T.; Wiswedel, B.; de Figueiredo, L. F.; Berthold, M.; Steinbeck, C. KNIME-CDK: Workflow-Driven Cheminformatics. *BMC Bioinf.* **2013**, *14*, 257.
- (29) McColm, G. L. An Introduction to Random Trees. *Res. Lang. Comput.* **2003**, *1* (3–4), 203–227.
- (30) Le Gall, J.-F. Random Trees and Applications. *Probab. Surveys* **2005**, *2*, 245–311.
- (31) López, V.; Fernández, A.; Moreno-Torres, J. G.; Herrera, F. Analysis of Preprocessing vs. Cost-Sensitive Learning for Imbalanced Classification. Open Problems on Intrinsic Data Characteristics. *Expert Syst. Appl.* **2012**, *39* (7), 6585–6608.
- (32) He, H.; Garcia, E. A. Learning from Imbalanced Data. *IEEE Trans. Knowledge Data Eng.* **2009**, *21* (9), 1263–1284.
- (33) Tetko, I. V.; Novotarskyi, S.; Sushko, I.; Ivanov, V.; Petrenko, A. E.; Dieden, R.; Lebon, F.; Mathieu, B. Development of Dimethyl Sulfoxide Solubility Models Using 163 000 Molecules: Using a Domain Applicability Metric to Select More Reliable Predictions. *J. Chem. Inf. Model.* **2013**, *53* (8), 1990–2000.
- (34) Jain, S.; Kotsampasakou, E.; Ecker, G. F. Comparing the Performance of Meta-Classifiers—a Case Study on Selected Imbalanced Data Sets Relevant for Prediction of Liver Toxicity. *J. Comput.-Aided Mol. Des.* **2018**, *32* (5), S83–S90.
- (35) Leonhardt, M.; Keiser, M.; Oswald, S.; Kühn, J.; Jia, J.; Grube, M.; Kroemer, H. K.; Siegmund, W.; Weitschies, W. Hepatic Uptake of the Magnetic Resonance Imaging Contrast Agent Gd-EOB-DTPA: Role of Human Organic Anion Transporters. *Drug Metab. Dispos.* **2010**, *38* (7), 1024–1028.
- (36) Peters, J.; Eggers, K.; Oswald, S.; Block, W.; Lütjohann, D.; Lämmer, M.; Venner, M.; Siegmund, W. Clarithromycin Is Absorbed by an Intestinal Uptake Mechanism That Is Sensitive to Major Inhibition by Rifampicin: Results of a Short-Term Drug Interaction Study in Foals. *Drug Metab. Dispos.* **2012**, *40* (3), S22–S28.
- (37) Kramer, C.; Kalliokoski, T.; Gedeck, P.; Vulpetti, A. The Experimental Uncertainty of Heterogeneous Public K_i Data. *J. Med. Chem.* **2012**, *55* (11), S165–S173.
- (38) Kalliokoski, T.; Kramer, C.; Vulpetti, A.; Gedeck, P. Comparability of Mixed IC50 Data—A Statistical Analysis. *PLoS One* **2013**, *8* (4), e61007.
- (39) Montanari, F.; Ecker, G. F. BCRP Inhibition: From Data Collection to Ligand-Based Modeling. *Mol. Inf.* **2014**, *33* (5), 322–331.

- (40) Lowe, D. M.; Corbett, P. T.; Murray-Rust, P.; Glen, R. C. Chemical Name to Structure: OPSIN, an Open Source Solution. *J. Chem. Inf. Model.* **2011**, *51* (3), 739–753.
- (41) Nozawa, T.; Tamai, I.; Sai, Y.; Nezu, J.-I.; Tsuji, A. Contribution of Organic Anion Transporting Polypeptide OATP-C to Hepatic Elimination of the Opioid Pentapeptide Analogue [d-Ala²,d-Leu⁵]-Enkephalin. *J. Pharm. Pharmacol.* **2003**, *55* (7), 1013–1020.
- (42) Jasial, S.; Hu, Y.; Bajorath, J. Assessing the Growth of Bioactive Compounds and Scaffolds over Time: Implications for Lead Discovery and Scaffold Hopping. *J. Chem. Inf. Model.* **2016**, *56* (2), 300–307.
- (43) Zdrazil, B.; Hellsberg, E.; Viereck, M.; Ecker, G. F. From Linked Open Data to Molecular Interaction: Studying Selectivity Trends for Ligands of the Human Serotonin and Dopamine Transporter. *MedChemComm* **2016**, *7* (9), 1819–1831.
- (44) Li, X.; Guo, Z.; Wang, Y.; Chen, X.; Liu, J.; Zhong, D. Potential Role of Organic Anion Transporting Polypeptide 1B1 (OATP1B1) in the Selective Hepatic Uptake of Hematoporphyrin Monomethyl Ether Isomers. *Acta Pharmacol. Sin.* **2015**, *36* (2), 268–280.
- (45) Hall, M. A. Correlation-Based Feature Selection for Machine Learning. Ph.D. Thesis, University of Waikato, Hamilton, New Zealand, 1999.
- (46) Leuthold, S.; Hagenbuch, B.; Mohebbi, N.; Wagner, C. A.; Meier, P. J.; Stieger, B. Mechanisms of PH-Gradient Driven Transport Mediated by Organic Anion Polypeptide Transporters. *Am. J. Physiol. Cell Physiol.* **2009**, *296* (3), C570–C582.
- (47) Kotsampasakou, E.; Escher, S. E.; Ecker, G. F. Linking Organic Anion Transporting Polypeptide 1B1 and 1B3 (OATP1B1 and OATP1B3) Interaction Profiles to Hepatotoxicity - The Hyperbilirubinemia Use Case. *Eur. J. Pharm. Sci.* **2017**, *100*, 9–16.
- (48) Thakkar, N.; Lockhart, A. C.; Lee, W. Role of Organic Anion-Transporting Polypeptides (OATPs) in Cancer Therapy. *AAPS J.* **2015**, *17* (3), 535–545.
- (49) Lancaster, C. S.; Sprowl, J. A.; Walker, A. L.; Hu, S.; Gibson, A. A.; Sparreboom, A. Modulation of OATP1B-Type Transporter Function Alters Cellular Uptake and Disposition of Platinum Chemotherapeutics. *Mol. Cancer Ther.* **2013**, *12* (8), 1537–1544.
- (50) Brenner, S.; Riha, J.; Giessrigl, B.; Thalhammer, T.; Grusch, M.; Krupitza, G.; Stieger, B.; Jäger, W. The Effect of Organic Anion-Transporting Polypeptides 1B1, 1B3 and 2B1 on the Antitumor Activity of Flavopiridol in Breast Cancer Cells. *Int. J. Oncol.* **2015**, *46* (1), 324–332.

3.2 Structural Dissection of 13-epiestrones Based on the Interaction with Human Organic Anion-transporting Polypeptide, OATP2B1

LACZKÓ-RIGÓ, Réka; JÓJÁRT, Rebeka; MERNYÁK, Erszébet; BÁKOS, Éva; TUERKOVA, Alzbeta; ZDRAZIL, Barbara; ÖZVEGY-LACZKA, Csilla. *The Journal of Steroid Biochemistry and Molecular Biology*, **2020**, 105652.

* Corresponding author: laczka.csilla@ttk.mta.hu

R. Laczkó-Rigó designed experimental methodology, performed EC₅₀ measurements, data visualization, and wrote original draft. R. Jójárt and E. Mernyák provided resources. A. Tuerkova, together with B. Zdrzil, designed SAR analysis. A. Tuerkova performed SAR analysis and wrote corresponding parts of the manuscript. C. Laczka, together with E. Bákos, conceived the research. The manuscript was written by the contribution of all authors.

The Supplementary Information can be found in Part V.

This article is reprinted with permission from Elsevier:

Laczkó-Rigó, R., Jójárt, R., Mernyák, E., Bakos, É., Tuerkova, A., Zdrzil, B., Özvegy-Laczka, C. (2020). Structural dissection of 13-epiestrones based on the interaction with human Organic anion-transporting polypeptide, OATP2B1. *The Journal of Steroid Biochemistry and Molecular Biology*, 105652.



Contents lists available at ScienceDirect

Journal of Steroid Biochemistry and Molecular Biology

journal homepage: www.elsevier.com/locate/jsbmb

Structural dissection of 13-epiestrones based on the interaction with human Organic anion-transporting polypeptide, OATP2B1

Réka Laczkó-Rigó^a, Rebeka Jójárt^b, Erzsébet Mernyák^b, Éva Bakos^a, Alzbeta Tuerkova^c, Barbara Zdrazil^c, Csilla Özvegy-Laczka^{a,*}

^a Membrane Protein Research Group, Institute of Enzymology, RCNS, H-1117, Budapest, Magyar tudósok krt. 2, Hungary

^b Department of Organic Chemistry, University of Szeged, Dóm tér 8, H-6720, Szeged, Hungary

^c Department of Pharmaceutical Chemistry, Division of Drug Design and Medicinal Chemistry, University of Vienna, Althanstraße 14, A-1090, Vienna, Austria

ARTICLE INFO

Keywords:

Organic anion-transporting polypeptide
13-epiestrones
Inhibitor
SAR

ABSTRACT

Human OATP2B1 encoded by the *SLCO2B1* gene is a multispecific transporter mediating the cellular uptake of large, organic molecules, including hormones, prostaglandins and bile acids. OATP2B1 is ubiquitously expressed in the human body, with highest expression levels in pharmacologically relevant barriers, like enterocytes, hepatocytes and endothelial cells of the blood-brain-barrier. In addition to its endogenous substrates, OATP2B1 also recognizes clinically applied drugs, such as statins, antivirals, antihistamines and chemotherapeutic agents and influences their pharmacokinetics. On the other hand, OATP2B1 is also overexpressed in various tumors. Considering that elevated hormone uptake by OATP2B1 results in increased cell proliferation of hormone dependent tumors (e.g. breast or prostate), inhibition of OATP2B1 can be a good strategy to inhibit the growth of these tumors.

13-epiestrones represent a potential novel strategy in the treatment of hormone dependent cancers by the suppression of local estrogen production due to the inhibition of the key enzyme of estrone metabolism, 17 β -hydroxysteroid-dehydrogenase type 1 (HSD17 β 1). Recently, we have demonstrated that various phosphonated 13-epiestrones are dual inhibitors also suppressing OATP2B1 function. In order to gain better insights into the molecular determinants of OATP2B1 13-epiestrone interaction we investigated the effect of C-2 and C-4 halogen or phenylalkynyl modified epiestrones on OATP2B1 transport function. Potent inhibitors (with EC₅₀ values in the low micromolar range) as well as non-inhibitors of OATP2B1 function were identified. Based on the structure-activity relationship (SAR) of the various 13-epiestrone derivatives we could define structural elements important for OATP2B1 inhibition. Our results may help to understand the drug/inhibitor interaction profile of OATP2B1, and also may be a useful strategy to block steroid hormone entry into tumors.

1. Introduction

Organic anion-transporting polypeptides (OATPs) encoded by the *SLCO* (solute carrier for organic anions) genes are membrane proteins that mediate the cellular uptake of large (> 300 Da) organic molecules in a Na⁺- and ATP-independent manner [1]. 11 human OATPs are known, that are highly variable in their tissue distribution and substrate recognition. Some OATPs are ubiquitously expressed in the human body (OATP4A1, OATP3A1), while the expression of others is restricted to a given organ, like OATP1B1 and OATP1B3 expression which is restricted to hepatocytes [2,3]. Also, based on the substrate interaction

profile, multispecific OATPs (1A2, 1B1, 1B3, and 2B1) recognizing a plethora of organic compounds (including clinically applied drugs), and OATPs with a more limited substrate recognition (e.g., the thyroid transporter OATP1C1) can be distinguished. Although not all of the 11 OATPs are properly characterized with regard to their expression and function, steroids (e.g. bile acids, estrone-3-sulfate (E1S) and dehydroepiandrosterone sulfate (DHEAS)) can be considered as general OATP substrates [4]. Consequently, the OATPs 1A2, 1B1, 2B1 and 4A1 are supposed to be key participants in the cellular uptake of the steroid hormone conjugates E1S and DHEAS [3,5]. Besides their role in the maintenance of steroid hormone homeostasis, multispecific OATPs,

Abbreviations: OATP, Organic anion-transporting polypeptide; E1S, estrone-3-sulfate; HSD17 β 1, 17 β -hydroxysteroid-dehydrogenase type 1; SAR, structure activity relationship; STS, steroid sulfatase

* Corresponding author.

E-mail address: laczka.csilla@ttk.mta.hu (C. Özvegy-Laczka).

<https://doi.org/10.1016/j.jsbmb.2020.105652>

Received 13 November 2019; Received in revised form 20 February 2020; Accepted 5 March 2020

Available online 06 March 2020

0960-0760/ © 2020 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1A2, 1B1, 1B3 and 2B1 are important determinants of pharmacokinetics [6]. In addition, OATPs are often up-regulated in tumors [7–9]. Hence, they are promising targets for anti-tumor therapy.

OATP2B1 is a ubiquitously expressed transporter, with highest protein levels in pharmacologically relevant barrier tissues, like the intestine, liver and blood-brain barrier [1]. Besides, it is also expressed in the placenta, mammary gland and in skeletal muscle cells [1]. OATP2B1 is a multispecific transporter that recognizes molecules with largely variable size and structure. The most relevant endogenous substances transported by OATP2B1 are taurocholate, leukotriene C₄, E1S, DHEAS and prostaglandin E₂ [10]. OATP2B1 also promotes cellular uptake of clinically applied drugs, like statins, antibiotics, anti-hypertensives, anti-inflammatory drugs and chemotherapeutics [10]. Considering its tissue distribution and substrate recognition pattern, OATP2B1 may be a key player in intestinal drug absorption and also drug transport across the blood–brain barrier [11,12]. On the other hand, OATP2B1 overexpression has been detected in tumors of the colon, bone, breast, prostate and also in gliomas [13–15]. Considering its transport of anti-cancer agents, OATP2B1 is one of the main candidates of tumor-targeted drug delivery. On the other hand, it has been shown that increased steroid hormone (E1S, DHEAS) uptake by OATP2B1 promotes growth of steroid-dependent tumors. Matsumoto and colleagues demonstrated that overexpression of OATP2B1 results in increased survival of breast cancer cells *in vitro* [14]. Moreover, *in vivo* data revealed that DHEAS uptake by OATP2B1 has crucial role in prostate cancer progression [16]. Also, the SLCO2B1 rs12422149 GG (Arg312Gln) genotype resulting in increased OATP2B1 function correlates with shorter time to progression in prostate cancer patients who received androgen deprivation therapy [17,18]. Therefore, inhibition of OATP2B1 function presents a possible strategy to suppress steroid hormone uptake and hence the proliferation of hormone dependent cancers.

13-epiestrones are stereoisomers of natural estrone, lacking hormonal activity [19,20]. Previous work has demonstrated that certain 13-epiestrones are potent inhibitors of the 17 β -hydroxysteroid-dehydrogenase type 1 (HSD17B1) and steroid sulfatase (STS) enzymes crucial in estrone metabolism [21]. These enzymes are responsible for local estrogen formation and generation of the transcriptionally active estradiol therefore promoting proliferation of hormone dependent cancers [22,23]. Hence their inhibition e.g. by 13-epiestrones can be a potential anti-tumor strategy.

Recently we found that phosphonated 13-epiestrones inhibit the function of OATP2B1 [24]. In the current work, in order to get a better insight into the molecular determinants involved in this inhibition we analyzed the interaction between OATP2B1 and a large set of 13-epiestrones containing modifications on C-3 and C-2 or C-4. In addition, we systematically investigated the influence of certain substituents on inhibitory activity by correlation analysis.

2. Materials and methods

2.1. Materials

Materials if not stated otherwise were purchased for Sigma Aldrich (Budapest, Hungary). 13-epiestrones investigated in this study were synthesized as described elsewhere [21,25].

2.2. Generation and maintenance of the cell lines

A431 (human epidermoid carcinoma) cells overexpressing human OATP2B1 or mock transfected controls used in the current study were generated earlier as described in [26]. Briefly, OATP2B1 expressing cells were generated by transposase mediated genomic insertion of the OATP2B1 cDNA (BC041095.1, HsCD00378878). As a negative control mock transfected (pSB-CMV) cells were used. After 2 weeks of puromycin (1 μ g/ml) selection, cells were sorted based on Live/Dead Green

uptake. After recovery, cell were grown in DMEM (Gibco, Thermo Fischer Scientific (Waltham, MA, US)) without puromycin supplemented with 10 % fetal calf serum, 2 mM L-glutamine, 100 U/ml penicillin, and 100 μ g/ml streptomycin at 37 °C with 5% CO₂ and 95 % humidity.

2.3. Western blot detection of OATP2B1 expression

OATP2B1 expression was confirmed by Western blot as described earlier [26]. Briefly, whole cell lysates of A431 cells were separated on 7.5 % SDS-PAGE gels and transferred onto a PVDF membrane. OATP2B1 was detected by using an anti-OATP2B1 antibody (a courtesy of Dr. Bruno Stieger, Department of Clinical Pharmacology and Toxicology, University Hospital, 8091 Zurich, Switzerland) [27]. As a secondary antibody HRP-conjugated anti-rabbit antibody (Jackson ImmunoResearch, Suffolk, UK) was used in a dilution of 20,000 \times . An anti- β -actin antibody (A1978, Sigma) and HRP-conjugated anti-mouse antibody (Jackson ImmunoResearch, Suffolk, UK, 20,000 \times dilution) were used to detect β -actin. Luminescence was detected using the Luminor Enhancer Solution kit by Thermo Fisher Scientific (Waltham, MA, US).

2.4. Fluorescent dye uptake determined by flow cytometry

The transport function of OATP2B1 was determined by flow cytometry. A431 cells (mock and OATP2B1 overexpressing) were collected after 0.1 % trypsin treatment. The cells were washed in Uptake buffer (125 mM NaCl, 4.8 mM KCl, 1.2 mM CaCl₂, 1.2 mM KH₂PO₄, 12 mM MgSO₄, 25 mM MES, and 5.6 mM glucose, with the pH adjusted to 5.5 using 1 M HEPES and 1 N NaOH). After washing 5 \times 10⁵ cells were incubated for 15 min at 37 °C with Zombie Violet (BioLegend®, San Diego, CA, US) (0.4 μ l ZV/5 \times 10⁵ cell) in a final volume of 100 μ l. The reaction was stopped by the addition of 1 ml ice-cold PBS (phosphate buffered saline) and the cells were kept on ice until the flow cytometry analysis. The fluorescence of 10,000 living cells was determined using Attune NxT Flow Cytometer (Invitrogen, Carlsbad, CA). Dead cells were excluded by propidium iodide (1 μ g/ml) labeling.

2.5. 96 well plate-based transport assay

Effect of 13-epiestrones on OATP2B1 function was determined by measuring Zombie Violet (BioLegend®, San Diego, CA, US) fluorescent dye uptake on microplates [26]. Briefly, A431 cells were seeded on 96 well-plates (8 \times 10⁴ cells in 200 μ l final volume/well) and cultured for 16–24 h at 37 °C, 5% CO₂ prior to the transport measurements. Next day after repeated washing with 200 μ l PBS, cells were pre-incubated with 50 μ l Uptake buffer (125 mM NaCl, 4.8 mM KCl, 1.2 mM CaCl₂, 1.2 mM KH₂PO₄, 12 mM MgSO₄, 25 mM MES, and 5.6 mM glucose, with the pH adjusted to 5.5 using 1 M HEPES and 1 N NaOH) containing the appropriate concentrations of the 13-epiestrones (0–100 μ M) for 5 min at 37 °C. Reaction was started by the addition of 770 \times diluted Zombie Violet dye in 50 μ l/well Uptake buffer followed by a 30 min incubation at 37 °C. The reaction was stopped by the addition of 200 μ l ice-cold PBS. After repeated washing, 200 μ l ice-cold PBS was added to each well and fluorescence in the wells was determined in an Enspire fluorescent plate reader (Perkin Elmer, Waltham, MA) at Ex/Em: 405/423 nm. Experiments were repeated at least three times.

2.6. Measurement of ³H-E1S uptake

A431 control and A431-OATP2B1 cells (10⁶ cells/sample) were incubated in the absence or presence of 2-bromo-13-epiestrone (final concentration 50 μ M) for 5 min at 37 °C in uptake buffer pH 5.5. Transport reaction was started by the addition of ³H-E1S (250 mCi/ml, final concentration 9.65 nM (Perkin Elmer, Waltham, MA)). After incubation for further 10 min at 37 °C, the reaction was stopped by the addition of 1 ml ice-cold PBS and the cells were centrifuged at 300 g.

The cell pellet was collected in 100 μ l PBS and pipetted into 1 ml Opti-Fluor (Perkin Elmer, Waltham, MA). Radioactivity was measured in a Wallac Liquid Scintillator Counter. Experiments were repeated three times.

2.7. Data calculation

Transport data were obtained by subtracting the fluorescence in mock transfected cells from that measured in OATP2B1 cells. Kinetic parameters of dye uptake and half inhibitory concentrations (EC_{50}) obtained from at least three independent experiments were determined by Hill fit using the GraphPad prism software (GraphPad, La Jolla, CA, USA).

2.8. Structure-activity relationship (SAR) analysis

Compounds were drawn using the structure editor integrated into the ChemSpider chemical structure database webservice (freely available at <https://www.chemspider.com/About.aspx>). Daylight SMILES structural format was generated for every unique compound of the dataset. KNIME Analytics Platform (version 3.4) [28] was used to create an automated workflow for Structure-Activity Relationship (SAR) analysis of 13-epiestrones. First, Daylight SMILES format for input compounds was converted into the canonical form ('RDKit Canon SMILES' node). Murcko scaffolds were generated and the maximum common substructure (MCS) was derived from the retrieved Murcko scaffolds ('RDKit MCS' node). MCS was used as a structural query for substructure mining in order to perform R-group decomposition ('RDKit R Group Decomposition' node).

Physicochemical descriptors (RDKit) for the substituents (R groups) at position C-2 and C-4 were calculated. Descriptor values were normalized ('Normalizer' node using Z-score normalization). The Pearson correlation coefficient was calculated to identify positive or negative correlations between pEC_{50} values and respective physicochemical descriptors at a given R-group position ('Linear Correlation' node).

3. Results

3.1. Effect of C-2 or C-4 halogenated 13-epiestrones on the transport activity of OATP2B1

Recently, we have reported that various phosphonated 13-epiestrones are potent inhibitors of OATP2B1 function [24]. In order to gain better insights into the molecular determinants of this inhibition, we investigated the inhibitory effect of various 2- or 4-halogenated 3-hydroxy- (3-OH) or 3-methoxy (3-OMe) 13-epiestrones (Fig. 1).

Interaction was measured in A431 mock transfected (control) and OATP2B1 overexpressing cell lines using the Zombie Violet (ZV) assay. The A431 cell line overexpressing OATP2B1 was generated earlier [26]. Prior to the interaction tests OATP2B1 overexpression was confirmed by Western blot (Fig. 2A). ZV is one of the newly identified fluorescent dyes applicable for testing drug interactions and function of OATP2B1, and a good alternative to the generally used radioactive functional assays [26]. Fig. 2B shows that OATP2B1 overexpression results in increased uptake of the viability dye, ZV.

When testing the inhibitory effect of the epiestrone derivatives, ZV uptake was measured in A431-OATP2B1 (and control) cells seeded in 96-well plates in the presence of increasing amounts of the tested compounds. As shown in Fig. 3, the compounds showed various degrees of transport inhibition. The first striking difference could be observed between the C-3 OH and OMe compounds. 13-epiestrone (C-3 OH, termed as **compound 1**) had practically no effect on ZV transport (EC_{50} around 50 μ M), while its methylether counterpart (3-OMe derivative, **compound 2**) resulted in a quite effective inhibition (EC_{50} 2.98 μ M) (see Fig. 3 and Table 1).

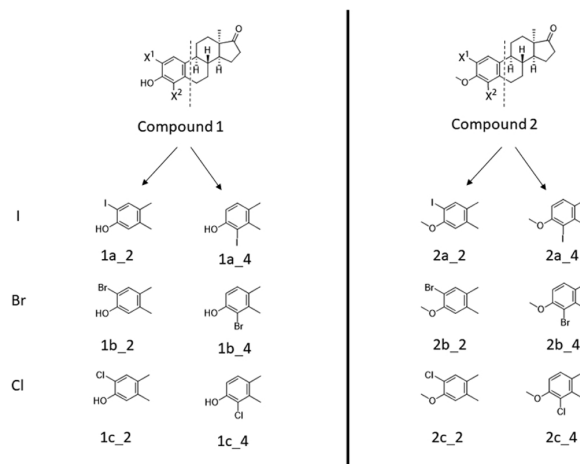


Fig. 1. Structure of the halogenated 13-epiestrones investigated in the current study. In the case of **compounds 1 and 2** $X^1 = H$ and $X^2 = H$.

Interestingly, introduction of a second modification, halogenation of either at C-2 or C-4 resulted in opposite changes in the inhibitory potential of 13-epiestrones. In the case of compound 1, halogenation at C-2 (compounds **1a_2**, **1b_2** and **1c_2**) resulted in a striking increase in inhibitory potential, with EC_{50} values between 0.5 and 2.1 μ M. However, the C-4 chlorinated derivative (compound **1c_4**) showed practically no effect on OATP2B1 transport function, while 4-iodo and 4-bromo 13-epiestrones (**1a_4** and **1b_4**) proved to be weak inhibitors. This reveals a strong dependence of the regioisomerism of the halogenated 3-OH compounds, the most potent inhibition detected with the 2-halogenated derivatives. In the case of 3-OMe derivatives, halogenation caused lower alteration in the inhibitory potential, although all of the compounds showed increased EC_{50} values compared to the initial compound. Iodinated derivatives (**2a_2** and **2a_4**) were more effective than the brominated or chlorinated compounds (**2b_2**, **2b_4** or **2c_2**, **2c_4**). Regioisomer specific inhibitory effect in the case of the 3-OMe variants could only be observed for the chlorinated compounds, with an approximately 5-fold increase in the EC_{50} of the C-4 vs. C-2 chlorinated epiestrones (**2c_2** and **2c_4**, Table 1).

3.2. Effect of C-2 or C-4 phenylalkynylated 13-epiestrones on the transport activity of OATP2B1

Next, we investigated the effect of the introduction of a phenylalkynyl group in position C-2 or C-4 on OATP2B1 function (Fig. 4).

In the case of 3-OH 13-epiestrones, introduction of a large ring resulted in various effects. In general, although some phenylalkynylated 13-epiestrones showed detectable interaction with OATP2B1, these compounds were less effective inhibitors than the C-2 halogenated epiestrones (see Table 2 and Fig. 5). The only exception is **1gS_4** containing a (4-methoxyphenyl)ethynyl substituent, that was almost as effective inhibitor as the C-2 halogenated 13-epiestrones. Interestingly, in the case of phenylalkynylated 3-OH 13-epiestrones no clear rule in preference in interaction with C-2 over C-4 modified compounds could be observed. Compounds **1eS_2** and **1gS_4** were more potent inhibitors than their C-2 counterparts (**1eS_2** and **1gS_2**). However, in the case of 13-epiestrones bearing fluorinated substituents (**1hS** and **1fS**), this tendency changed, C-2 compounds being slight inhibitors (EC_{50} around 10 μ M) compared to C-4 modified derivatives lacking any inhibitory potential (at least in the concentrations tested).

In the case of the 3-OMe compounds with a second, subst. phenylalkynyl modification, similarly to that observed for the halogenated

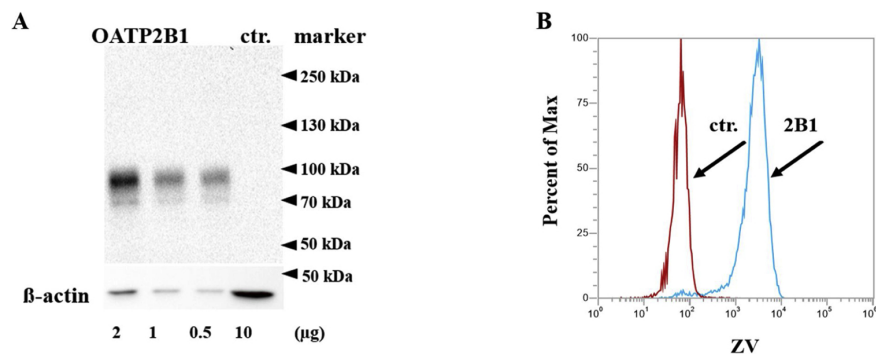


Fig. 2. A) Western blot detection of OATP2B1 expressed in A431 cells. OATP2B1 was detected by an anti-OATP2B1 antibody [27], and β -actin was used as a loading control. Control (ctr.) stands for mock transfected A431 cells. Multiple migratory bands may represent differentially glycosylated forms of OATP2B1. B) Zombie Violet uptake in A431-OATP2B1 and control cells. Histograms show the uptake of ZV (250x dilution) in the cells incubated with the fluorescent dye for 15 min at 37 °C in uptake buffer (pH 5.5). Living (propidium-iodide negative) cells are shown. Mock transfected cells are indicated with a red line and OATP2B1 transfected are with blue.

compounds, a decrease or even a complete loss of inhibition compared to the parental compound 3-OMe 13-epi estrone could be observed. The only exceptions were compounds **2gS_2** and **2gS_4** that proved to be effective inhibitors. A clear rule of regioselectivity could not be observed for the 3-OMe phenylalkynyl modified compounds.

3.3. Effect of 2-bromo 13-epi estrone on OATP2B1-mediated estrone-3-sulfate uptake

Inhibition of hormone uptake by OATP2B1 in tumor cells could be the major goal of the newly identified inhibitors. Therefore, in order to

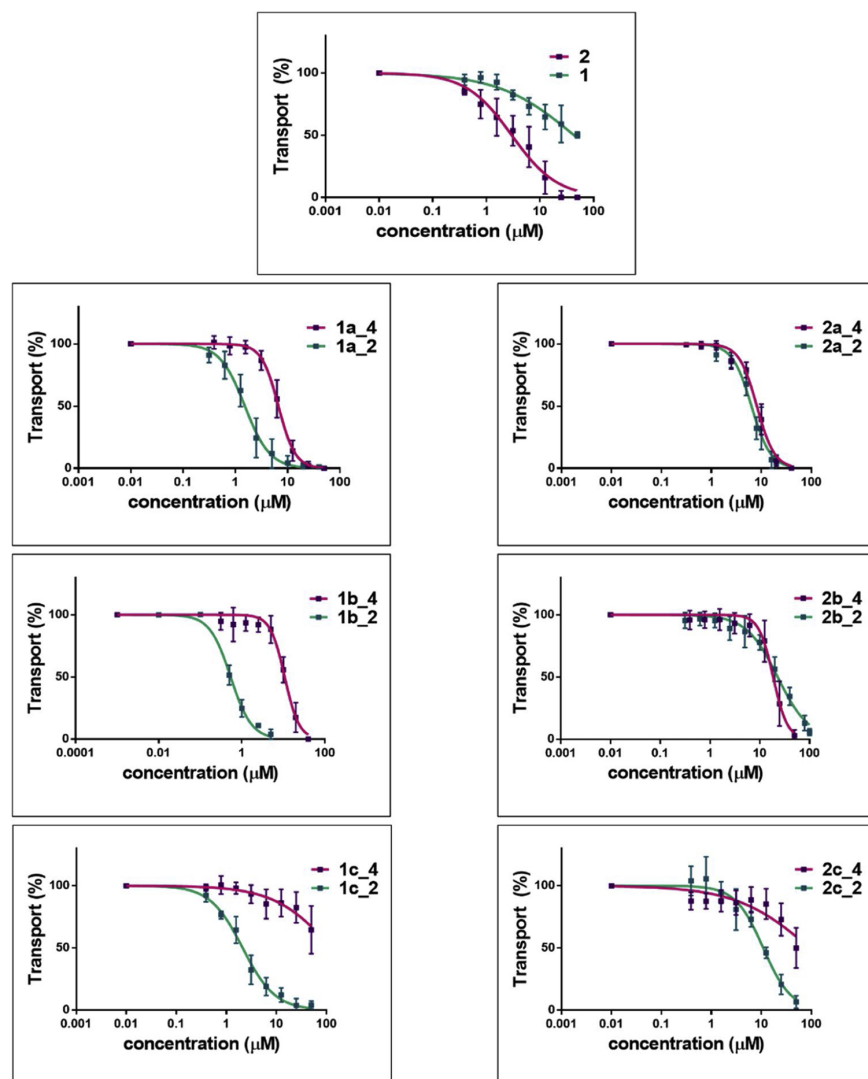


Fig. 3. Inhibition of dye uptake in A431-OATP2B1 cells by halogenated 13-epi estrones. A431-OATP2B1 cells and their mock transfected counterparts were incubated in the presence or absence of increasing amounts of the 13-epi estrones (1–100 μ M, as indicated on the x axis) with the Zombie Violet dye for 30 min at 37 °C. Fluorescence was measured in an Enspire plate reader. Fluorescence obtained in mock transfected A431 cells was subtracted from that measured in A431-OATP2B1 cells. Transport was calculated based on the fluorescence measured in the absence of 13-epi estrones (100 %). Data points show the average \pm SD values obtained in at least three independent biological replicates.

Table 1
Inhibition of OATP2B1 activity by C-2 or C-4 halogenated 13-epiestrones.

compound "name"	EC ₅₀ ± SD (μM)	compound "name"	EC ₅₀ ± SD (μM)	compound "name"	EC ₅₀ ± SD (μM)	compound "name"	EC ₅₀ ± SD (μM)
1	50			2	2.98 ± 0.05		
1a_2	1.52 ± 0.01	1a_4	6.63 ± 0.01	2a_2	6.34 ± 0.02	2a_4	8.14 ± 0.02
1b_2	0.54 ± 0.02	1b_4	10.8 ± 0.02	2b_2	22.88 ± 0.02	2b_4	18.6 ± 0.02
1c_2	2.11 ± 0.02	1c_4	> 50	2c_2	10.89 ± 0.03	2c_4	> 50

The inhibitory effect of 13-epiestrones was measured using *Zombie Violet* as a test substrate [26]. The kinetic parameters of inhibition shown on Fig. 3 were determined by the Graphpad Prism software.

determine whether the best performing newly identified inhibitor, 2-bromo-13-epiestrone (compound **1b_2**) can be applied to inhibit hormone uptake, we determined its effect on ³H-E1S uptake in A431 control and A431-OATP2B1 cells. Fig. 6 shows that compound **1b_2** can attenuate E1S uptake mediated by OATP2B1, therefore it is a good candidate to block hormone uptake in OATP2B1 expressing cells.

3.4. SAR analysis

Since the compounds under study do all possess a common core structure (scaffold) with largest variance in terms of different substituents in positions C-2 and C-4, it appears interesting to perform Structure-Activity Relationship (SAR) analysis for positions C-2 and C-4 separately. The rationale behind is that the increase or decrease of bioactivities within a congeneric SAR series of compounds can possibly be explained by the variations in physicochemical properties at a specific substitution site (R or X-group site). However, since the initial (parent) compounds 3-OH and 3-OMe are possessing very different inhibitory potential (EC₅₀ value of compound **1** (3-OH) is 50 μM while for compound **2** (3-OMe) is 2.98 μM), we have performed the SAR analysis for 3-OH and 3-OMe derivatives separately.

In the case of the 3-OH derivatives, the SAR analysis (Table 3) for substituents at position C-2 shows that the number of atoms and heavy atoms in the substituent is negatively correlated with activity ($R = -0.89$ and -0.84). Of equal effect is the number of aromatic carbocycles in that side chain: possessing no rings is more favorable ($R = -0.90$). Further, molar refractivity ('SMR') which reflects the charge distribution and hence corresponds to the polarizability of a given functional group is inversely correlated to bioactivity: less polarizable is more favourable ($R = -0.88$). Also Labute's Approximative Surface Area ('LabuteASA') ($R = -0.82$) [29] - a measure of the size of a molecules' surface area - as well as the partition coefficient ('SlogP') are negatively

correlated ($R = -0.73$) to activity, corresponding to a favorable lower size and lipophilicity of substituents at position C-2. Concrete values of Pearson correlation coefficients for given descriptors are listed in Table 3 (a heat map representation of correlation values is given in Supplementary Figure S1).

The only significant positive correlation at position C-2 was identified for the HallKier Alpha index ('HallKier α', $R = 0.88$), as introduced by Hall and Kier (equation nr. 58 in provided reference) [30]. HallKier α belongs to the class of topological descriptors which can quantify molecular shape similarity within a set of molecules. HallKier α relates to the size contribution of a query fragment to C(sp³)-hybridized atoms, which are taken as a reference (HallKier α for sp³ carbon equals to 0). HallKier α thus encodes the effect of both covalent radius and hybridization state of a given group of atoms. From Supplementary Table 1 it becomes clear that the halogenated derivatives are generally having positive HallKier α values and the phenylalkynyl compounds are showing negative values. Interestingly, in the case of 3-OMe derivatives none of the investigated physicochemical properties of the C-2 substituents showed a meaningful correlation with the bioactivity values of these compounds.

In contrast to substituents at position C-2, the SAR analysis for substituents at position C-4 did not allow to prioritize physicochemical features influencing the overall bioactivity on OATP2B1 (Table 3 and Supplementary Figure S1). In general, correlations of physicochemical features and bioactivity for substituents in position C-4 are comparable for both 3-OH and 3-OMe derivatives. Most strikingly, the number of heteroatoms of substituents at position C-4 is slightly negatively correlated with bioactivity for both 3-OH and 3-OMe derivatives ($R = -0.47$ and -0.45). Due to the chemical composition of our compounds, this effect points to an unfavorable effect of electronegative atoms at this position (Cl, I, Br, F) which is inverse to the trends observed for position C-2 (where halogens seem to be favorable).

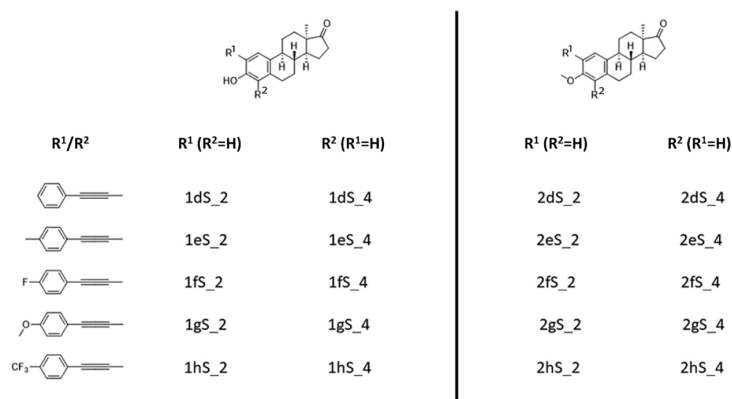


Fig. 4. Structure of the phenylalkynyl modified 13-epiestrones investigated in the current study.

Table 2
Inhibition of OATP2B1 activity by C-2 or C-4 phenylalkynylated 13-epiestrones.

compound "name"	EC ₅₀ ± SD (μM)	compound "name"	EC ₅₀ ± SD (μM)	compound "name"	EC ₅₀ ± SD (μM)	compound "name"	EC ₅₀ ± SD (μM)
1dS_2	15.53 ± 0.03	1dS_4	11.55 ± 0.02	2dS_2	> 50	2dS_4	12.83 ± 0.04
1eS_2	> 50	1eS_4	11.36 ± 0.02	2eS_2	20.77 ± 0.03	2eS_4	11.86 ± 0.02
1fS_2	12.95 ± 0.01	1fS_4	> 50	2fS_2	> 50	2fS_4	8.73 ± 0.02
1gS_2	11.08 ± 0.04	1gS_4	4.57 ± 0.01	2gS_2	3.79 ± 0.05	2gS_4	3.4 ± 0.02
1hS_2	9.27 ± 0.04	1hS_4	> 50	2hS_2	10.5 ± 0.03	2hS_4	46.09 ± 0.04

4. Discussion

Recognizing numerous clinically applied drugs and promoting their intestinal, hepatic and central nervous system (through the blood-brain-barrier) uptake, OATP2B1 is a key determinant of drug pharmacokinetics [31]. Hence understanding the mechanism of its substrate/inhibitor recognition can promote drug development and may also help in predicting/avoiding adverse effects caused by OATP2B1 mediated drug-drug interactions. In addition, OATP2B1 is a dedicated conjugated steroid hormone transporter. Its steroid hormone substrates are E1S, pregnenolone-sulfate and DHEAS [32,33]. OATP2B1 has been identified as a key uptake transporter of DHEAS and E1S in the placenta and mammary gland that are largely dependent on these hormone precursors [15,34]. In addition, OATP2B1 expressed in endothelial cells of the blood-brain barrier is considered as an important mediator of the uptake of the neuroactive steroids DHEAS and pregnenolone-sulfate into the brain [35]. On the other hand, increased steroid hormone uptake by OATP2B1 may also be favorable for tumor progression, as was demonstrated in breast and prostate cancer [36,37]. Therefore, inhibition of OATP2B1 function may be an alternative/successful strategy to inhibit the growth of various tumors.

During the last two decades, since OATP2B1 was cloned [27], numerous inhibitors of OATP2B1 have been described. Most of these are clinically applied drugs, like cyclosporin A, rifampicin and statins [38,39]. However these compounds are also inhibiting other drug transporters, like P-glycoprotein [40] and additional OATPs or even CYP enzymes [6,41], therefore they are lacking OATP2B1 specificity. In addition, various steroids, e.g. estrone or testosterone that are themselves not transported by OATP2B1 have been documented as OATP2B1 inhibitors [32]. However, besides again the lack of specificity, these steroids are not effective inhibitors, since only low levels of inhibition could be observed even at concentrations well above their physiological occurrence. Grube et al. documented only 20–30 % decrease in OATP2B1-mediated E1S uptake by the application of 10 or 100 μM testosterone or estrone, respectively [32].

13-epiestrones represent a new class of OATP2B1 inhibitors. They have no steroidogenic effect, hence their application may be void of side effects. In our preliminary work we found that phosphonated 13-epiestrones are potent inhibitors of OATP2B1 function, with EC₅₀ values in the micromolar range [24]. In order to map the molecular determinants of this inhibition, here we analyzed the inhibitory effect of a series of 3-hydroxy or 3-methylether 13-epiestrones containing a second, C-2 or C-4 halogen or phenylalkynyl modification. **Compound 1**, (3-OH 13-epiestrone) showed no interaction with OATP2B1 (EC₅₀ around 50 μM), but the 3-methoxy counterpart **compound 2** (3-OMe 13-epiestrone) performed a strong interaction (EC₅₀ 2.98 μM). Introduction of a second modification on C-2 or C-4 resulted in various effects. In the case of 3-OMe 13-epiestrones, the second modification issued in a decrease or loss of inhibitory activity with the exception of 2-iodo or 4-iodo (**2a_2** and **2a_4**), (4-fluorophenyl)ethynyl (**2fS_4**), (2-methoxyphenyl)ethynyl (**2gS_2**) and (4-methoxyphenyl)ethynyl

(**2gS_4**) derivatives having similar EC₅₀ values as the initial compound 3-OMe 13-epiestrone. In contrast, in the case of 13-epiestrone, the second modification had various effects depending on the site (C-2 or C-4) or nature (halogen or phenylalkynyl) of the substitution. In general, introduction of a phenylalkynyl substituent did not result in potent inhibitors. The only exception was the C-4 modified compound **1gS_4** having an EC₅₀ value well below 10 μM. However, the most striking change in the inhibitory effect was observed in the case of the 2-halogenated 13-epiestrones (**1a_2**, **1b_2** and **1c_2**). These compounds potentially inhibited OATP2B1 function with EC₅₀ values between 0.5 and 2.1 μM. In addition, halogenated 13-epiestrones revealed a strong regioselectivity, 4-halogenated compounds showing no or very weak inhibition of OATP2B1 activity. This C-2 halogen preference has already been observed in the case of 2-iodo-estrone-3-sulfate [42]. Banerjee and colleagues have demonstrated that 2-[¹²⁵I]-estrone-3-sulfate is transported by OATP2B1 while transport in the case of its 4-iodo counterpart could not be observed.

SAR analysis by R-group decomposition was performed in order to elucidate molecular determinants potentially being responsible for bioactivity of 13-epiestrones. Different potency of respective compounds was correlated to the subtle changes in physico-chemical properties at a specific R or X -group position. It has to be emphasized that the observed trends are showing correlations of side chain features and bioactivity, but these correlations are not necessarily pointing to a causal relationship of activity and chemical descriptor value. In other words, some of the correlations we see might be artefacts caused by high intercorrelation of related features (e.g. molecular weight and lipophilicity are often intercorrelated). The presence of halogens in position C-2 was identified to drive the activity against OATP2B1. The importance of fluor and 4-fluorophenyl functional groups for OATP2B1 substrate activity was already demonstrated by the substructural fragment analysis performed by Shaikh et al. [43]. Our findings are suggesting the likelihood of halogen bond formation in OATP2B1-ligand binding complexes. As an outlook, we suggest to prove such a hypothesis by e.g. molecular docking combined with quantum mechanics techniques. SAR analysis of functional groups at position C-4 delivered negative correlation with the number of heteroatoms. These trends, however, appear to be less pronounced for position C-4 substituents and therefore it is required to repeat the analyses with a bigger data set showing a larger range of structural variations in this position for further investigations. SAR analysis revealed that 3-OH derivatives are showing more pronounced positive and negative correlation with different physicochemical properties at position C-2. In general, 3-OH derivatives are more sensitive to the substitutions at position C-2 when compared to 3-OMe derivatives.

In our previous study we investigated 3-hydroxy, 3-methylether and 3-benzylether (3-OBn) 13-epiestrones with a C-2 or C-4 diethyl phosphono or diphenylphosphine oxide substitution [24]. In that study, in accordance with our current findings, we found that the second (C-2 or C-4) modification results in dramatic increase in the inhibitory potential of 13-epiestrone (3-OH). Also in harmony with our current results,

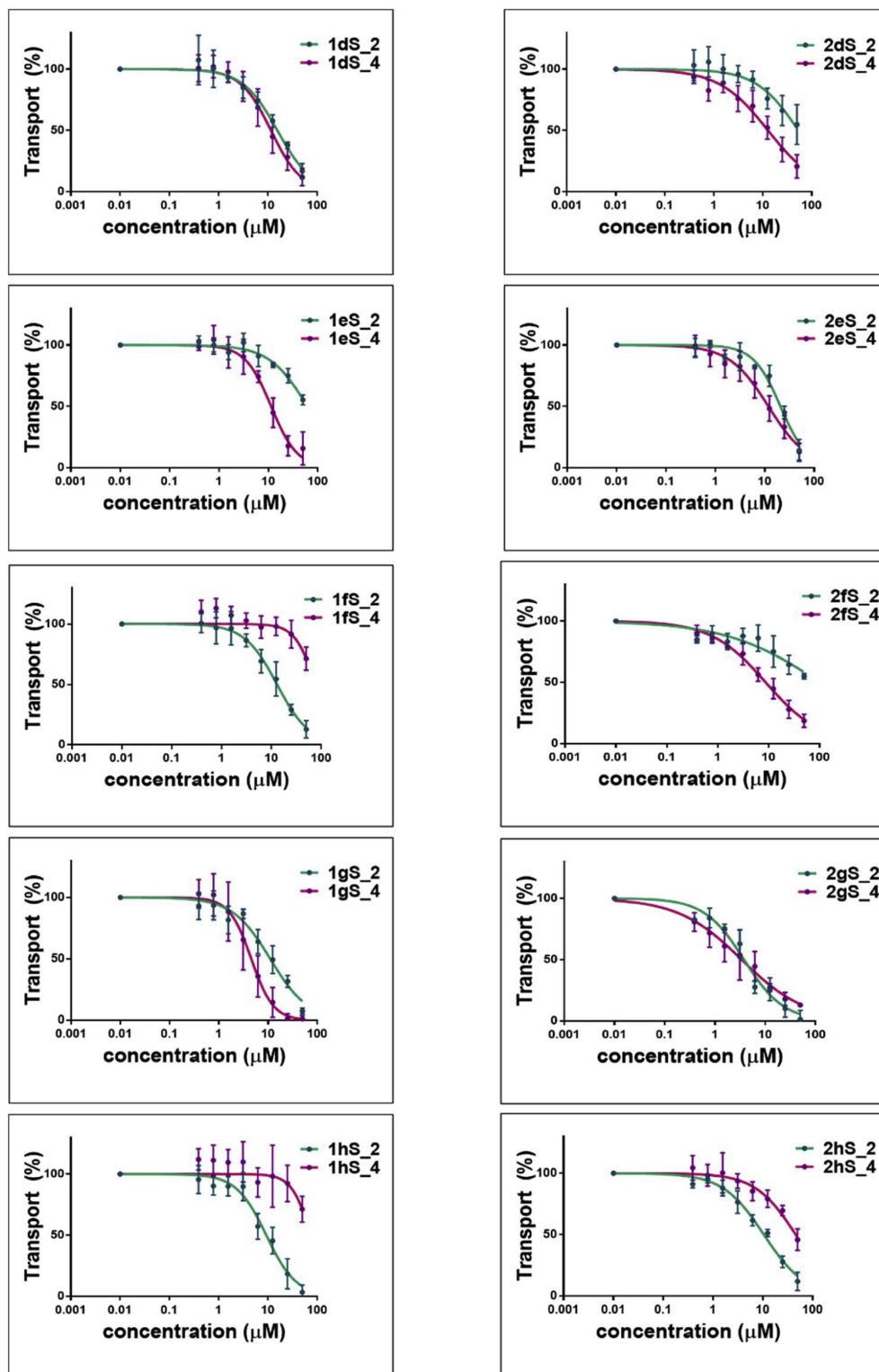


Fig. 5. Inhibition of dye uptake in A431-OATP2B1 cells by phenylalkynylated 13-epiestrones. Inhibition of Zombie Violet uptake in A431-OATP2B1 and mock transfected cells was measured as described at Fig. 3. Average \pm SD values obtained in at least three independent biological replicates are shown.

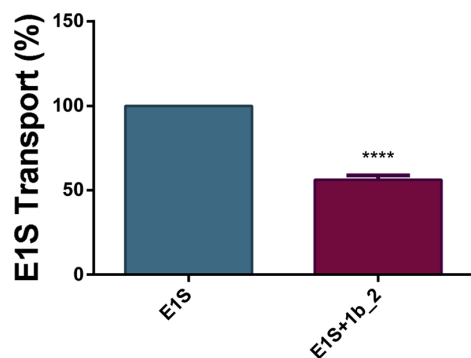


Fig. 6. Inhibition of ^3H -E1S uptake in A431-OATP2B1 cells by 2-bromo-13-epiestrone (1b₂). A431-OATP2B1 cells and their mock transfected counterparts were incubated in the presence or absence of 50 μM 1b₂ with 9.65 nM ^3H -E1S for 10 min at 37 °C. The radioactivity was measured by Wallac Liquid Scintillator Counter. Radioactivity obtained in mock transfected A431 cells was subtracted from that measured in A431-OATP2B1 cells. Transport was calculated based on the measured radioactivity of ^3H -E1S in the absence of 2-bromo-13-epiestrone (100 %). Data points show the average \pm SD values obtained in three independent biological replicates. Significance was calculated with GraphPad Prism, using unpaired t-test, **** p < 0.0001.

Table 3

Pearson correlation coefficients for physico-chemical descriptors at position C-2 and C-4.

A) 3-OH derivatives		
Descriptor name	Pearson correlation coefficient	
	Position C-2	Position C-4
NumAtoms	−0.89	0.01
NumHeavyAtoms	−0.84	−0.13
NumHeteroAtoms	0.23	−0.47
NumAromaticCarbocycles	−0.90	−0.06
SMR	−0.88	0.08
LabuteASA	−0.82	−0.03
SlogP	−0.73	−0.26
HallKierAlpha	0.88	0.13
B) 3-OMe derivatives		
Descriptor name	Pearson correlation coefficient	
	Position C-2	Position C-4
NumAtoms	−0.05	0.31
NumHeavyAtoms	−0.06	0.20
NumHeteroAtoms	0.39	−0.45
NumAromaticCarbocycles	−0.21	0.30
SMR	−0.05	0.39
LabuteASA	−0.01	0.28
SlogP	−0.08	−0.04
HallKierAlpha	0.16	−0.23

the second modification could not further improve the potent inhibition by 3-OMe 13-epiestrone. However, in the case of phosphonated 13-epiestrones, the OATP2B1 inhibitory action did not substantially depend on the regioisomerism (C-2 vs. C-4).

HSD17 β 1 and STS are key enzymes in local estrogen production. Inhibition of their activity by specific or more desirably by dual inhibitors could be a good strategy to prevent tumor progression [44]. The most promising STS inhibitor STX-64, Irosustat performed well in phase I trial, however the phase II trial was not that satisfactory [44], suggesting the need for combined treatment with HSD17 β 1 inhibitors [45]. Moreover, it has been demonstrated that blocking the aromatase pathway resulted in the upregulation of the STS enzyme and OATPs [46]. Therefore, simultaneous blockage of the different estrogen metabolic pathways and the function of the steroid uptake transporter

Table 4

Comparison of the inhibitory effect of 13-epiestrones on OATP2B1, HSD17 β 1 and STS activity.

C-3	C-2	C-4	Inhibition of OATP2B1 (determined in the current study)	Inhibition of HSD17 β 1 (as described earlier in [21,25])	STS (as described earlier in [21])
OH	H	H	non	+	non
	halogen	H	+ / + +	+ +	non
	phenylalkynyl	H	non	+ +	+ / non
	H	halogen	+ / non	+	+ / non
	H	phenylalkynyl	+ / non	non	non
OMe	H	H	+	+	n.d.
	halogen	H	+ / non	non	n.d.
	phenylalkynyl	H	non	non	n.d.
	H	halogen	+ / non	+ +	n.d.
	H	phenylalkynyl	+ / non	non	n.d.

Columns labeled with C-2, C-3 or C-4 indicate modifications on the 2nd, 3rd or 4th carbon of 13-epiestrone. n.d.: no data available.

+ : EC50 below 10 μM .

+ + : EC50 below 1 μM .

OATPs could be the only successful strategy to inhibit hormone dependent cancers. Our experiments show that the newly identified best performing inhibitor, 2-bromo-13-epiestrone (compound 1b₂) can be used to inhibit OATP2B1-mediated E1S uptake (Fig. 6). Some of the inhibitors identified in our current study, including compound 1b₂, are also potent inhibitors of the HSD17 β 1 enzyme (see Table 4), therefore they can be good candidate dual inhibitors to be tested in hormone dependent cell lines.

Although OATP2B1 is not related evolutionally to the STS and HSD17 β enzyme families, considering their overlapping inhibitor specificities, one may speculate that knowledge gathered from the inhibitor recognition profile of these enzymes can be used to design effective inhibitors of OATP2B1. This is especially important since a protein structure of OATPs is not yet available. Therefore, we have compared the inhibition data obtained in the current study for OATP2B1 with that previously measured for HSD17 β 1 and STS enzymes [21,25]. Table 4 shows, that inhibition of HSD17 β 1 reveals few similar features to that of the inhibition of OATP2B1.

Namely, HSD17 β 1 also has a C-2 preference and both 2-halogenated and phenylalkynyl conjugates potently inhibit HSD17 β 1 function. Also, similarly to OATP2B1, 3-OMe 13-epiestrones are less effective inhibitors of HSD17 β 1. However, C-4 halogenation of 3-OMe 13-epiestrones also results in effective inhibitors that is in contrast to that observed for OATP2B1. Unfortunately, data about the effect of the 13-epiestrones investigated in the current study on STS activity are incomplete. Nevertheless, although estrone-3-sulfate is a common substrate of OATP2B1 and STS, based on the interaction with halogenated 13-epiestrones, opposite inhibitor preference could be observed for OATP2B1 and STS. STS was only inhibited by some of the C-4 halogenated compounds that were not the most effective inhibitors of OATP2B1. We suggest that a larger data set of HSD17 β 1, STS, OATP and 13-epiestrone interactions should be generated in order to determine whether common trends in structural elements important for their inhibition can be observed. Also it would be interesting to investigate the inhibitory effect of 13-epiestrones on the function of other OATPs up-regulated in hormone dependent cancers, OATP1A2, OATP1B3, OATP3A1 and OATP4A1 [47]. On the other hand, hepatic OATPs, 1B1, 1B3 and 2B1 have overlapping substrate specificities, hence, although desirable, specific inhibitors amenable to distinguish between their function are scarce. The OATP2B1 inhibitor 13-epiestrones (1a₂, 1b₂ and 1c₂) identified in the current study can be good candidates to be tested for OATP1B interaction.

In summary, we identify potent inhibitors of OATP2B1. The EC₅₀ of the most potent inhibitor 2-bromo-13-epiestrone falls within the range

of previously documented OATP2B1 inhibitors (BSP 1.26 μM [26], antivirals: 0.5–1 μM , erlotinib: 0.03 μM [48] and E1S: 0.56 μM [26]. However, although potent inhibitors were identified, our assay cannot distinguish between an inhibition caused by transported substrates or non-competitive inhibitors. Further experiments, e.g. measurement of direct uptake of the best performing 13-epiestrones (showing the highest inhibition) are needed to clarify this issue. Still, one may speculate that if OATP2B1 can mediate the uptake of these HSD17 β 1 inhibitors, a more potent anti-tumor effect can be achieved in tumors expressing both HSD17 β 1 and OATP2B1. Since certain 2-halogenated-13-epiestrones, and the previously investigated phosphonated 13-epiestrones [24] are dual inhibitors of HSD17 β 1 and OATP2B1, their effect on the survival of hormone dependent cell lines with OATP2B1 overexpression, and/or HSD17 β 1 expression is reasonable to be investigated.

CRedit authorship contribution statement

Réka Laczkó-Rigó: Methodology, Visualization, Writing - original draft. **Rebeka Jójárt:** Resources. **Erzsébet Mernyak:** Resources, Writing - review & editing. **Éva Bakos:** Methodology, Supervision, Conceptualization. **Albeta Tuerkova:** Methodology, Formal analysis. **Barbara Zdrzil:** Methodology, Formal analysis, Writing - review & editing. **Csilla Özvegy-Laczka:** Conceptualization, Writing - review & editing.

Acknowledgements

This work has been supported by research grants from the National Research, Development and Innovation Office (OTKA FK 128751 and SNN 124329). E. M. and Cs. Ö-L. are recipients of the János Bolyai fellowship of the Hungarian Academy of Sciences. This work also received funding from the Austrian Science Fund (FWF) (Grant P 29712).

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:<https://doi.org/10.1016/j.jsmb.2020.105652>.

References

- [1] B. Hagenbuch, B. Stieger, The SLCO (former SLC21) superfamily of transporters, *Mol. Aspects Med.* 34 (2013) 396–412.
- [2] J. König, Y. Cui, A.T. Nies, D. Keppler, A novel human organic anion transporting polypeptide localized to the basolateral hepatocyte membrane, *Am. J. Physiol. Gastrointest. Liver Physiol.* 278 (2000) G156–64.
- [3] M. Roth, A. Obaidat, B. Hagenbuch, OATPs, OATs and OCTs: the organic anion and cation transporters of the SLCO and SLC22A gene superfamilies, *Br. J. Pharmacol.* 165 (2012) 1260–1287.
- [4] A. Obaidat, M. Roth, B. Hagenbuch, The expression and function of organic anion transporting polypeptides in normal tissues and in cancer, *Annu. Rev. Pharmacol. Toxicol.* 52 (2012) 135–151.
- [5] A. Koenen, K. Kock, M. Keiser, W. Siegmund, H.K. Kroemer, M. Grube, Steroid hormones specifically modify the activity of organic anion transporting polypeptides, *Eur. J. Pharm. Sci.* 47 (2012) 774–780.
- [6] A. Kallikowski, M. Niemi, Impact of OATP transporters on pharmacokinetics, *Br. J. Pharmacol.* 158 (2009) 693–705.
- [7] V. Buxhofer-Ausch, L. Secky, K. Wlcek, M. Svoboda, V. Kounnis, E. Briassoulis, A.G. Tzakos, W. Jaeger, T. Thalhammer, Tumor-specific expression of organic anion-transporting polypeptides: transporters as novel targets for cancer therapy, *J. Drug Deliv.* 2013 (2013) 863539.
- [8] R.R. Schulte, R.H. Ho, Organic anion transporting polypeptides: emerging roles in cancer pharmacology, *Mol. Pharmacol.* 95 (2019) 490–506.
- [9] K. Wlcek, M. Svoboda, T. Thalhammer, F. Sellner, G. Krupitza, W. Jaeger, Altered expression of organic anion transporter polypeptide (OATP) genes in human breast carcinoma, *Cancer Biol. Ther.* 7 (2008) 1450–1455.
- [10] D. Kovács, I. Patik, C. Özvegy-Laczka, The role of organic anion transporting polypeptides in drug absorption, distribution, excretion and drug-drug interactions, *Expert Opin. Drug Metab. Toxicol.* 13 (2017) 409–424.
- [11] T. Nakanishi, I. Tamai, Genetic polymorphisms of OATP transporters and their impact on intestinal absorption and hepatic disposition of drugs, *Drug Metab. Pharmacokinet.* 27 (2012) 106–121.
- [12] B. Gao, S.R. Vavricka, P.J. Meier, B. Stieger, Differential cellular expression of

- organic anion transporting peptides OATP1A2 and OATP2B1 in the human retina and brain: implications for carrier-mediated transport of neuropeptides and neurosteroids in the CNS, *Pflügers Arch.* 467 (2015) 1481–1493.
- [13] J. Kindla, T.T. Rau, R. Jung, P.A. Fasching, R. Strick, R. Stoeck, A. Hartmann, M.F. Fromm, J. König, Expression and localization of the uptake transporters OATP2B1, OATP3A1 and OATP5A1 in non-malignant and malignant breast tissue, *Cancer Biol. Ther.* 11 (2011) 584–591.
 - [14] J. Matsumoto, N. Ariyoshi, M. Sakakibara, T. Nakanishi, Y. Okubo, N. Shiina, K. Fujisaki, T. Nagashima, Y. Nakatani, I. Tamai, H. Yamada, H. Takeda, I. Ishii, Organic anion transporting polypeptide 2B1 expression correlates with uptake of estrone-3-sulfate and cell proliferation in estrogen receptor-positive breast cancer cells, *Drug Metab. Pharmacokinet.* 30 (2015) 133–141.
 - [15] F. Pizzagalli, Z. Varga, R.D. Huber, G. Folkers, P.J. Meier, M.V. St-Pierre, Identification of steroid sulfate transport processes in the human mammary gland, *J. Clin. Endocrinol. Metab.* 88 (2003) 3902–3912.
 - [16] S.M. Green, A. Kaipainen, K. Bullock, A. Zhang, J.M. Lucas, C. Matson, W.A. Banks, E.A. Mostaghel, Role of OATP transporters in steroid uptake by prostate cancer cells *in vivo*, *Prostate Cancer Prostatic Dis.* 20 (2017) 20–27.
 - [17] N. Fujimoto, T. Kubo, H. Inatomi, H.T. Bui, M. Shiota, T. Shio, T. Matsumoto, Polymorphisms of the androgen transporting gene SLCO2B1 may influence the castration resistance of prostate cancer and the racial differences in response to androgen deprivation, *Prostate Cancer Prostatic Dis.* 16 (2013) 336–340.
 - [18] M. Yang, W. Xie, E. Mostaghel, M. Nakabayashi, L. Werner, T. Sun, M. Pomerantz, M. Freedman, R. Ross, M. Regan, N. Sharifi, W.D. Figg, S. Balk, M. Brown, M.E. Taplin, W.K. Oh, G.S. Lee, P.W. Kantoff, SLCO2B1 and SLCO1B3 may determine time to progression for patients receiving androgen deprivation therapy for prostate cancer, *J. Clin. Oncol.* 29 (2011) 2565–2573.
 - [19] D. Ayan, J. Roy, R. Maltais, D. Poirier, Impact of estradiol structural modifications (18-methyl and/or 17-hydroxy inversion of configuration) on the *in vitro* and *in vivo* estrogenic activity, *J. Steroid Biochem. Mol. Biol.* 127 (2011) 324–330.
 - [20] B. Schonecker, C. Lange, M. Kotteritzsch, W. Gunther, J. Weston, E. Anders, H. Gohl, Conformational design for 13alpha-steroids, *J. Org. Chem.* 65 (2000) 5487–5497.
 - [21] I. Bacsa, B.E. Herman, R. Jojart, K.S. Herman, J. Wolfing, G. Schneider, M. Varga, C. Tomboly, T.L. Rizner, M. Szecsi, E. Mernyak, Synthesis and structure-activity relationships of 2- and/or 4-halogenated 13beta- and 13alpha-estrone derivatives as enzyme inhibitors of estrogen biosynthesis, *J. Enzyme Inhib. Med. Chem.* 33 (2018) 1271–1282.
 - [22] J.M. Tian, B. Ran, C.L. Zhang, D.M. Yan, X.H. Li, Estrogen and progesterone promote breast cancer cell proliferation by inducing cyclin G1 expression, *Braz. J. Med. Biol. Res.* 51 (2018) 1–7.
 - [23] M. Yang, J. Wang, L. Wang, C. Shen, B. Su, M. Qi, J. Hu, W. Gao, W. Tan, B. Han, Estrogen induces androgen-repressed SOX4 expression to promote progression of prostate cancer cells, *Prostate* 75 (2015) 1363–1375.
 - [24] R. Jojart, S. Pecs, G. Keglevich, M. Szecsi, R. Rigo, C. Özvegy-Laczka, G. Kecskemeti, E. Mernyak, Pd-Catalyzed microwave-assisted synthesis of phosphonated 13alpha-estrone derivatives as potential OATP2B1, 17beta-HSD1 and/or STS inhibitors, *Beilstein J. Org. Chem.* 14 (2018) 2838–2845.
 - [25] I. Bacsa, R. Jojart, J. Wolfing, G. Schneider, B.E. Herman, M. Szecsi, E. Mernyak, Synthesis of novel 13alpha-estrone derivatives by Sonogashira coupling as potential 17beta-HSD1 inhibitors, *Beilstein J. Org. Chem.* 13 (2017) 1303–1309.
 - [26] I. Patik, V. Szekely, O. Nemet, A. Szepesi, N. Kucsma, G. Varady, G. Szakacs, E. Bakos, C. Özvegy-Laczka, Identification of novel cell-impermeant fluorescent substrates for testing the function and drug interaction of Organic Anion-Transporting Polypeptides, OATP1B1/1B3 and 2B1, *Sci. Rep.* 8 (2018) 2630.
 - [27] G.A. Kullak-Ublick, M.G. Ismail, B. Stieger, L. Landmann, R. Huber, F. Pizzagalli, K. Fattinger, P.J. Meier, B. Hagenbuch, Organic anion-transporting polypeptide B (OATP-B) and its functional comparison with three other OATPs of human liver, *Gastroenterology* 120 (2001) 525–533.
 - [28] M.R.C. Berthold, N. Dill, F. Gabriel, T.R. Kotter, T.O. Meinel, P. Thiel, K. B. Wiswedel, KNIME - the konstan information miner version 2.0 and beyond *AcM SIGKDD explorations Newsletter*, 11 (2009), pp. 26–31.
 - [29] P. Labute, A widely applicable set of descriptors, *J. Mol. Graph. Model.* 18 (2000) 464–477.
 - [30] L.H.K. Hall, L. B, The molecular connectivity chi indexes and kappa shape indexes in structure-property modeling, *Rev. Comput. Chem.* (1991) 367–422.
 - [31] S.J. McFeely, L. Wu, T.K. Ritchie, J. Unadkat, Organic anion transporting polypeptide 2B1 - more than a glass-full of drug interactions, *Pharmacol. Ther.* 196 (2019) 204–215.
 - [32] M. Grube, K. Kock, S. Karner, S. Reuther, C.A. Ritter, G. Jedlitschky, H.K. Kroemer, Modification of OATP2B1-mediated transport by steroid hormones, *Mol. Pharmacol.* 70 (2006) 1735–1741.
 - [33] I. Tamai, T. Nozawa, M. Koshida, J. Nezu, Y. Sai, A. Tsuji, Functional characterization of human organic anion transporting polypeptide B (OATP-B) in comparison with liver-specific OATP-C, *Pharm. Res.* 18 (2001) 1262–1269.
 - [34] M.V. St-Pierre, B. Hagenbuch, B. Ugele, P.J. Meier, T. Stallmach, Characterization of an organic anion-transporting polypeptide (OATP-B) in human placenta, *J. Clin. Endocrinol. Metab.* 87 (2002) 1856–1863.
 - [35] M. Grube, P. Hagen, G. Jedlitschky, Neurosteroid transport in the brain: role of ABC and SLC transporters, *Front. Pharmacol.* 9 (2018) 354.
 - [36] W. Al Sarakbi, R. Mokbel, M. Salhab, W.G. Jiang, M.J. Reed, K. Mokbel, The role of STS and OATP-B mRNA expression in predicting the clinical outcome in human breast cancer, *Anticancer Res.* 26 (2006) 4985–4990.
 - [37] T. Nozawa, M. Suzuki, H. Yabuuchi, M. Irokawa, A. Tsuji, I. Tamai, Suppression of cell proliferation by inhibition of estrone-3-sulfate transporter in estrogen-dependent breast cancer cells, *Pharm. Res.* 22 (2005) 1634–1641.

- [38] J. König, H. Glaeser, M. Keiser, K. Mandery, U. Klotz, M.F. Fromm, Role of organic anion-transporting polypeptides for cellular mesalazine (5-aminosalicylic acid) uptake, *Drug Metab. Dispos.* 39 (2011) 1097–1102.
- [39] I.Y. Gong, R.B. Kim, Impact of genetic variation in OATP transporters to drug disposition and response, *Drug Metab. Pharmacokinet.* 28 (2013) 4–18.
- [40] R.B. Kim, Drugs as P-glycoprotein substrates, inhibitors, and inducers, *Drug Metab. Rev.* 34 (2002) 47–54.
- [41] A. Koenen, H.K. Kroemer, M. Grube, H.E. Meyer zu Schwabedissen, Current understanding of hepatic and intestinal OATP-mediated drug-drug interactions, *Expert Rev. Clin. Pharmacol.* 4 (2011) 729–742.
- [42] N. Banerjee, T.R. Wu, J. Chio, R. Kelly, K.A. Stephenson, J. Forbes, C. Allen, J.F. Valliant, R. Bendayan, (125)I-Labelled 2-Iodoestrone-3-sulfate: synthesis, characterization and OATP mediated transport studies in hormone dependent and independent breast cancer cells, *Nucl. Med. Biol.* 42 (2015) 274–282.
- [43] N. Shaikh, M. Sharma, P. Garg, Selective fusion of heterogeneous classifiers for predicting substrates of membrane transporters, *J. Chem. Inf. Model.* 57 (2017) 594–607.
- [44] X. Sang, H. Han, D. Poirier, S.X. Lin, Steroid sulfatase inhibition success and limitation in breast cancer clinical assays: an underlying mechanism, *J. Steroid Biochem. Mol. Biol.* 183 (2018) 80–93.
- [45] T.L. Rizner, T. Thalhammer, C. Ozvegy-Laczka, The importance of steroid uptake and intracrine action in endometrial and ovarian cancers, *Front. Pharmacol.* 8 (2017) 346.
- [46] T. Higuchi, M. Endo, T. Hanamura, T. Gohno, T. Niwa, Y. Yamaguchi, J. Horiguchi, S. Hayashi, Contribution of estrone sulfate to cell proliferation in aromatase inhibitor (AI)-Resistant, hormone receptor-positive breast Cancer, *PLoS One* 11 (2016) e0155844.
- [47] N. Banerjee, N. Miller, C. Allen, R. Bendayan, Expression of membrane transporters and metabolic enzymes involved in estrone-3-sulphate disposition in human breast tumour tissues, *Breast Cancer Res. Treat.* 145 (2014) 647–661.
- [48] R.A. Johnston, T. Rawling, T. Chan, F. Zhou, M. Murray, Selective inhibition of human solute carrier transporters by multikinase inhibitors, *Drug Metab. Dispos.* 42 (2014) 1851–1857.

3.3 A Ligand-based Computational Drug Repurposing Pipeline Using KNIME and Programmatic Data Access: Case Studies for Rare Diseases and COVID-19

TUERKOVA, Alzbeta*; ZDRAZIL, Barbara*. *Journal of Cheminformatics*, **2020**, 12.1: 1-20.

* *Corresponding authors: alzbeta.tuerkova@univie.ac.at; barbara.zdrazil@univie.ac.at*

A. Tuerkova and B. Zdrazil conceptualized and designed the study. A. Tuerkova generated the KNIME workflows, performed the data integration, processing and analyses. B. Zdrazil provided advice. The manuscript was written through contributions of both authors. Both authors read and approved the final manuscript.

The Supplementary Information can be found in Part V.

The following article is reprinted from:

Tuerkova, A., Zdrazil, B. (2020). A ligand-based computational drug repurposing pipeline using KNIME and Programmatic Data Access: case studies for rare diseases and COVID-19. *Journal of cheminformatics*, 12(1), 1-20.

EDUCATIONAL

Open Access



A ligand-based computational drug repurposing pipeline using KNIME and Programmatic Data Access: case studies for rare diseases and COVID-19

Alzbeta Tuerkova* and Barbara Zdrazil*

Abstract

Biomedical information mining is increasingly recognized as a promising technique to accelerate drug discovery and development. Especially, integrative approaches which mine data from several (open) data sources have become more attractive with the increasing possibilities to programmatically access data through Application Programming Interfaces (APIs). The use of open data in conjunction with free, platform-independent analytic tools provides the additional advantage of flexibility, re-usability, and transparency. Here, we present a strategy for performing ligand-based in silico drug repurposing with the analytics platform KNIME. We demonstrate the usefulness of the developed workflow on the basis of two different use cases: a rare disease (here: Glucose Transporter Type 1 (GLUT-1) deficiency), and a new disease (here: COVID 19). The workflow includes a targeted download of data through web services, data curation, detection of enriched structural patterns, as well as substructure searches in DrugBank and a recently deposited data set of antiviral drugs provided by Chemical Abstracts Service. Developed workflows, tutorials with detailed step-by-step instructions, and the information gained by the analysis of data for GLUT-1 deficiency syndrome and COVID-19 are made freely available to the scientific community. The provided framework can be reused by researchers for other in silico drug repurposing projects, and it should serve as a valuable teaching resource for conveying integrative data mining strategies.

Keywords: Drug repurposing, Data integration, Data mining, Data access, Application programming interface, Substructure search, Rare disease, KNIME workflow, COVID-19, SARS-CoV-2, GLUT-1 deficiency syndrome, ChEMBL, Open targets platform, DrugBank, PDB, UniProtKB, Guide-to-pharmacology, PubChem

Background

Computer-aided mining of biomedical data is an emerging field in cheminformatics and drug design which has reshaped current drug development [1–3]. Open access to various life-science repositories, such as ChEMBL [4], PubChem [5], UniProt [6], or DrugBank [7], provides a

competitive advantage when using data-driven drug discovery approaches as opposed to non-integrative approaches [8]. Furthermore, many databases enable programmatic access of the stored data through an Application Programming Interface (API). Consequently, it is of importance to find appropriate tools to analyze gathered data in an automated way. The Konstanz Integration Miner (KNIME) is an open-source data pipelining and analytics platform which enables the creation of (semi)automated workflows to process, transform, analyze, and visualize the data as well as the generation and

*Correspondence: alzbeta.tuerkova@univie.ac.at; barbara.zdrazil@univie.ac.at
Department of Pharmaceutical Chemistry, Division of Drug Design and Medicinal Chemistry, University of Vienna, Althanstraße 14, 1090 Vienna, Austria



© The Author(s) 2020. This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

deployment of approximative mathematical models [9]. In the recent past, the KNIME community has released a plethora of cheminformatics extensions, such as the RDKit [10], Chemistry Development Kit (CDK) [11], Indigo [12], or Vernalis [13] toolkits.

Large-scale data fusion supplied with cheminformatics data analyses can uncover underlying patterns within the data and can pave the way for the development of novel medicine. Such a strategy can be leveraged for drug repurposing (also known as drug repositioning) strategies, in which a re-evaluation of an already approved drug can lead to a treatment for another disease [14]. This approach is particularly useful to, e.g., discover a cure for orphan diseases [15], or to find drug candidates that are worth further investigations for an ongoing pandemic, such as Coronavirus disease 2019 (COVID-19).

With the rapid increase of the availability of biomedical data in the open domain, computational drug repurposing approaches now strongly benefit from interconnecting different types of data entities, including genes, tissue expression data, targets, drugs, phenotypes, and diseases, to deliver an indication about a drugs' mode-of-action. Method-wise, computational drug repurposing techniques range from data (text) mining, and different machine learning approaches, to network analyses and structure-based approaches [16]. For example, Li et al. combined data originating from text mining with protein interaction networks to develop a drug-target connectivity map for a certain disease [17]. Machine learning methods used in drug repurposing strategies include, *inter alia*, support vector machines [18], classification models [19], and currently also deep neural networks [20] to predict drug-disease relationships. Network analyses enable to model complex functional similarities between various biological entities, such as drugs, genes, proteins, or entire protein families [21]. An orthogonal approach to ligand-based strategies, is to perform structure-based virtual screening by using a consensus inverse docking strategy, as demonstrated by Wang et al. [22].

Semi-automated drug repositioning pipelines are uniting the advantages of computational workflows (e.g., provided by using the open source tool KNIME) with the availability of big open data sources that can be accessed programmatically. They make access to data resources easier and thus lower the barriers for effective data usage for non-data scientists. Also, their usage shortens the time period from data collection to the identification of hidden relationships in the data. In addition, such workflows are easily reproducible, and can be adapted according to individual project needs [23].

In this study, we are providing a general strategy and a step-by-step tutorial for automated data access and data integration from multiple open data sources (which are

providing an API), along with data processing and cheminformatics data analysis by using the pipelining tool KNIME. Individual operations, such as the specification and execution of API requests, extraction of properties through JSON/XPath queries, structural data standardization, identification of enriched structural fragments, and substructure searches in external data sources, are thoroughly described and demonstrated herein.

Protein and ligand information related to GLUT-1 deficiency syndrome and to COVID-19 have been chosen as individual use cases to demonstrate the usefulness of the approach. GLUT-1 deficiency syndrome is a rare disease caused by genetic variation of glucose transporter member 1 (SLC2A1), which leads to impaired transport of glucose (<https://ghr.nlm.nih.gov/condition/glut1-deficiency-syndrome>). For COVID-19, to date only data for suggested targets can be used (with relatively little knowledge about the strength of the target-disease associations). Just recently, about 66 druggable protein targets with potential interest for SARS-CoV-2 treatment have been reported [24].

The Open Targets Platform integrates public domain data to enable target identification and prioritization by providing association scores between targets and diseases [25]. Targets represented in the Open Targets Platform can be genes, transcripts or proteins integrated through the Ensembl gene ID (<https://www.ensembl.org/index.html>).

In this study, we used highly scored proteins from the Open Targets Platform for the diseases under investigation. In case of COVID-19, protein targets listed in the UniProtKB pre-release web page (available at https://covid-19.uniprot.org/uniprotkb?query=*) were additionally used as a starting point.

API calls were specified to map UniProt IDs of the targets to available structural data in the Protein Data Bank (PDB) [26]. Ligands co-resolved with a protein structure were extracted as separate entities. For sake of data augmentation, ligand bioactivity measurements (such as Ki, IC50, or Km end-points) for the protein targets under study were retrieved from ChEMBL [4], PubChem [5], and Guide-to-Pharmacology (IUPHAR) [27]. After data cleaning and chemical structure standardization, Bemis-Murcko scaffolds [28] were extracted from the ligands in the data set and grouped by similarity into structural queries for subsequent substructure searches in DrugBank [7] and the CAS COVID-19 antiviral candidate compounds data set (available upon request at <https://www.cas.org/covid-19-antiviral-compounds-datas-et>). These searches led to the identification of structurally analogous compounds which could potentially show similar pharmacological action at targets associated with GLUT-1 deficiency syndrome or COVID-19. A list of

identified hits, is provided as an output of the workflow. A schematic overview of the whole data-driven drug repurposing workflow is depicted in Fig. 1.

Taken together, the developed data mining pipeline is a useful resource for any in silico drug repurposing project and is exemplified on the basis of a drug repositioning strategy for GLUT-1 deficiency syndrome and the Coronavirus Disease 2019 (COVID-19). The step-by-step instructions allow for an easy implementation for other drug discovery projects along these lines and they shall give especially guidance to students or researchers new to the field of data-driven drug discovery. All workflows can be accessed via an open GitHub Repository (available at <https://github.com/AlzbetaTuerkova/Drug-Repurposing-in-KNIME>).

Methods

As the drug-repurposing strategy applied here is mainly conceived for educational purposes, we introduce a step-by-step tutorial for a guided development of a KNIME workflow. In addition, the workflow developed herein is fully versatile and it can thus be reproduced for other diseases of interest. Basic knowledge of configuration and execution of standard nodes (e.g., the 'Row Filter' node, the 'GroupBy' node, the 'Joiner' node, the 'Pivoting' node), import of external data sets into a KNIME workflow (e.g., the 'SDF reader' node, the 'File Reader' node), handling different structural formats, as well as working with specific data types in KNIME, is expected here as a prerequisite.

When integrating data from diverse sources, it becomes beneficial to query databases programmatically, i.e., without the need of laborious manual data download and data integration. UniProtKB and other databases used in this example enable targeted access of the stored data through an Application Programming Interface (API).

In the KNIME workflow discussed herein, a triad of KNIME nodes is consecutively executed (1) to specify the API request (via the 'String Manipulation' node), (2) to retrieve data from web services (via the 'GET request' node), and (3) to perform XPath/JSON queries to extract useful properties for a given protein (via the 'XPath' or 'JSONPath' node, respectively). The corresponding part of the KNIME workflow is depicted in Fig. 2.

1. Step: Mapping target identifiers of the Open Targets Platform to UniProt

The workflow discussed herein, allows two different sorts of input: (1) Automated retrieval of targets associated

with a certain disease via the Open Targets Platform and (2) importing an external data set with a list of protein targets.

In option (1), the disease identifiers from the Open Targets Platform for GLUT-1 deficiency syndrome (Orphanet_71277) and COVID-19 (MONDO_0100096) have been specified as input in the 'Table Creator' node. Next, an API request to fetch disease records was created using the 'String Manipulation' node. The join() function in the 'String Manipulation' node is used and a corresponding Open Targets Platform disease ID is forwarded to the string as a variable (\$disease_id\$ column). Additional parameters used in this API request are the maximum number of associated drug targets ('size', here set to 10,000), and the association score, which enables to prioritize the drug targets on basis of their available evidence for a disease ('scorevalue_min', here set to 0.99): join("https://platform-api.opentargets.io/v3/platform/public/association/filter?disease=", \$disease_id\$, "&size=10000&scorevalue_min=0.99").

As an output of the 'String Manipulation' node, a column with the respective API requests is appended to the output table, such as: https://platform-api.opentargets.io/v3/platform/public/association/filter?disease=EFO_0001360&size=10000&scorevalue_min=0.99.

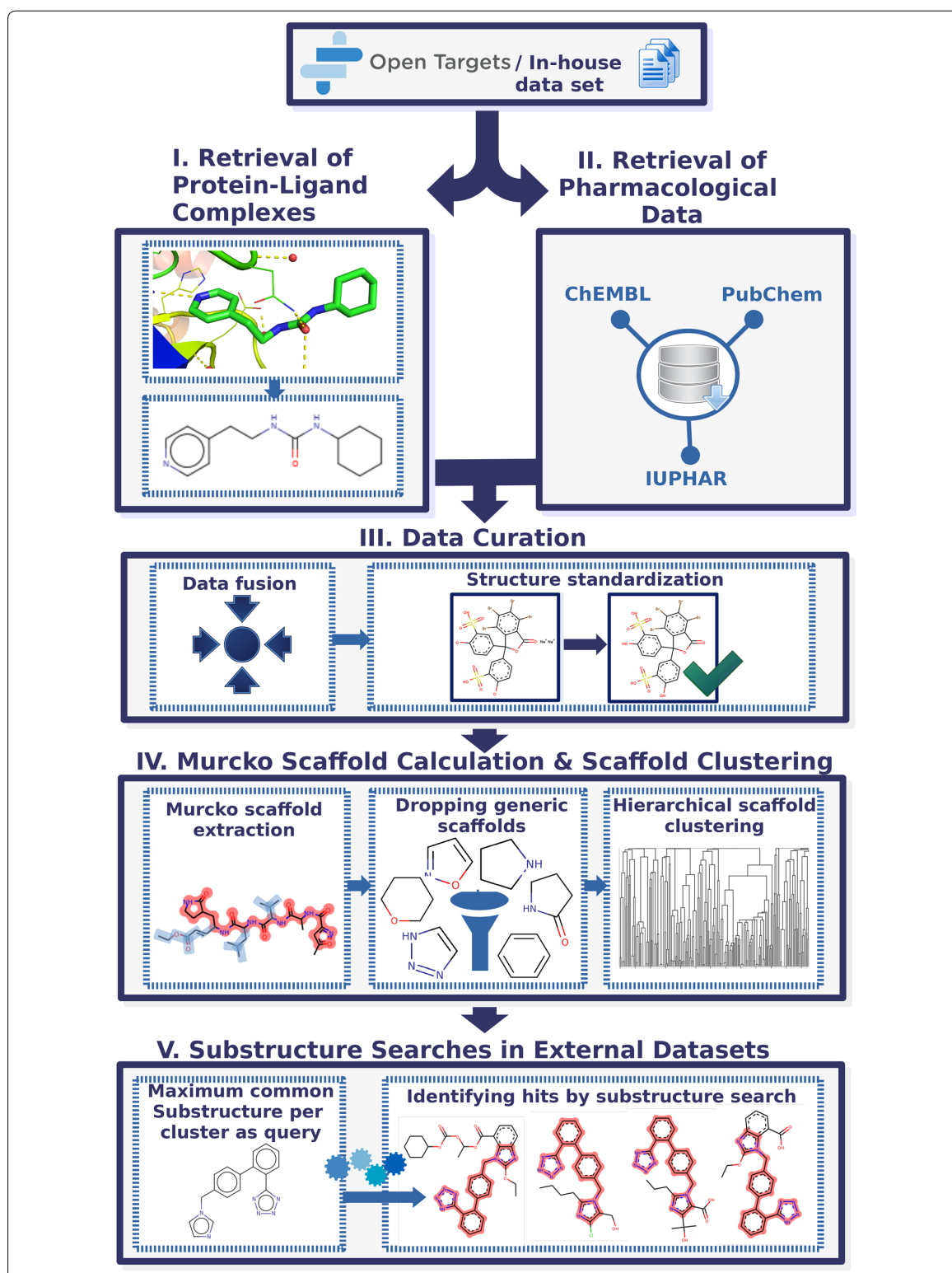
By executing the API request (via the 'GET Request' node), a JSON file is downloaded from the Open Targets Platform and appended to the output table as a separate column. Additionally, columns reporting the content type (here 'application/json'), and the HTTP status code are appended (Fig. 3). There exist five classes of HTTP status codes: (1) Informational responses (100–199), (2) Successful responses (200–299), (3) Redirects (300–399), (4) Client errors (400–499), and (5) Server errors (500–599). The information provided about the status of the request can be used to filter out any useless data entries. It is recommended to increase the timeout in the 'GET Request' configuration as the default specification (2 s) is usually insufficient to receive all requested data.

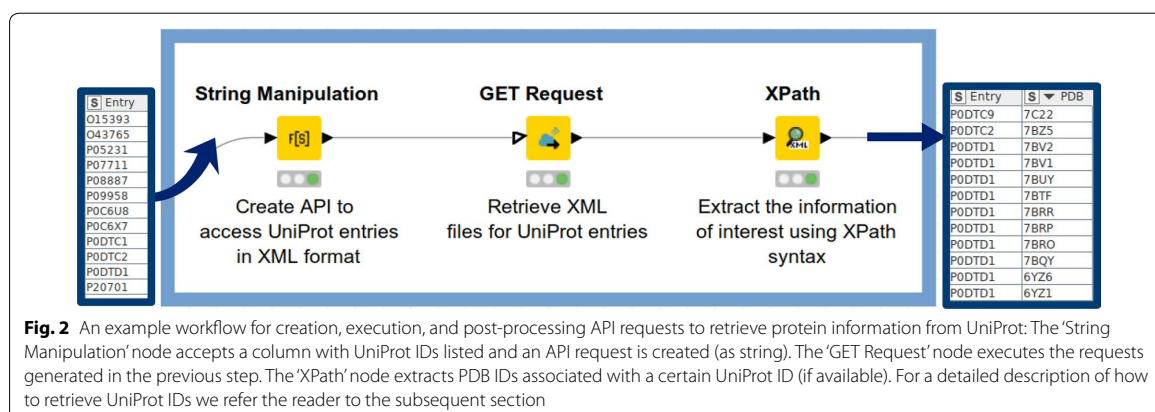
Subsequently, the 'JSON Path' node is used to extract the information of interest on the basis of querying different JSON objects. The 'JSON Path' node enables to create JSON Path queries in both dot-notation and bracket-notation (depending on how the properties of an object are specified in the syntax). Here, the bracket notation is applied to extract target identifiers, target names, and gene symbol by using the following JSON paths:

```
$['data'][*]['target']['id']
$['data'][*]['target']['gene_info']['name']
```

(See figure on next page.)

Fig. 1 Schematic overview of the data-driven drug-repurposing workflow





Row ID	S disease_id	I Status	S Content type	JSON body
Row0	EFO_0001360	200	application/json	{ "from": 0, "took": 49, "data_version": "20.06", "query": { "sort": ["harmonic-sum.overall"], "search": null, "rna_expression_level": 0, "protein_expression_tissue": [], "scorevalue_types": ["overall"], "datatype": [], "fields": null, "format": "json". } }

Fig. 3 An example of the output table generated after the execution of the 'GET Request' node: Status, content type, and JSON file are appended to the table as separate columns

`['data']['target']['gene_info']['symbol']`

Output values are appended to separate cells as a collection data type. The 'Ungroup' node is subsequently used to transform collections of values into individual rows.

Next, cross-references for all human target entries in the Open Targets Platform can be fetched via the UniProt web services. Here, a corresponding API request was executed to retrieve the mappings for targets (UniProt target IDs are mapped to Open Targets Platform target IDs): [https://www.uniprot.org/uniprot/?query=organism:9606+AND+database:OpenTargets&format=xls&columns=id,database\(OpenTargets\),reviewed](https://www.uniprot.org/uniprot/?query=organism:9606+AND+database:OpenTargets&format=xls&columns=id,database(OpenTargets),reviewed).

Due to the potential workflow overload, we recommend to download a mapping file (XLS format) and

forward it to the workflow via the 'File Reader' node and later join the two data sets via the 'Joiner' node.

Option (2) is to use a user-specified list of UniProt IDs in a data table format. In this contribution, this step is exemplified by the use case for proteins that are listed to be of potential interest for treating COVID-19 (53 entries available at https://covid-19.uniprot.org/uniprotkb?query=*). The CSV/TSV file is read in by a 'File Reader' node.

2. Step: Retrieving protein–ligand structural data from the Protein Data Bank

UniProt IDs for targets of interest were used to retrieve available protein–ligand complexes stored in the Protein Data Bank (PDB) [29].

Based on the same strategy as in step 1, a column with the respective API requests is appended to the output

table. An example for such an API request looks like this: <https://www.uniprot.org/uniprot/F8W8F0.xml>.

When executing the workflow with COVID-19 pre-release data provided by UniProtKB, the API request has to be adopted in the following manner: <https://www.ebi.ac.uk/uniprot/api/covid-19/uniprotkb/accession/O15393.xml>.

By executing the API requests (via the 'GET Request' node), the XML file is downloaded from UniProt and appended to the output table as XML cell. Similar to the 'JSON Path' described in the previous step, the 'XPath' node (XPath 1.0 version) is used to extract the information of interest on the basis of querying different XML elements and the associated XML attributes. One can define an XPath query within the 'XPath' node from scratch. Another way is to perform a double-click on a specific section in the XML-Cell Preview table and the XPath query is generated automatically. The XPath query below is used to retrieve all available PDB IDs for a given UniProt ID:

```
/dns:uniprot/dns:entry/
dns:dbReference[@type='PDB']/@id
```

The 'dns' prefix corresponds to the namespace used in the XPath query. Here, <http://uniprot.org/uniprot> is used as a namespace. Namespaces are defined automatically and are listed in the node configuration.

The example XPath query shows that PDB IDs are integrated within the <dbReference>XML element. However, UniProt entries consist of multiple <dbReference> elements which are pointing to different data sources, such as PubMed, GO, InterPro, Pfam, or PDB:

```
<dbReference type="PubMed"
id="12730500">
  <dbReference type="GO" id="GO:0039579">
    <dbReference type="InterPro"
id="IPR036333">
      <dbReference type="Pfam" id="PF06478">
        <dbReference type="PDB" id="6NUR">
```

A key task is to query data from XML elements which do possess the 'PDB' attribute exclusively. The '@' character is used to specify certain XML attributes in the XPath query. Therefore, dbReference[@type='PDB'] is forwarded to the XPath query to get all PDB IDs by querying the @id attribute.

Due to the possible synchronization delay of UniProt releases with other cross-referenced databases, an additional alternative approach has been used to fetch PDB data. Specifically, PDBe graph APIs were used for this purpose. The PDB entities are returned in JSON format by default. Below an example is provided for a request to fetch protein structures for the ACE2 receptor (UniProt ID: Q9BYF1) via PDBe graph APIs: https://www.ebi.ac.uk/pdbe/graph-api/mappings/best_structures/Q9BYF1.

Similar to the 'XPath' node for processing XML documents, KNIME also provides the 'JSON Path' node which is used to process JSON data. The 'JSON Path' node enables to create JSON Path queries in both dot notation and bracket notation (depending on how the properties of an object are specified in the syntax). In the discussed KNIME workflow herein, the bracket notation is applied to extract the PDB IDs:

```
$..[*].['pdb_id']
```

Since the data are listed as a collection column type, the 'JSON Path' node is followed by the 'UnGroup' node to list multiple PDB IDs per protein target into separate rows. After concatenating data ('Concatenate' node) retrieved from PDBe graph APIs, duplicates for a respective target were removed by grouping the data by target UniProt ID and PDB IDs ('GroupBy' node). The 'PDB ID' column is used to create the Uniform Resource Locator (URL) path to extract different properties by using the same strategy as shown in Fig. 2. An example of such URL is given below: <https://files.rcsb.org/view/2VYI.pdb>.

The 'PDB Loader' and the 'PDB Property Extractor' nodes are available from the KNIME repository (created by Vernalis, Cambridge, UK) to facilitate analysis of PDB data in KNIME (Fig. 4). These nodes were employed in order to explore properties of the PDB files, such as the experimental method used (X-ray diffraction, solution NMR, cryo-EM, theoretical models), the number of stored models, the resolution of structures, Space groups, R-factor, and so on.

Next, the available PDB structures were examined for their availability of co-resolved ligands. Ligand information (in JSON format) can be received through the RCSB PDB RESTful Web services by creating the following request: <https://data.rcsb.org/rest/v1/core/entry/2VYI>.

Following JSON Path node is used to retrieve a collection of bound ligands, if available:

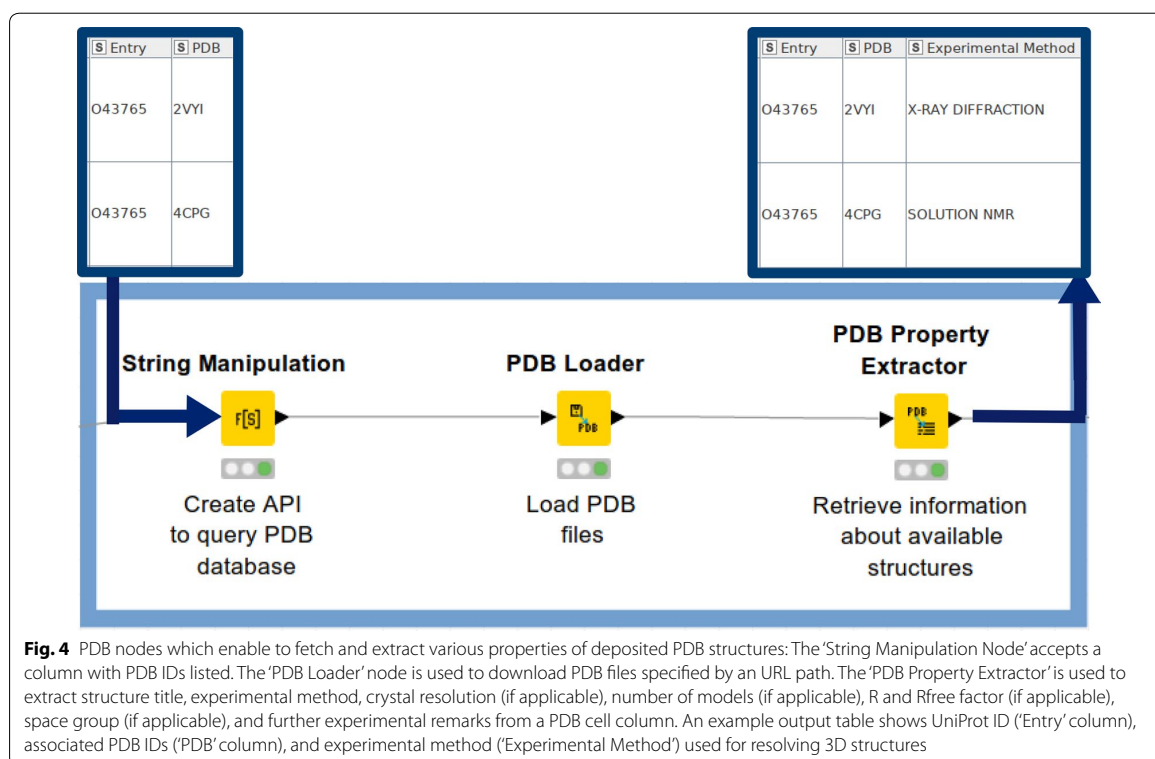
```
$['rcsb_entry_info']
['nonpolymer_bound_components']
```

Available ligands are listed using their shortcuts (e.g., BME, NAG, XU3). An API request is subsequently created and executed to fetch ligand information (in JSON format): <https://data.rcsb.org/rest/v1/core/chemcomp/NAG>.

The following JSON path node is used to retrieve the SMILES code for a specific ligand:

```
$['rcsb_chem_comp_descriptor']
['smiles']
```

Subsequently, PDB entries without a co-resolved ligand are filtered out (by applying the 'RowFilter' node). The 'GroupBy' node is used to keep unique ligand structures per protein target (grouping by UniProt ID and smiles



string). This procedure might also retrieve salts, solvents, and/or co-crystallizing compounds, as they are identified as 'ligands' in PDB. Although the salts and unconnected fragments are stripped during the structure standardization procedure (as described in Sect. 3), it is generally advisable to cross-check the output table to eliminate retained co-crystallizing agents (e.g., isonicotinamide).

3. Step: Fetching ligand bioactivity data from open bioactivity data sources via programmatic data access

Orthogonal to fetching ligand data for drug targets of interest from their protein structures, ligands and their experimental bioactivity measurements can also be collected from open pharmacological databases. In this example, data is retrieved from ChEMBL (version 26) [4, 30], PubChem [5], and IUPHAR (also known as Guide-to-Pharmacology, version 2020.2) [27] by using the respective web services via the 'Get Request' and 'XPath' nodes in KNIME. Automated data access can be achieved by using predefined identifiers for targets, ligands (such as ligand structure, available bioactivities, or molecule names), biochemical assays, and so on.

The KNIME workflow for fetching ChEMBL data allows to map UniProt IDs of protein targets to target ChEMBL IDs and subsequent retrieval of ligand

bioactivities and their respective structural information (here: canonical smiles), document ChEMBL IDs, and Pubmed IDs for the primary publication. A major challenge is the limited number of bioactivities (up to 1000 bioactivities) that are being fetched per single call. The KNIME workflow therefore has to be adopted to fetch all available data without manual intervention. The metanode that does the trick (termed 'Get bioactivities per target') works as follows:

1. A single XML file per target is downloaded and the number of bioactivities integrated within the <total_count> XML element is extracted.
2. The number of iterations needed to fetch all available bioactivities per target is calculated by dividing the number of bioactivities by 1000 and then rounding the result up (ceil() function in the 'Math Formula' node).
3. A recursive loop is used in order to process protein targets one-by-one.
4. A nested loop is used within a recursive loop where the API call is modified in a way that it dynamically changes the 'offset' parameter per each iteration. The 'offset' parameter determines the number of bioactivities that should be skipped before downloading the

next portion of bioactivities for a given target. After the loop ends, all information needed is extracted from the collected XML files by the 'XPath' node.

This procedure shall be illustrated on basis of an example: There are 2410 bioactivities for protein X available. Thus, three iterations are needed to fetch all data available for protein X if offset is set to 1000. Within each iteration, a column is appended to the table containing the API call with the corresponding offset parameter, i.e.

https://www.ebi.ac.uk/chembl/api/data/activity?target_chembl_id=CHEMBL5118&limit=1000&offset=0 (iteration#1).

https://www.ebi.ac.uk/chembl/api/data/activity?target_chembl_id=CHEMBL5118&limit=1000&offset=1000 (iteration#2).

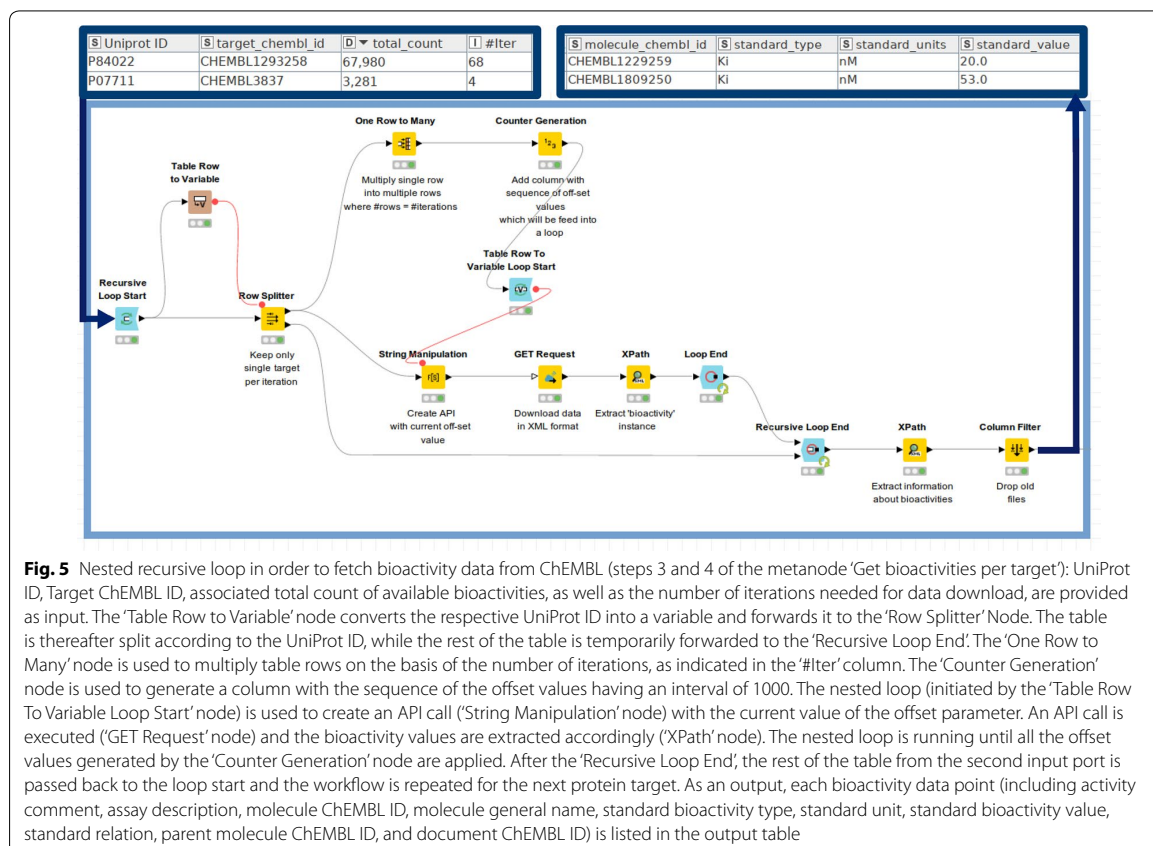
https://www.ebi.ac.uk/chembl/api/data/activity?target_chembl_id=CHEMBL5118&limit=1000&offset=2000 (iteration#3).

At the end of the loop, 2410 bioactivities have been collected for protein X and these are processed as indicated in the description above.

Step 3 and 4 from the metanode 'Get bioactivities per target', as described above, are visually depicted in Fig. 5.

In case of PubChem, UniProt IDs are mapped to 'PubChem Assay IDs' (AID) in the first step. Further, AIDs are mapped to available compounds by 'PubChem Compound ID' (CID), including bioactivity measurements and associated PubMed IDs. Compound structures and names are retrieved in the next step. In some cases, compound names in PubChem are included in the form of molecule ChEMBL IDs. If this condition is true, the ChEMBL is additionally queried to download a compound name, if available.

In order to query IUPHAR data, the UniProt ID is mapped to the IUPHAR target ID. API calls have a specific syntax for accessing substrates, e.g.: <http://www.guidetopharmacology.org/services/targets/2421/substrates> and for accessing inhibitors, e.g.: <http://www.guidetopharmacology.org/services/targets/2421/interactions>,



where “2421” is an identifier for a specific target ID. Compound ID, PubMed ID, affinity, affinity type (corresponding to a certain end-point), and action (corresponding to a certain activity annotation) were retrieved by using the ‘JSON Path’ node. Retrieval of the ligand structural format is done by an additional API call on basis of the respective ligand ID.

Bioactivity values are converted to their negative logarithmic representation and binary labels (‘1’ for active and ‘0’ for inactive) are assigned on the basis of an activity cut-off. In this example, all compounds possessing a negative logarithmic value greater than 9 (i.e., < 1 nM) were labeled as ‘1’, while the rest was labeled as ‘0’.

After merging the output tables from ChEMBL, PubChem, and IUPHAR, the data is grouped to keep unique ligands per target and median values for binary activity labels (by using the ‘GroupBy’ node). In addition, only active ligands per target (label ‘1’) are kept and the final table is concatenated with ligand structures from PDB entries.

A prerequisite for merging ligand data from diverse sources is standardization of the molecular structures. A similar curation strategy like the one published by Gadaleta et al. [31] was applied:

1. Characters encoding stereoisomerism in SMILES format (@; \; /) are removed by using the ‘String Replacer’ node since for subsequent operations this information is not needed.
2. Salts are stripped by using the ‘RDkit Salt Stripper’ node. (This node works with pre-defined sets of different salts/salt mixtures by default. If requested, additional salt definitions can be forwarded to the node.)
3. Salt components are listed in the output table using the ‘Connectivity’ node (CDK plugin) followed by the ‘Split Collection Column’ node
4. The ‘RDKit Structure Normalizer’ node neutralizes charges and checks for atomic clashes, etc. Additional criteria for compound quality check can be adjusted in the ‘Advanced’ section of the node configuration.
5. The ‘Element Filter’ node keeps compounds containing the following elements only: H,C,N,O,F,Br,I,Cl,P,S).
6. InChI, InChIKey, and Canonical smiles formats are finally created from the standardized compounds.

Steps 2–4 are visually depicted in Fig. 6.

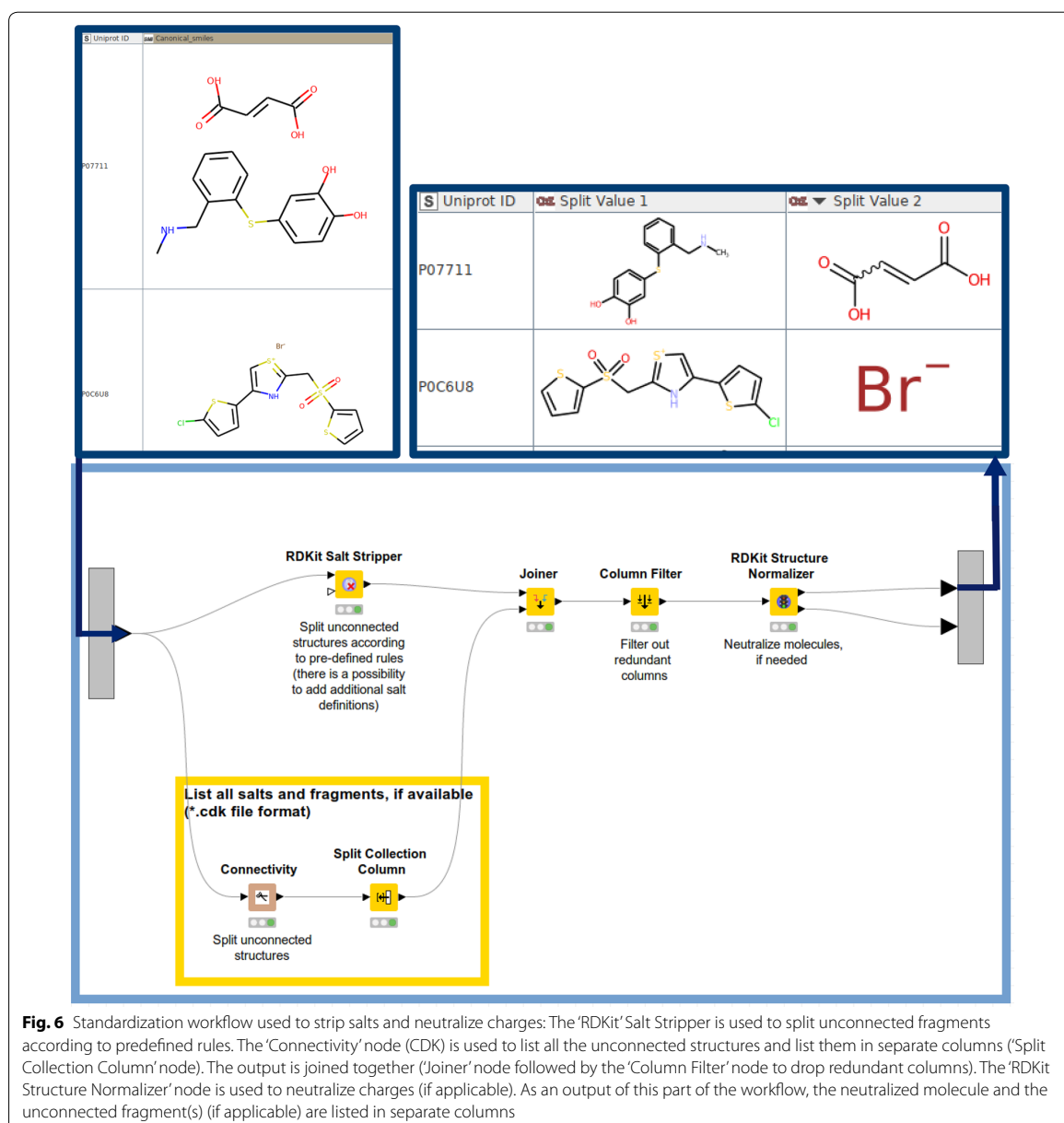
4. Step: Substructure searches to identify potentially interesting compounds for drug repurposing

Finally, the merged data sets are used to generate structural queries in SMARTS format in order to perform substructure searches in DrugBank (version 5.1.6, approx. 10,000 compounds, structures in SDF format are publicly available at <https://www.drugbank.ca/releases/latest#structures>) and in the COVID-19 antiviral candidate compound data set provided by the Chemical Abstracts Service (approx. 50,000 compounds, available upon request at <https://www.cas.org/covid-19-antiviral-compounds-dataset>).

Bemis-Murcko scaffolds are extracted (‘RDKit Find Murcko Scaffolds’ node) in order to get a quick overview of the structural diversity of the curated data set. Scaffolds possessing too generic structures (i.e., a single aromatic ring) can be filtered out (by using the ‘RDKit Descriptors Calculator’ node in conjunction with the ‘Row Filter’ node) and remaining ones can be explored with respect to their structural similarity in the context of a certain target. This step is done by (1) calculating molecular distances (the ‘MoSS MCSS Molecule Similarity’ node), (2) hierarchical clustering (the ‘Hierarchical Clustering [DistMatrix]’ node), and (3) assigning a threshold (here: distance threshold=0.5) for cluster assignment (the ‘Hierarchical Cluster Assigner’ node). The ‘MoSS MCSS Molecule Similarity’ node is used to calculate similarities between Murcko scaffolds by taking the size of their Maximum Common Substructure (MCS) as a similarity metric. Molecular similarities are then evaluated on the basis of a distance matrix. The respective part of the workflow is depicted in Fig. 7.

Next, looping over distinct clusters of associated Bemis-Murcko scaffolds for a respective target is done in order to create a maximum common substructure (the ‘RDKit MCS’ node) from all associated scaffolds belonging to a respective cluster. Recursive loops are extensions to regular loops which can be used in conjunction with a ‘Row Splitter’ node to separate the current row from the rest of the table. After termination of the current iteration, the rest of the table is forwarded to the loop start and the next row is used for the subsequent iteration (see Fig. 8). Generated substructures for a certain target are appended to the output table in SMARTS format.

For the substructure searches in DrugBank and the CAS data set loops are being used as well (Fig. 9). The ‘Table Row To Variable Loop Start’ forwards each substructure as a query to the ‘RDKit Substructure Filter’ node as a flow variable which then examines whether a particular substructure is contained in the data sets from DrugBank or CAS. Extracted compounds are being forwarded to the ‘RDKit molecule highlighting’ node which

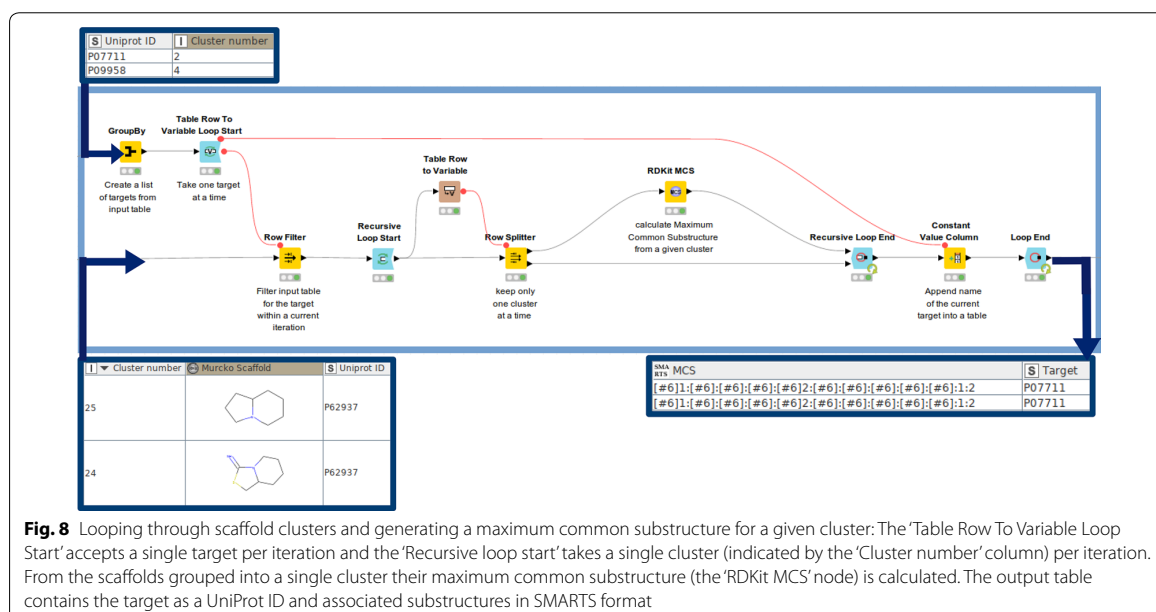
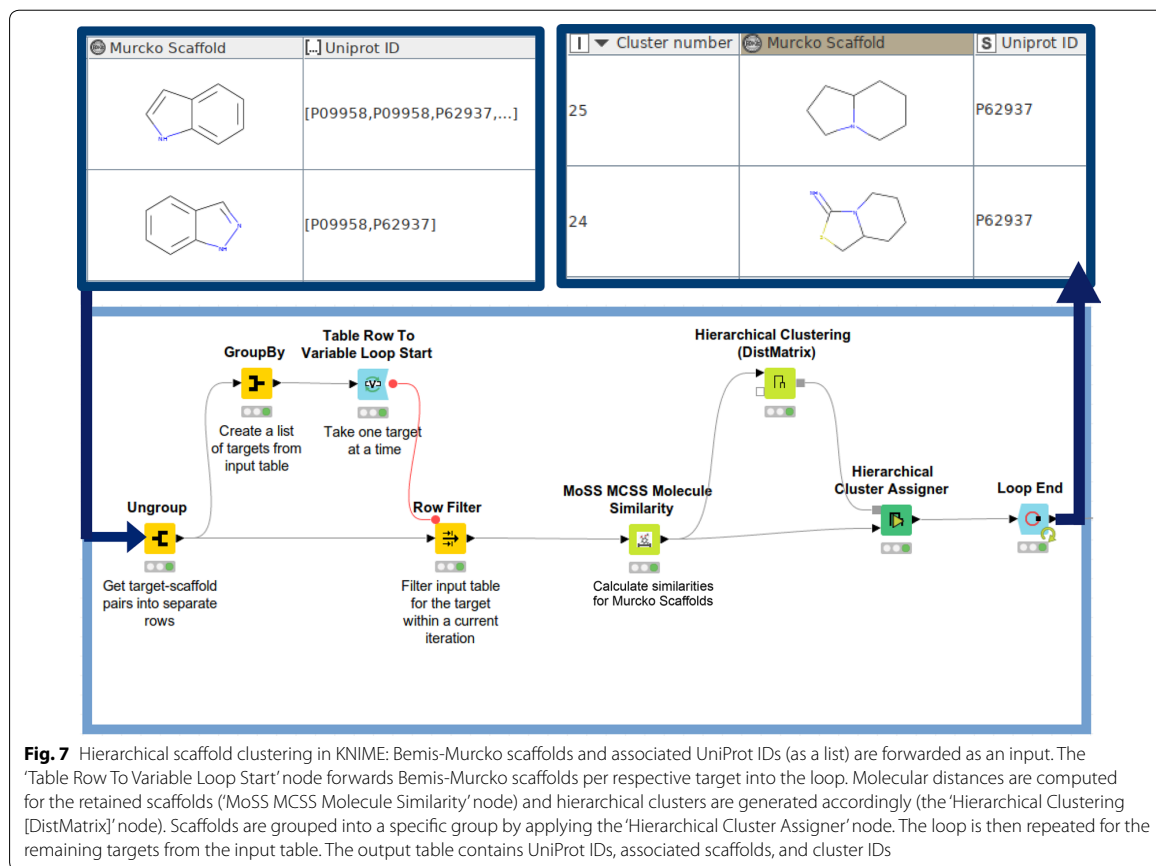


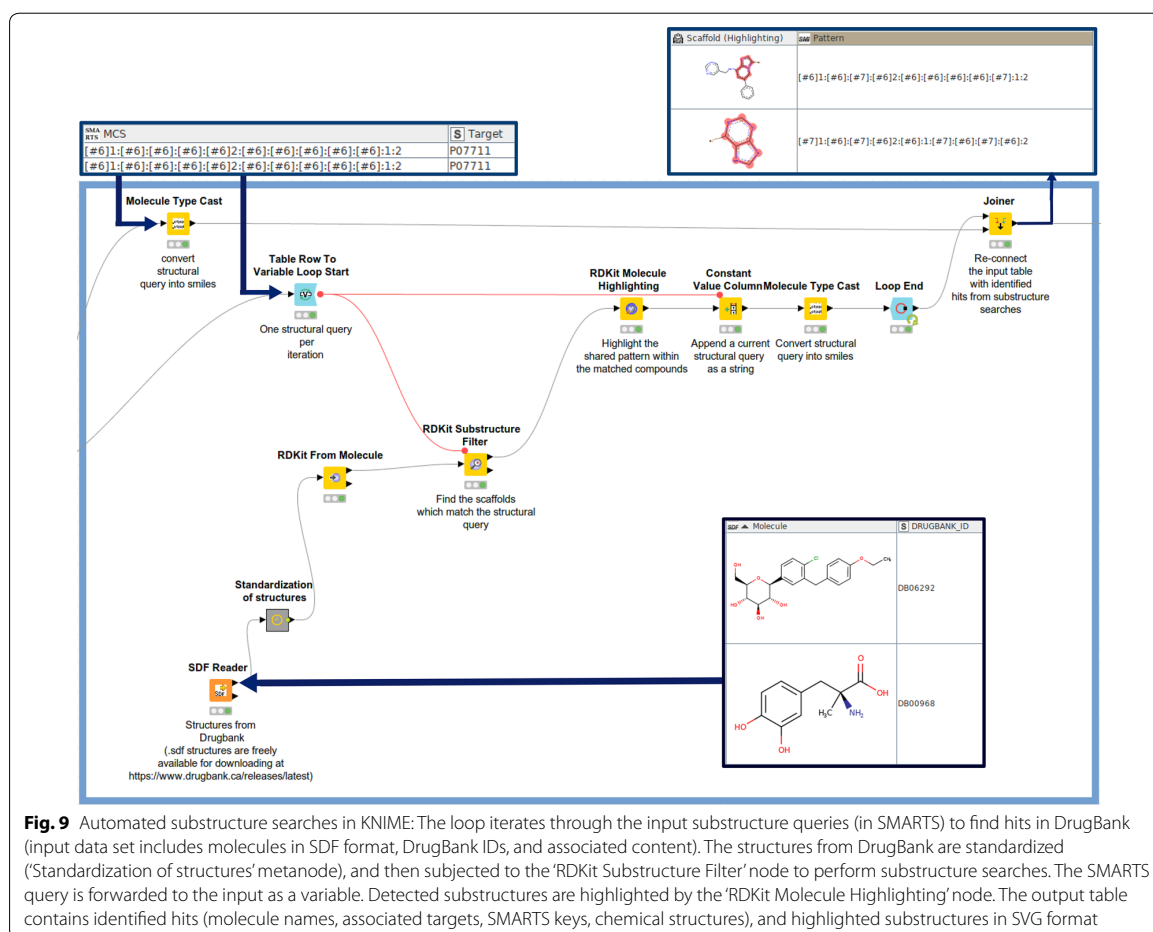
visualizes the highlighted substructure within the respective compounds.

Software

KNIME workflows were built in KNIME version 4.1.2. The KNIME workflows are freely available from GitHub

(<https://github.com/AlzbetaTuerkova/Drug-Repurposing-in-KNIME>). The published workflow can be either used as a single pipeline, or as multiple stand-alone workflows (1) to gather data from PDB, (2) to retrieve ligand bioactivities from ChEMBL, PubChem, and IUPHAR, and (3) to perform substructure searches, by providing the needed data input, respectively.





Results and discussion

In this contribution, a semiautomatic KNIME workflow for drug repurposing based on publicly available structural- and bioactivity-ligand data is presented. The pipeline includes automatic mapping of UniProtKB entries and PDB via cross-referencing, programmatic data access via the data sources' web services (exemplified for ChEMBL, PubChem, and IUPHAR), fully automatic data curation (including data integration, chemical data standardization, removal of duplicates, and cut off setting for assigning activity labels), the identification of common structural patterns in SMARTS format, and substructure searches (here in DrugBank and the CAS data set of antiviral drugs) in order to identify interesting compounds for further investigations. The drug repurposing pipeline developed here is showcased by applying it on one rare disease as well as one new disease: Glucose Transporter Type 1 (GLUT-1) deficiency syndrome and COVID-19.

Retrieval of COVID-19 data

The Universal Protein Resource KnowledgeBase (UniProtKB) is a freely accessible database for protein sequence and annotation data. The UniProt ID (e.g., P59596, P59637, P0C6X7) is a protein identifier which can be used to retrieve comprehensive information about a given protein, including protein names and synonyms, taxonomy, function, cellular localization, available three-dimensional structures, as well as cross-references to other databases. Cross-referenced databases include (but are not limited to) sequence databases (e.g. GenBank [32], CCDS [33]), 3D structure databases (e.g., Protein Data Bank [29], ModBase [34], SWISS-MODEL-Workspace [35]), protein-protein interaction databases (e.g., Biogrid [36], IntAct [37], STRING [38]), and chemistry databases (e.g., BindingDB, [39] ChEMBL, [4] DrugBank [7]). In a first instance, content from a pre-release UniProt web page (available at https://covid-19.uniprot.org/uniprotkb?query=*) was used as an input for the data

mining pipeline to gather and analyze data for proteins potentially interesting for the treatment of infections with human SARS-CoV-2 (53 proteins, Additional file 1: Table S1). As seen from Additional file 1: Table S1, available protein templates include 14 SARS-CoV-2, 15 SARS-CoV, and 24 structures with origin *Homo Sapiens*.

Listed UniProt IDs were used to retrieve protein structures stored in PDB (1084 structures, 953 unique structures). From these sources, 151 unique ligands could be extracted, yielding 87 unique Murcko scaffolds. From the orthogonal approach—the automatic gathering of ligand bioactivity data from ChEMBL, PubChem, and IUPHAR via its webservices—3951 unique ligands with (median) activity value < 1 nM were identified (2555 unique Murcko scaffolds).

As an alternative solution for generating a list of targets associated to COVID-19, 55 human protein targets with the association score of at least 0.99 were retrieved from the Open Targets Platform (see Additional file 1: Table S2). Interestingly, the interleukin-6 receptor subunit alpha (UniProt ID P08887) was identified as a sole target which was also listed at the UniProt pre-release web page. Such a different constitution of the input data between the UniProt pre-release webpage and the Open Targets Platform could be explained by the fact that target-disease association scores in Open Targets are based on a cumulative score collecting different sources of evidence (such as genetic associations, somatic mutations, drugs available in ChEMBL, pathways & system biology, RNA expression data, text mining, animal models). However, in the case of COVID-19 to date only association scores for evidence from drugs in ChEMBL and text mining are available, which restricts the highly scored targets to the ones already described in literature. The approach did not allow for prioritization of, e.g., ACE2 receptor, as

its association score possesses a value of only 0.11 in the Open Targets Platform (accessed Sept. 2020). This use case might illustrate the ultimate benefit when combining protein-disease association data from various independent sources.

Listed targets originating from the Open Target platform have become a source of multiple PDB structures (571 structures, 502 unique structures). In total, 85 unique ligands could be extracted, (45 unique Bemis-Murcko scaffolds). By applying integrative mining of bioactivity data from public databases, 3207 unique ligands with (median) activity value < 1 nM were fetched (1710 unique Bemis-Murcko scaffolds).

The highly ranked targets (based on the number of retrieved compounds) from either resource are listed in Table 1.

Analysis of COVID-19 data sets

Numbers of unique compounds per individual COVID-19 drug target that could be fetched from the different data sources are listed in Additional file 1: Tables S3 and S4. In case of targets retrieved from the UniProt pre-release web page (Additional file 1: Table S3), PubChem is the predominant source of ligands (9751 unique compounds). At the other end of the scale, IUPHAR provides 19 unique compounds only. Inspecting the origin of data for the respective protein targets, it becomes apparent that the ligand information for human SARS-CoV-2 solely originates from PDB structures (see Additional file 1: Table S3, entries ending with “_SARS2”). Notably, the majority of structures for SARS-CoV-2—such as PDB IDs 6W4B [40], 6Y2E, or 6Y2G for replicase polyprotein 1a [41] were refined via molecular replacement based on the homology to SARS-COV. It therefore seems

Table 1 Number of compounds available from different data sources (PDB, ChEMBL, IUPHAR, PubChem) for the five top-ranked protein targets retrieved from both the UniProt pre-release web page and the Open Targets Platform

Target shortcut	Target source	PDB	ChEMBL	IUPHAR	PubChem	# Unique active compounds
PPIA_HUMAN	UniProt pre-release	57	2	1	3123	3183
CATL1_HUMAN	UniProt pre-release	25	38	4	946	1003
ITAL_HUMAN	UniProt pre-release	13	94	2	550	564
FURIN_HUMAN	UniProt pre-release	4	10	1	448	463
R1AB_CVHSA	UniProt pre-release	37	187	0	47	227
GABRG2_HUMAN	Open Targets Platform	152	0	677	2	831
MMP13_HUMAN	Open Targets Platform	319	3	80	28	430
GABRB1_HUMAN	Open Targets Platform	121	0	166	0	287
DPP4_HUMAN	Open Targets Platform	140	2	29	14	185
GABRA1_HUMAN	Open Targets Platform	3	4	172	2	181

to be beneficial to integrate data from diverse sources, especially including PDB as a source for most up-to-date compound information.

Across all data sources, the largest number of ligand bioactivity measurements (in case of targets from UniProt pre-release web page) was gathered for human peptidyl-prolyl cis-trans isomerase A (UniProt ID P62937; 3183 unique compounds), followed by human procathepsin L (UniProt ID P07711; 1003 unique compounds), human integrin alpha-L (UniProt ID P20701; 564 unique compounds), human furin (UniProt ID P09958; 463 unique compounds), SARS replicase polyprotein 1ab (UniProt ID P0C6X7; 227 unique compounds), human angiotensin-converting enzyme 2 (ACE2; UniProt ID Q9BYF1; 172 unique compounds), SARS replicase polyprotein 1a (UniProt ID P0C6U8; 141 unique compounds), and human mothers against decapentaplegic homolog 3 (UniProt ID P84022; 71 unique compounds). For other potential COVID-19 targets, only a neglectable number of compounds was retrieved. The ACE2 receptor is considered a relevant therapeutic target due to its interaction with spike glycoprotein of coronaviruses when entering host cells [42]. Replicase polyproteins 1a and 1ab are attractive targets to treat COVID-19 given their crucial role in replication and transcription of viral RNAs [43]. A current study has suggested a potential role of integrins as alternative receptors for SARS-CoV-2, as the spike glycoprotein contains an integrin-binding motif [44].

From the data retrieved from the Open Targets Platform, a complete list of unique compounds per individual target is included in Additional file 1: Table S4. The highest number of unique compounds was retrieved for gamma-aminobutyric acid type A receptor subunit gamma 2 (UniProt ID P18507; 831 unique compounds). In general, different gamma-aminobutyric receptor subunits have been ranked high in terms of the number of gathered compounds.

COVID-19 case: substructure searches in external data sets

Chemical (molecular) similarity is a traditional concept in the field of cheminformatics [45]. It is used to identify structural analogs which might exert similar biological action on similar biological targets [46]. Common cheminformatics similarity approaches are based on the global similarity of a molecule. For example, fingerprint-based descriptors are used to evaluate compound similarity by quantifying the presence/absence of the specific structural features (e.g., distinct functional groups in a molecule). On the contrary, molecular graph-based methods do capture a specific molecular topology and hence account for the local similarity of molecules [47]. Graph-based methods are therefore a robust tool to, e.g., distinguish between different structural isomers (such as

n-pentane and dimethylpropane). Here, Maximum Common Substructures (MCS) of a compound collection were used as structural keys for detecting new potential drug candidates. Such substructure searches are especially useful for drug repositioning strategies, since they more likely capture the local similarity of chemical compounds and therefore allow for more flexibility than global similarity measures (especially if there are large differences of the size of compounds that are being compared).

In a first instance, the Bemis-Murcko scaffold for identified ligands was extracted. For each target, scaffolds were grouped into hierarchical clusters by considering their Maximum Common Substructure (MCS) as a measure of similarity. Afterwards, looping in KNIME was applied to generate one MCS (in SMARTS) per cluster (and target). For details see the Methods Section. In total, 257 distinct MCSs were calculated. A complete list of MCSs can be found in Additional file 1: File S1.

Structural queries generated in the previous step helped identify 7836 compounds from DrugBank and 36,521 compounds from the CAS data set. A complete list of hits found by the substructure searches is provided in Additional file 1: File S2 (DrugBank) and Additional file 1: File S3 (CAS data set). Out of those hits, 135 compounds were retrieved from both DrugBank and the CAS data set (Additional file 1: File S4) and were identified on basis of 18 distinct MCSs (Table 2). Identified MCSs can be combined into five separate clusters (Table 2): (1) Hits identified on basis of the open-chain structural keys (59 hits), (2) Nucleoside/nucleotide analogs (53 hits), (3) miscellaneous, which contain ubiquitous substructures (22 hits), (4) cyclopropane-containing hits (3 hits), and (5) adamantane derivatives (3 hits). Supplementary Figure S1 shows examples of identified hits for the most pronounced clusters. It has to be noted, that the searches do also retrieve compounds that were part of the list of structural queries that were used as an input. For example, remdesivir was rediscovered as part of the substructure searches but it was also included in the original input file. However, for the COVID-19 use case, only less than 2% of hits (820 out of 43,259 compounds) were already part of the input query file.

GLUT-1 deficiency syndrome

Glucose transporter type 1 (GLUT1) deficiency syndrome is characterized by the impairment of glucose transport that might be attributed to mutations in the SLC2A1 gene. Indeed, glucose transporter 1 (GLUT1, encoded by SLC2A1 gene) has been identified as a sole target associated with this disease (with an association score of 1.00 in the Open Targets Platform). Glucose transporter 1 (GLUT-1) is a member of the SLC2A transporter subfamily, being ubiquitously expressed

Table 2 COVID-19 case: Five clusters of enriched Maximum Common Substructures which were retrieved from DrugBank and the CAS data set

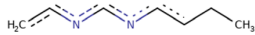


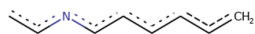


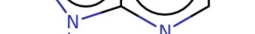

Cluster number	Maximum Common Substructure	SMARTS String	# Hits	Targets
1		[#6]([#7]:[#6]:[#7]-;[#6]-;[#6]-[#6]-[#6]):[#6]	53	PPIA_HUMAN
		[#6](-;[#6]-;[#6]-[#6]-[#6]=[#6]-[#6]-;[#6]-;[#6]	5	PPIA_HUMAN
		[#6];-;[#6]-;[#7]-;[#6];-;[#6];-;[#6];-;[#6];-;[#6];-;[#6]	1	PPIA_HUMAN
2		[#6]1:[#7]:[#6]:[#7]:[#6]2:[#6]:1:[#7]:[#6]:[#7]:2-[#6]1-[#6]-[#6]-[#6]-1	36	R1AB_SARS2
		[#7]1:[#6]:[#7]:[#6]2:[#6]:1:[#7]:[#6]:[#7]:[#6]:2	11	PPIA_HUMAN
		[#6]1:[#7]:[#6]:[#7](-[#6]2-[#8]-[#6]-[#6]-2):[#7]:1	3	PPIA_HUMAN
		[#6]1:[#7]:[#6]:[#7]:[#6]([#6]:1):[#7]([#6]-[#6]1-[#8]-[#6]-[#6]-1	2	PPIA_HUMAN
		[#6]1(-[#6]2:[#6]:[#6]:[#6]3:[#6]:[#7]:[#6]:[#7]:2:3)-[#8]-[#6]-[#6]-[#6]-1	1	R1AB_SARS2

Table 2 (continued)

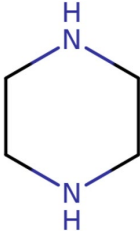
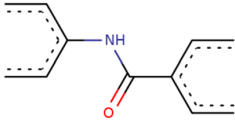
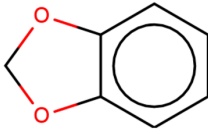
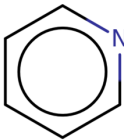
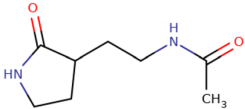
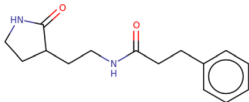
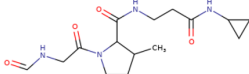
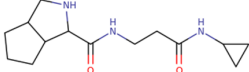
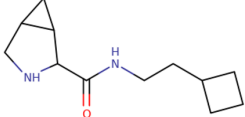

Cluster number	Maximum Common Substructure	SMARTS String	# Hits	Targets
3		[#7]1-[#6]-[#6]-[#7]-[#6]-[#6]-1	8	R1AB_SARS2
		[#6]-[#6]-[#6](-[#6](=[#8])-[#7]-[#6](-[#6]-[#6])-[#6]-[#6])-[#6]-[#6])	4	R1AB_SARS2
		[#6]1:[#6]:[#6]:[#6]2:[#6]([#6]:1)-[#8]-[#6]-[#8]-2	3	R1AB_SARS2
		[#6]1:[#7]:[#6]:[#6]:[#6]:[#6]:1	3	PPIA_HUMAN
		[#6](-[#6]-[#6]1-[#6]-[#6]-[#7]-[#6]-1=[#8])-[#7]-[#6](=[#8])-[#6]	2	R1A_SARS, R1AB_SARS
		[#6](-[#6]-[#6]1-[#6]-[#6]-[#7]-[#6]-1=[#8])-[#7]-[#6](=[#8])-[#6]-[#6]-[#6]1:[#6]:[#6]:[#6]:[#6]:1	2	R1A_SARS, R1AB_SARS
4		[#6](-[#7]-[#6](=[#8])-[#6]1-[#6](-[#6]-[#6]-[#7]-1-[#6](=[#8])-[#6]-[#7]-[#6]=[#8])-[#6])-[#6](=[#8])-[#7]-[#6]1-[#6]-[#6]-1	1	R1AB_SARS2
		[#6](-[#7]-[#6](=[#8])-[#6]1-[#6]2-[#6]-[#6]-[#6]-2-[#6]-[#7]-1)-[#6]-[#6](=[#8])-[#7]-[#6]1-[#6]-[#6]-1	1	R1AB_SARS2
		[#7]1-[#6]-[#6]2-[#6](-[#6]-1-[#6](=[#8])-[#7]-[#6]-[#6]-[#6]1-[#6]-[#6]-[#6]-1)-[#6]-2	1	R1AB_SARS2
5		[#6]12-[#6]-[#6]3-[#6]-[#6](-[#6]-1)-[#6]-[#6](-[#6]-2)-[#6]-3	3	PPIA_HUMAN

Table 2 (continued)

The structural fragment, SMARTS string, the number of identified hits, and the protein target(s) for which these hits have been found, are given

in different tissues, including fetal tissues, mammary glands, placenta, brain, or epithelial cells [48]. GLUT-1 is an essential transmembrane protein for basal glucose uptake.

Symptoms of GLUT1 deficiency syndrome are predominantly seizures, epilepsy and cognitive deficit. GLUT-1 deficiency syndrome is treatable via ketogenic diet [49]. Furthermore, several drugs (e.g., Triheptanoin, DrugBank ID DB11677) have been tested in clinical trials for their efficacy. Up to now, no drug candidate has been found to become an effective treatment for GLUT-1 deficiency syndrome. This neurologic disorder belongs to the group of rare diseases and therefore represents an interesting case study for our drug repurposing pipeline.

Retrieval of GLUT-1 data

By mapping GLUT-1 retrieved from the Open Target Platform to UniProt IDs, protein structures stored in PDB (4 unique structures) have been fetched. From these sources, 4 unique ligands could be extracted, yielding 3

unique Bemis-Murcko scaffolds. Integrative mining of bioactivity data delivered 653 unique compounds from ChEMBL (394 unique Murcko scaffolds), 243 unique compounds from PubChem (115 unique Murcko scaffolds), and 2 unique compounds from IUPHAR (2 unique Murcko scaffolds) with activity < 1 μ M. The threshold for activity label assignment was adopted due to the specific activity range characteristic for membrane transporters [50, 51].

GLUT-1 case: substructure searches in DrugBank

Hierarchical clustering of available Murcko scaffolds delivered 94 fragments used for substructure searches in DrugBank. 18 different fragments (depicted in Fig. 10) have been enriched in 539 unique compounds retrieved from DrugBank (Additional file 1: File S5). 14% of the retrieved hits (28 out of 200 compounds in total) were already part of the input query file and were therefore rediscovered as part of the substructure searches.

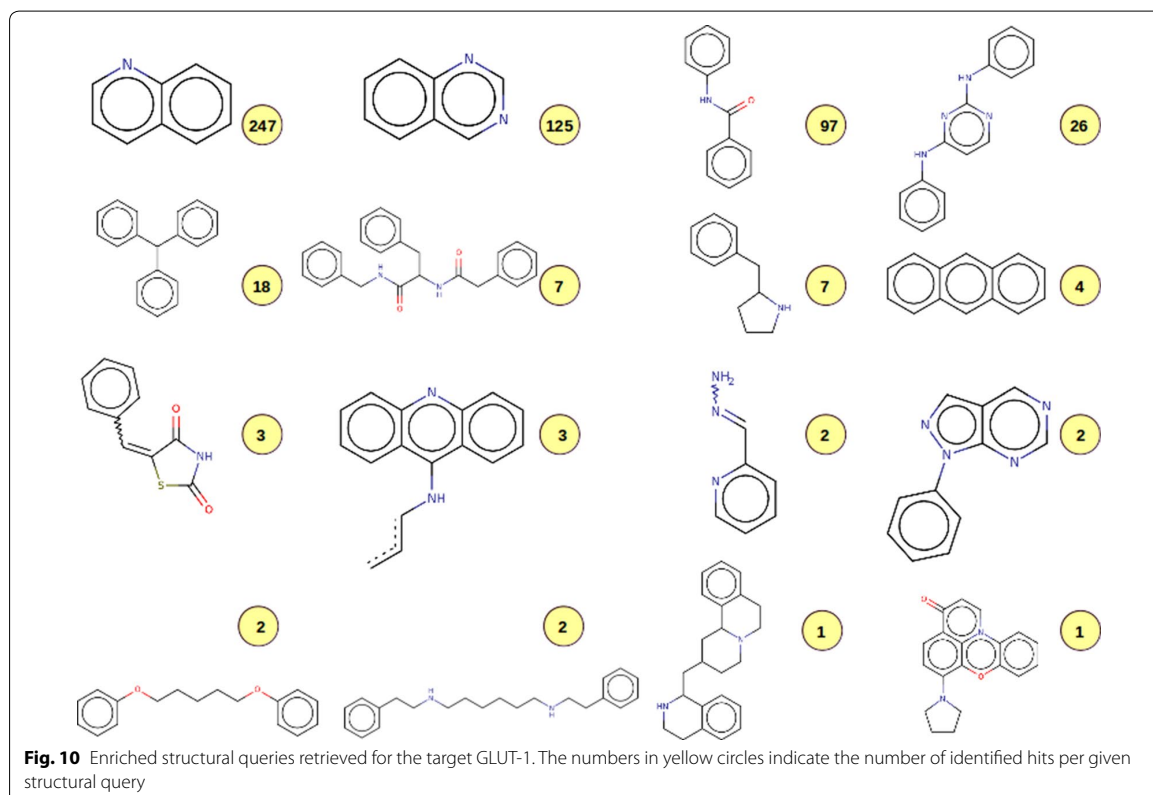


Fig. 10 Enriched structural queries retrieved for the target GLUT-1. The numbers in yellow circles indicate the number of identified hits per given structural query

It appears interesting that most of the identified hits do contain a quinoline (n=247 hits) or quinazoline (n=125) scaffold. These heterocyclic compounds are broadly pharmacologically active [51, 52]. Interestingly, quinoline/quinazoline analogs have been inspected to become promising anticonvulsant agents, as indicated in different studies [53–55]. Since 90% of all patients with this syndrome also develop frequent seizures (<https://medlineplus.gov/genetics/condition/glut1-deficiency-syndrome/>) these classes of compounds could be interesting for future investigations on GLUT-1 deficiency syndrome.

Experiences when using the workflow in the classroom

The workflow described herein has been used in the summer semester 2020 (April 20–24) in the framework of the course “Experimental Methods in Drug Discovery and Preclinical Drug Development” which is part of the English-language Master’s Degree Program *Drug Discovery and Development* at the University of Vienna (<https://drug-dd.univie.ac.at/>). Due to the requirements of social distancing caused by the COVID-19 pandemic, this course was conceptualized as a virtual classroom. The students have attended online sessions, in which the authors of this manuscript have explained the various steps of the workflow. Tutorials and the different parts of the workflow (available at <https://github.com/AlzbetaTuerkova/Drug-Repurposing-in-KNIME>) have been handed out daily in order to not overwhelm the students. On the last day of this 5-days course, each student had to select one of the hits retrieved by the substructure search and dedicate some time to literature searches. Finally, every student submitted a report summarizing what is known about the selected compound and its potential usefulness for COVID-19 treatment (according to what was known in April 2020). Based on the feedback that was provided by students after the course was finalized, the pace of teaching was evenly distributed over the course schedule. The only exception was the step when bioactivity data was retrieved from ChEMBL and PubChem on day 3 of the course. Specifically, some students found it difficult to grasp the essence of the application and execution of the recursive and/or nested loops. In conclusion, the course did provide insights into a variety of KNIME nodes, which can be exploited further for future drug discovery applications.

Summary and conclusions

In this educational paper, we are describing a semi-automatic KNIME workflow for ligand-based *in silico* drug repurposing. The consecutive data mining steps include integration, curation, and analysis of bioassay data from the open domain for specific targets of interest, as well

as the generation of structural queries for automated substructure searches in collections of approved, withdrawn, and/or experimental drugs. Targeted access of data through APIs has been implemented at several stages of the KNIME workflow. Incorporation of API calls into KNIME allows repeating the whole procedure in an automated fashion, e.g., when new data is becoming available. As a consequence of the current COVID-19 pandemic, the cheminformatics analyses performed as a use case herein was tailored to ligand and protein data currently available for drug repurposing strategies in the framework of this life-threatening disease. As a side effect of analyzing the data, we are providing insights into enriched chemical substructures for proposed drug targets of SARS-CoV-2. In addition, the workflow has been used to detect data coverage and enriched clusters for the treatment of a rare disease, GLUT-1 deficiency syndrome. The material has been used successfully for teaching undergraduate students the use of programmatic data access via KNIME workflows and subsequent data analysis steps. The workflows, tutorials, and the information gained on COVID-19 and GLUT-1 data are freely available to the scientific community for follow-up studies or may be tailored to specific needs of other use cases (available at <https://github.com/AlzbetaTuerkova/Drug-Repurposing-in-KNIME>).

Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1186/s13321-020-00474-z>.

Additional file 1. Supplementary Information—(1) a list of drug targets with potential interest for treatment of COVID-19, available from https://covid-19.uniprot.org/uniprotkb?query=* and (2) from the Open Targets Platform, (3) number of unique ligands gathered from PDB, ChEMBL, PubChem, and IUPHAR for COVID-19 targets from UniProt pre-release web page and (4) from the Open Targets Platform (5) examples of identified drugs with the highlighted structural query, and (6) description of the supplementary data files.

Abbreviations

API: Application Programming Interface; KNIME: Konstanz Information Miner; CDK: Chemistry Development Kit; UniProtKB: The Universal Protein Resource KnowledgeBase; COVID-19: Coronavirus Disease 2019; SARS-CoV-2: Severe Acute Respiratory Syndrome Coronavirus 2; PDB: Protein Data Bank; NMR: Nuclear Magnetic Resonance; Cryo-EM: Cryo-Electron Microscopy; RCSR: Research Collaboratory for Structural Bioinformatics; AID: Assay ID; CID: Compound ID; CAS: Chemical Abstract Service; MCS: Maximum Common Substructure; PCA: Principal Component Analysis; LabuteASA: Labute’s Accessible Surface Area; SMR: Molecular Refractivity; TPSA: Topological Polar Surface Area; GLUT-1: Glucose Transporter Type 1; URL: Uniform Resource Locator.

Acknowledgements

The authors acknowledge active involvement of undergraduate students participating in the course “Experimental Methods in Drug Discovery and Preclinical Drug Development” in the summer semester 2020 at the University of Vienna in testing and applying the developed tutorial and KNIME workflow.

Authors' contributions

AT and BZ conceptualized and designed the study. AT generated the KNIME workflows, performed the data integration, processing and analyses. BZ provided advice. The manuscript was written through contributions of both authors. Both authors read and approved the final manuscript.

Funding

No funding was received for the present study.

Competing interests

The authors declare no competing interests.

Received: 6 July 2020 Accepted: 9 November 2020

Published online: 25 November 2020

References

- Karaman B, Sippl W (2019) Computational drug repurposing: current trends. *Curr Med Chem* 26(28):5389–5409 ([[cito:citesAsAuthority](#)][[cito:agreesWith](#)])
- Bajorath J (2017) Compound data mining for drug discovery. In: Keith JM (ed) *Bioinformatics: volume II: structure, function, and applications*. Springer, New York, NY, pp 247–256
- Agatonovic-Kustrin S, Morton D (2016) Chapter 9—data mining in drug discovery and design. In: Puri M, Pathak Y, Sutariya VK, Tipparaju S, Moreno W (eds) *Artificial neural network for drug design, delivery and disposition*. Academic Press, Boston, pp 181–193 ([[cito:citesAsAuthority](#)][[cito:agreesWith](#)])
- Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Félix E et al (2019) ChEMBL: towards the deposition of bioassay data. *Nucleic Acids Res* 47(D1):D930–D940 ([[cito:usesDataFrom](#)][[cito:citesAsDataSource](#)])
- Kim S, Chen J, Cheng T, Gindulyte A, He J, He S et al (2019) PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res* 47(D1):D1102–D1109 ([[cito:usesDataFrom](#)][[cito:citesAsDataSource](#)])
- Consortium TU (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* 47(D1):D506–D515 ([[cito:usesDataFrom](#)][[cito:citesAsDataSource](#)])
- Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR et al (2018) DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* 46(D1):D1074–D1082 ([[cito:usesDataFrom](#)][[cito:citesAsDataSource](#)])
- Qian T, Zhu S, Hoshida Y (2019) Use of big data in drug development for precision medicine: an update. *Expert Rev Precis Med Drug Dev* 4(3):189–200 ([[cito:citesAsAuthority](#)][[cito:agreesWith](#)])
- Berthold MR, Cebon N, Dill F, Gabriel TR, Köttler T, Meinel T et al (2009) KNIME—the Konstanz information miner: version 2.0 and beyond. *ACM SIGKDD Explor Newsl* 11(1):26–31 ([[cito:usesMethodIn](#)])
- Landrum G. RDKit Documentation. p 159. [[cito:usesMethodIn](#)]
- Beiske S, Meinel T, Wiswedel B, de Figueiredo LF, Berthold M, Steinbeck C (2013) KNIME-CDK: Workflow-driven cheminformatics. *BMC Bioinform* 14(1):257 ([[cito:usesMethodIn](#)])
- Pavlov D, Rybalkin M, Karulin B, Kozhevnikov M, Savelyev A, Churinov A (2011) Indigo: universal cheminformatics API. *J Cheminformatics* 3(Suppl 1):P4 ([[cito:citesAsAuthority](#)])
- Roughley S. Five Years of the KNIME Vernalis Cheminformatics Community Contribution. *Curr Med Chem*. 2018; [[cito:citesAsAuthority](#)]
- Pushpakom S, Iorio F, Eyers PA, Escott KJ, Hopper S, Wells A et al (2019) Drug repurposing: progress, challenges and recommendations. *Nat Rev Drug Discov* 18(1):41–58 ([[cito:citesAsAuthority](#)][[cito:discusses](#)])
- Fetro C, Scherman D (2020) Drug repurposing in rare diseases: myths and reality. *Therapies* 75(2):157–160 ([[cito:citesAsAuthority](#)][[cito:discusses](#)])
- Jarada TN, Rokne JG, Alhajj R (2020) A review of computational drug repositioning: strategies, approaches, opportunities, challenges, and directions. *J Cheminform* 12(1):46 ([[cito:citesAsAuthority](#)][[cito:discusses](#)])
- Li J, Zhu X, Chen JY (2009) Building disease-specific drug-protein connectivity maps from molecular interaction networks and PubMed abstracts. *PLOS Comput Biol* 5(7):1000450 ([[cito:citesAsAuthority](#)][[cito:discusses](#)])
- Shawe-Taylor J, Cristianini N, editors. *Support Vector Machines*. In: An introduction to support vector machines and other kernel-based learning methods. Cambridge: Cambridge University Press; 2000. p. 93–124. <https://www.cambridge.org/core/books/an-introduction-to-support-vector-machines-and-other-kernelbased-learning-methods/support-vector-machines/DD4EA48AA6C383944EA67BF8A7BEC6CC> [[cito:citesAsAuthority](#)][[cito:discusses](#)]
- Susnow RG, Dixon SL (2003) Use of robust classification techniques for the prediction of human cytochrome P450 2D6 inhibition. *J Chem Inf Comput Sci* 43(4):1308–1315 ([[cito:citesAsAuthority](#)][[cito:discusses](#)])
- Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T (2018) The rise of deep learning in drug discovery. *Drug Discov Today* 23(6):1241–1250 ([[cito:citesAsAuthority](#)][[cito:discusses](#)])
- Wu C, Gudivada RC, Aronow BJ, Jegga AG (2013) Computational drug repositioning through heterogeneous network clustering. *BMC Syst Biol* 7(5):56 ([[cito:citesAsAuthority](#)][[cito:discusses](#)])
- Wang F, Wu F-X, Li C-Z, Jia C-Y, Su S-W, Hao G-F et al (2019) ACID: a free tool for drug repurposing using consensus inverse docking strategy. *J Cheminform* 11(1):73 ([[cito:citesAsAuthority](#)][[cito:discusses](#)])
- Steinmetz FP, Mellor CL, Meinel T, Cronin MTD (2015) Screening chemicals for receptor-mediated toxicological and pharmacological endpoints: using public data to build screening tools within a KNIME Workflow. *Mol Inform* 34(2–3):171–178 ([[cito:citesAsAuthority](#)][[cito:agreesWith](#)])
- Gordon DE, Jang GM, Bouhaddou M, Xu J, Obernier K, White KM et al (2020) A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* 30:1–13 ([[cito:citesAsAuthority](#)][[cito:discusses](#)][[cito:agreesWith](#)])
- Carvalho-Silva D, Pierleoni A, Pignatelli M, Ong C, Fumis L, Karamanis N et al (2019) Open Targets Platform: new developments and updates two years on. *Nucleic Acids Res* 47(D1):D1056–D1065 ([[cito:usesDataFrom](#)][[cito:citesAsDataSource](#)][[cito:discusses](#)])
- Goodsell DS, Zardecki C, Costanzo LD, Duarte JM, Hudson BP, Persikova I et al (2020) RCSB Protein Data Bank: enabling biomedical research and drug discovery. *Protein Sci* 29(1):52–65 ([[cito:usesDataFrom](#)][[cito:citesAsDataSource](#)])
- Pawson AJ, Sharman JL, Benson HE, Faccenda E, Alexander SPH, Buneman OP et al (2014) The IUPHAR/BPS Guide to PHARMACOL-OGY: an expert-driven knowledgebase of drug targets and their ligands. *Nucleic Acids Res* 42(D1):D1098–D1106 ([[cito:usesDataFrom](#)][[cito:citesAsDataSource](#)])
- Bemis GW, Murcko MA (1996) The properties of known drugs. 1. Molecular frameworks. *J Med Chem* 39(15):2887–2893 ([[cito:usesMethodIn](#)])
- Burley SK, Berman HM, Bhikadiya C, Bi C, Chen L, Di Costanzo L et al (2019) RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res* 47(D1):D464–D474 ([[cito:usesDataFrom](#)][[cito:citesAsDataSource](#)])
- Davies M, Nowotka M, Papadatos G, Dedman N, Gaulton A, Atkinson F et al (2015) ChEMBL web services: streamlining access to drug discovery data and utilities. *Nucleic Acids Res* 43(2):W612–W620 ([[cito:usesMethodIn](#)])
- Gadaleta D, Lombardo A, Toma C, Benfenati E (2018) A new semi-automated workflow for chemical data retrieval and quality checking for modeling applications. *J Cheminformatics* 10(1):1–13 ([[cito:usesMethodIn](#)])
- Sayers EW, Cavanaugh M, Clark K, Ostell J, Pruitt KD, Karsch-Mizrachi I (2019) GenBank. *Nucleic Acids Res* 47(D1):D94–D99 ([[cito:citesAsAuthority](#)])
- Pujar S, O'Leary NA, Farrell CM, Loveland JE, Mudge JM, Wallin C et al (2018) Consensus coding sequence (CCDS) database: a standardized set of human and mouse protein-coding regions supported by expert curation. *Nucleic Acids Res* 46(D1):D221–D228 ([[cito:citesAsAuthority](#)])
- Pieper U, Webb BM, Dong GQ, Schneidman-Duhovny D, Fan H, Kim SJ et al (2014) ModBase, a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res* 42(D1):D336–D346 ([[cito:citesAsAuthority](#)])
- Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumieny R et al (2018) SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res* 46(W1):W296–303 ([[cito:citesAsAuthority](#)])

36. Oughtred R, Stark C, Breitkreutz B-J, Rust J, Boucher L, Chang C et al (2019) The BioGRID interaction database: 2019 update. *Nucleic Acids Res* 47(D1):D529–D541 ([[citesAsAuthority](#)])
37. Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F et al (2014) The MintAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res* 42(D1):D358–D363 ([[citesAsAuthority](#)])
38. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J et al (2019) STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* 47(D1):D607–D613 ([[citesAsAuthority](#)])
39. Gilson MK, Liu T, Baitaluk M, Nicola G, Hwang L, Chong J (2016) BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res* 44(D1):D1045–D1053 ([[citesAsAuthority](#)])
40. Littler DR, Gully BS, Colson RN, Rossjohn J (2020) Crystal structure of the SARS-CoV-2 non-structural protein 9, Nsp9. *iScience* 23(7):101258 ([[citesAsAuthority](#)])
41. Zhang L, Lin D, Sun X, Curth U, Drosten C, Sauerhering L et al (2020) Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α -ketoamide inhibitors. *Science* 368(6489):409–412 ([[citesAsAuthority](#)][[citesDiscusses](#)])
42. Yang J, Petitjean SJL, Koehler M, Zhang Q, Dumitru AC, Chen W et al (2020) Molecular interaction and inhibition of SARS-CoV-2 binding to the ACE2 receptor. *Nat Commun* 11(1):4541 ([[citesAsAuthority](#)][[citesDiscusses](#)])
43. Fang SG, Shen H, Wang J, Tay FPL, Liu DX (2008) Proteolytic processing of polyproteins 1a and 1ab between non-structural proteins 10 and 11/12 of Coronavirus infectious bronchitis virus is dispensable for viral replication in cultured cells. *Virology* 379(2):175–180 ([[citesAsAuthority](#)][[citesDiscusses](#)])
44. Sigrist CJ, Bridge A, Le Mercier P (2020) A potential role for integrins in host cell entry by SARS-CoV-2. *Antiviral Res* 177:104759 ([[citesAsAuthority](#)][[citesDiscusses](#)])
45. Johnson MA, Maggiora GM (1990) Concepts and applications of molecular similarity. Wiley, New York, p 420 ([[citesAsAuthority](#)][[citesDiscusses](#)])
46. Martin YC, Kofron JL, Traphagen LM (2002) Do structurally similar molecules have similar biological activity? *J Med Chem* 45(19):4350–4358 ([[citesAsAuthority](#)][[citesDiscusses](#)])
47. Cao Y, Jiang T, Girke T (2008) A maximum common substructure-based algorithm for searching and predicting drug-like compounds. *Bioinformatics* 24(13):i366–i374 ([[usesMethodIn](#)][[citesDiscusses](#)])
48. Wood IS, Trayhurn P (2003) Glucose transporters (GLUT and SGLT): expanded families of sugar transport proteins. *Br J Nutr* 89(1):3–9 ([[citesAsAuthority](#)][[citesDiscusses](#)])
49. Klepper J, Leiendecker B, Bredahl R, Athanassopoulos S, Heinen F, Gertsen E et al (2002) Introduction of a ketogenic diet in young infants. *J Inher Metab Dis* 25(6):449–460 ([[citesAsAuthority](#)][[citesDiscusses](#)])
50. Tanoli Z, Alam Z, Ianevski A, Wennerberg K, Vähä-Koskela M, Aittokallio T (2020) Interactive visual analysis of drug–target interaction networks using Drug Target Profiler, with applications to precision medicine and drug repurposing. *Brief Bioinform* 21(1):211–220 ([[citesAsAuthority](#)][[citesDiscusses](#)])
51. Wei C-X, Bian M, Gong G-H (2015) Current research on antiepileptic compounds. *Molecules* 20(11):20741–20776 ([[citesAsAuthority](#)][[citesAgreesWith](#)])
52. Ugale VG, Bari SB (2014) Quinazolines: new horizons in anticonvulsant therapy. *Eur J Med Chem* 10(80):447–501 ([[citesAsAuthority](#)][[citesAgreesWith](#)])
53. Cui L-J, Xie Z-F, Piao H-R, Li G, Chai K-Y, Quan Z-S (2005) Synthesis and anticonvulsant activity of 1-substituted-7-benzoyloxy-4,5-dihydro-[1,2,4]triazolo[4,3-a]quinoline. *Biol Pharm Bull* 28(7):1216–1220 ([[citesAsAuthority](#)][[citesAgreesWith](#)])
54. Xie Z-F, Chai K-Y, Piao H-R, Kwak K-C, Quan Z-S (2005) Synthesis and anticonvulsant activity of 7-alkoxy-4,5-dihydro-[1,2,4]triazolo[4,3-a]quinolines. *Bioorg Med Chem Lett* 15(21):4803–4805 ([[citesAsAuthority](#)][[citesAgreesWith](#)])
55. Jin H-G, Sun X-Y, Chai K-Y, Piao H-R, Quan Z-S (2006) Anticonvulsant and toxicity evaluation of some 7-alkoxy-4,5-dihydro-[1,2,4]triazolo[4,3-a]quinoline-1(2H)-ones. *Bioorg Med Chem* 14(20):6868–6873 ([[citesAsAuthority](#)][[citesAgreesWith](#)])

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions



3.4 Differences in Metformin and Thiamine Uptake between Human and Mouse Organic Cation Transporter OCT1: Structural Determinants and Potential Consequences for Intrahepatic Concentrations

MEYER, Marleen J.; TUERKOVA, Alzbeta, RÖMER, Sarah; WENZEL, Christoph; SEITZ, Tina; GAEDCKE, Jochen; OSWALD, Stefan; BROCKMÖLLER, Jürgen; ZDRAZIL, Barbara; TZVETKOV, Mladen V. *Drug Metabolism and Disposition*, **2020**.

* *Corresponding author: mladen.tzvetkov@med.uni-greifswald.de*



M.J. Meyer, B. Zdrazil, J. Brockmöller, and M.V. Tzvetkov participated in research design. M.J. Meyer, A. Tuerkova, C. Wenzel conducted experiments. A. Tuerkova, S. Römer, T. Seitz, J. Gaedcke, and B. Zdrazil contributed new reagents or analytic tools. M.J. Meyer, A. Tuerkova, S. Römer, C. Wenzel, S. Oswald, B. Zdrazil, and M.V. Tzvetkov performed data analysis. M.J. Meyer, A. Tuerkova, S. Römer, S. Oswald, J. Brockmöller, B. Zdrazil, and M.V. Tzvetkov wrote the manuscript.

The Supplementary Information can be found in Part V.

The following article is reprinted from:

Meyer, M. J., Tuerkova, A., Römer, S., Wenzel, C., Seitz, T., Gaedcke, J., ... Tzvetkov, M. V. (2020). Differences in Metformin and Thiamine Uptake between Human and Mouse Organic Cation Transporter 1: Structural Determinants and Potential Consequences for Intrahepatic Concentrations. *Drug Metabolism and Disposition*, 48(12), 1380-1392.

Differences in Metformin and Thiamine Uptake between Human and Mouse Organic Cation Transporter 1: Structural Determinants and Potential Consequences for Intrahepatic Concentrations^S

Marleen J. Meyer, Alzbeta Tuerkova, Sarah Römer, Christoph Wenzel, Tina Seitz, Jochen Gaedcke, Stefan Oswald, Jürgen Brockmöller,  Barbara Zdravil, and  Mladen V. Tzvetkov

Institute of Pharmacology, Center of Drug Absorption and Transport (C_DAT), University Medicine Greifswald, Greifswald, Germany (M.J.M., S.R., C.W., S.O., M.V.T.); Department of Pharmaceutical Chemistry, Division of Drug Design and Medicinal Chemistry, University of Vienna, Vienna, Austria (A.T., B.Z.); and Department of General, Visceral, and Pediatric Surgery (J.G.) and Institute of Clinical Pharmacology (T.S., J.B.), University Medical Center Göttingen, Göttingen, Germany

Received July 6, 2020; accepted September 28, 2020

ABSTRACT

The most commonly used oral antidiabetic drug, metformin, is a substrate of the hepatic uptake transporter OCT1 (gene name *SLC22A1*). However, OCT1 deficiency leads to more pronounced reductions of metformin concentrations in mouse than in human liver. Similarly, the effects of OCT1 deficiency on the pharmacokinetics of thiamine were reported to differ between human and mouse. Here, we compared the uptake characteristics of metformin and thiamine between human and mouse OCT1 using stably transfected human embryonic kidney 293 cells. The affinity for metformin was 4.9-fold lower in human than in mouse OCT1, resulting in a 6.5-fold lower intrinsic clearance. Therefore, the estimated liver-to-blood partition coefficient is only 3.34 in human compared with 14.4 in mouse and may contribute to higher intrahepatic concentrations in mice. Similarly, the affinity for thiamine was 9.5-fold lower in human than in mouse OCT1. Using human-mouse chimeric OCT1, we showed that simultaneous substitution of transmembrane helices TMH2 and TMH3 resulted in the reversal of affinity for metformin. Using homology modeling, we suggest several explanations, of which a different interaction of Leu155 (human

TMH2) compared with Val156 (mouse TMH2) with residues in TMH3 had the strongest experimental support. In conclusion, the contribution of human OCT1 to the cellular uptake of thiamine and especially of metformin may be much lower than that of mouse OCT1. This may lead to an overestimation of the effects of OCT1 on hepatic concentrations in humans when using mouse as a model. In addition, comparative analyses of human and mouse orthologs may help reveal mechanisms of OCT1 transport.

SIGNIFICANCE STATEMENT

OCT1 is a major hepatic uptake transporter of metformin and thiamine, but this study reports strong differences in the affinity for both compounds between human and mouse OCT1. Consequently, intrahepatic metformin concentrations could be much higher in mice than in humans, impacting metformin actions and representing a strong limitation of using rodent animal models for predictions of OCT1-related pharmacokinetics and efficacy in humans. Furthermore, OCT1 transmembrane helices TMH2 and TMH3 were identified to confer the observed species-specific differences in metformin affinity.

Introduction

Metformin is the most commonly prescribed oral antidiabetic drug. It reduces plasma glucose and has favorable effects on lipid metabolism.

This work was supported in part by European Regional Development Fund (ERDF) [Grant GHS-18-0021] to M.V.T.

Part of this work has been presented in the poster “OCT1 of mice and men – species-specific differences in the function of OCT1” by Meyer, MJ; Bolesta, M; Schreier, P; Krätzner, R; Seitz, T; Brockmöller, J; Tzvetkov, MV at the 11th International BioMedical Transporters Conference, Aug. 4–8, 2019 in Lucerne, Switzerland. This work is part of the Ph.D. thesis of M.J.M.

<https://doi.org/10.1124/dmd.120.000170>

^SThis article has supplemental material available at dmd.aspetjournals.org.

Metformin acts both in the liver and in the gut (Rena et al., 2017). In hepatocytes, metformin decreases glucose production and lipogenesis both by AMPK-dependent and AMPK-independent mechanisms.

At physiologic pH, metformin is almost entirely present as organic cation (pK_a of 12.04, estimated 99.998% positively charged molecules). Therefore, metformin depends highly on transporter proteins to enter hepatocytes. The organic cation transporter 1 (OCT1, gene name *SLC22A1*), which is strongly expressed in the sinusoidal membrane of hepatocytes, has been demonstrated to be the major hepatic uptake transporter of metformin (Wang et al., 2002; Shu et al., 2007). In humans, OCT1 is genetically highly variable. In total, 9% of Europeans and white Americans are carriers of two reduced function or loss-of-function OCT1 alleles and are so-called poor OCT1

ABBREVIATIONS: AMPK, AMP-activated protein kinase; DPBS, Dulbecco's phosphate-buffered saline; GLUT3, glucose 3 transporter; h, human; HBSS, Hanks' buffered salt solution; HEK, human embryonic kidney; IS, internal standard; IVIVE, in vitro to in vivo extrapolation; K_p , liver-to-blood partition coefficient; LC-MS/MS, liquid chromatography tandem mass spectrometry; m, mouse; MFS, major facilitator superfamily; OCT, organic cation transporter; TMH, transmembrane helix PET, positron emission tomography; DAPI, 4',6'-diamidino-2-phenylindole; THTR, thiamine transporter; SLC, solute carrier.

transporters (Kerb et al., 2002; Shu et al., 2003; Tzvetkov et al., 2012; Seitz et al., 2015). In some specific populations, like Surui Indians, this percentage may increase up to 80% (Seitz et al., 2015). These genetic OCT1 variants are expected to affect the hepatic uptake and thus the efficacy of metformin.

Indeed, OCT1 knockout in mice reduced metformin concentrations in the liver by up to 30-fold (Wang et al., 2002, 2003) and abolished the glucose-lowering effects of metformin (Shu et al., 2007). OCT1 deficiency in mice has been suggested to lead to an absolute lack of metformin uptake into hepatocytes (Wang et al., 2002). Also in humans, genetic variants leading to decreased OCT1 activity were associated with reduced intrahepatic concentrations of metformin (Sundelin et al., 2017). However, despite some initial reports of reduced response to metformin in poor OCT1 transporters (Shu et al., 2007), larger studies and meta-analyses could not confirm the association of OCT1 genetic variants with reduced efficacy of metformin in humans (Zhou et al., 2009; Dujic et al., 2017).

Similarly, OCT1 was identified as a relevant transporter of thiamine (vitamin B1) in the mouse liver (Chen et al., 2014). OCT1 knockout in mice resulted in higher thiamine plasma levels, likely because of decreased hepatic extraction (Chen et al., 2014; Liang et al., 2018). However, a recent study in humans showed no differences in the plasma concentrations of thiamine or its metabolites in poor OCT1 transporters (Jensen et al., 2020). Taken together, this questions the suitability of mouse as a model for studying effects of OCT1 on metformin or thiamine-driven biologic processes without understanding the causes of the species differences.

One possible explanation for the variable results between human and mouse may be differences in the kinetics of OCT1-mediated uptake of metformin and thiamine. Although species differences in organ-specific OCT1 expression are well characterized (Gorboulev et al., 1997; Zhang et al., 1997; Green et al., 1999; Schmitt et al., 2003), there is only very limited data reporting differences in transport activity. The amino acid identity between the human and mouse OCT1 orthologs is 77%. Since the exact mechanism of substrate interaction with OCT1 is not known, it is difficult to predict to what extent the 23% different amino acids between the two orthologs can confer differences in uptake. Furthermore, metformin and thiamine were suggested to share common sites of ligand-transporter interaction within OCT1 (Chen et al., 2014). Therefore, similar differences in the uptake between human and mouse could be expected for these two substrates.

The aim of this study was to compare the uptake of metformin and thiamine by human and mouse OCT1 *in vitro* to explore underlying mechanisms causing differences in the hepatic concentrations in humans and mice. This should help to better interpret the data of mouse models and should improve the translation of mouse pharmacokinetic data to humans. Furthermore, we used the differences in uptake between human and mouse OCT1 as a tool to improve our understanding of the transport mechanism of OCT1.

Materials and Methods

Reagents. Metformin hydrochloride, thiamine hydrochloride, ammoniumbicarbonate, dithiothreitol, and iodoacetamide were obtained from Sigma-Aldrich (Taufkirchen, Germany); bufomarin hydrochloride was obtained from Wako Chemicals (Neuss, Germany); and thiamine-d3 hydrochloride was obtained from Toronto Research Chemicals (North York, ON, Canada). All chemicals used in this study were purchased from commercial sources and had purities of 95% or higher. Dulbecco's modified Eagle's medium (DMEM), Hanks' buffered salt solution (HBSS), and additives for cell culturing were obtained from Life Technologies (Darmstadt, Germany). Dulbecco's phosphate-buffered saline (DPBS) was obtained from PAN-Biotech (Aidenbach, Germany). Poly-D-lysine (1–5 kDa), HEPES, bicinchoninic acid, and copper sulfate pentahydrate were

obtained from Sigma-Aldrich. Twelve-well plates were obtained from Nunc (Langensfeld, Germany), and tissue culture flasks were from Sarstedt (Nümbrecht, Germany). Acetonitrile, methanol, and formic acid in LC-MS/MS grade and sodium chloride were obtained from Merck (Darmstadt, Germany). SDS (ultrapure) was obtained from AppliChem (Darmstadt, Germany). Sequencing Grade Modified Trypsin and ProteaseMAX surfactant were obtained from Promega (Mannheim, Germany).

Generation of OCT1 Constructs. For overexpression of OCT1 in human embryonic kidney (HEK) 293 cells, pcDNA5/FRT expression vectors (Thermo Fisher Scientific, Darmstadt, Germany) containing wild-type, mutant, or chimeric OCT1 constructs were generated as follows or as described previously (Tzvetkov et al., 2012; Seitz et al., 2015). Human-mouse chimeric OCT1 was generated by restriction of human and mouse OCT1 genes with Bsu36I and BsaBI, separating the OCT1 gene into three fragments: from N terminus to large intracellular loop, from transmembrane helix (TMH) 7 to TMH9, and from TMH10 to C terminus (Fig. 3A). The fragments were ligated back together in the correct order but with different combinations of the species, and the resulting chimeric OCT1 genes were cloned into the pcDNA5/FRT vector after restriction of both gene and vector with HindIII and EcoRV. These constructs were then used for targeted chromosomal integration into HEK293 cells. Human-mouse chimeric OCT1 constructs with single TMH substitutions were generated using the overlap extension method (Horton et al., 1989) and primers listed in Supplemental Table 1. Point mutations in human and mouse OCT1 genes were introduced by site-directed mutagenesis in pcDNA5/FRT vectors containing human or mouse OCT1 wild-type genes, using primers listed in Supplemental Table 1. All generated constructs were validated by capillary sequencing of the complete open reading frame of OCT1 before transfection into HEK293 cells.

Cell Lines and Cell Culturing. HEK293 cells stably overexpressing human OCT1, mouse OCT1, rat OCT1, human-mouse chimeric OCT1, or human OCT2 were generated by targeted chromosomal integration using the Flip-In System (Life Technologies) as described previously (Tzvetkov et al., 2012; Seitz et al., 2015). Cells were cultured in Dulbecco's modified Eagle's medium supplemented with 10% FBS, 100 U/ml penicillin, and 100 µg/ml streptomycin at 37°C and 5% CO₂.

Transient Transfection of T-REx-293 Cells for Cellular Uptake Experiments. For transient transfection of OCT1 constructs into HEK293 cells for uptake experiments, 5×10^5 T-REx-293 cells (Life Technologies) were seeded per well of a 12-well plate precoated with poly-D-lysine. At 24 hours later, the cells were transfected with 100 µl of reaction mix per well, containing 2 µg pcDNA5/FRT vector with the OCT1 construct of interest, 0.5 µg pGFP-tpz vector, and 6.25 µl Lipofectamine 2000 (Thermo Fisher Scientific), according to the manufacturer's instructions. At 48 hours later, transfection efficacy was assessed microscopically by visualizing the GFP signal of the cotransfected GFP vector, and the cells were used for uptake experiments.

Cellular Uptake Experiments. At 48 hours prior to the experiment, 6×10^5 cells were seeded per well of a 12-well plate. When using transiently transfected cells, 5×10^5 T-REx-293 cells were seeded per well of a 12-well plate 72 hours prior to the experiment, and they were transfected 24 hours later as described above. Twelve-well plates were precoated with poly-D-lysine.

Cellular uptake experiments were performed at 37°C and pH 7.4 using HBSS supplemented with 10 mM HEPES (in the following referred to as HBSS+). Cells were washed once with 1 ml of prewarmed (37°C) HBSS+, and uptake was started by adding 400 µl prewarmed HBSS+ containing the substrate. Uptake was stopped after 2 minutes by adding 2 ml ice-cold HBSS+. Cells were washed twice with 2 ml ice-cold HBSS+ and were lysed in 500 µl 80% acetonitrile supplemented with internal standard (Table 1). Intracellular substrate concentrations were measured by LC-MS/MS as described below and afterward were normalized to the total amount of protein in the sample as measured using the bicinchoninic acid assay (Smith et al., 1985).

Quantification of Intracellular Substrate Concentration by LC-MS/MS. For quantification of intracellular substrate concentrations, the cell debris was removed by centrifugation of the cell lysate at 16,000g for 15 minutes. In total, 350 µl of the supernatant was evaporated to dryness under nitrogen flow at 40°C. The sample was reconstituted in 200 µl 0.1% formic acid, and 5 or 15 µl was injected into the LC-MS/MS system for metformin and thiamine, respectively.

For LC-MS/MS quantification, an API 4000 tandem mass spectrometer (AB SCIEX, Darmstadt, Germany) was used. Samples were separated on a Brownlee SPP RP-Amide column (4.6 × 100 mm, 2.7 µm; PerkinElmer, Rodgau,

TABLE 1
Parameters of quantitative LC-MS/MS analyses

Analyte	Quantifier Precursor Ion to Product Ion (m/z)	Retention Time (min)	IS	IS Precursor Ion to Product Ion (m/z)	Retention Time IS (min)	Mobile Phase (% Organic Solvent) ^a	Flow (μl/min)
Metformin	130.1 > 71	2.88	Buformin	158.1 > 60	4.0	3	300
Thiamine	265.3 > 122	2.47	Thiamine-d3	269.1 > 125	2.47	3	350

m/z, mass-to-charge ratio.

^aSix parts acetonitrile + one part methanol.

Germany) using a mobile phase of 0.1% (v/v) formic acid and varying concentrations of organic solvent (parameters are listed in Table 1).

Immunocytochemical Staining and Confocal Microscopy Analysis of OCT1-Overexpressing Cells. For immunocytochemical staining of OCT1, 6×10^5 HEK293 cells stably overexpressing human, mouse, or human-mouse chimeric OCT1 were seeded onto coverslips in 12-well plates 48 hours prior to the experiment. Coverslips were precoated with poly-D-lysine. Cells were washed twice with 1 ml DPBS for 10 minutes and were fixed with 100% ethanol for 20 minutes at -20°C . After washing three times with DPBS for 5 minutes, cell membranes were permeabilized with DPBS/0.4% Tween 20 for 10 minutes. Cells were washed three times with DPBS for 5 minutes and blocked with blocking buffer (DPBS/5% FBS) for 1–3 hours. Cells were incubated with the primary antibodies diluted in blocking buffer (according to Supplemental Table 2) in a humid chamber overnight. The next day, after washing three times with DPBS for 5 minutes, the cells were incubated with the secondary antibodies diluted in blocking buffer (according to Supplemental Table 2) for 1 to 2 hours protected from light. After washing three times with DPBS for 5 minutes, coverslips were mounted with ROTI-Mount FluorCare DAPI (Carl Roth, Karlsruhe, Germany) onto microscope slides. The cells were analyzed using a laser scanning microscope (LSM780; Carl Zeiss, Jena, Germany), and the images were processed using the Fiji distribution of ImageJ2 (Schindelin et al., 2012; Rueden et al., 2017).

Quantification of OCT1 Protein Abundance by Targeted Proteomics.

Normal human liver tissue was obtained as excess material, which had to be removed for technical reasons during liver surgery. Patients had given their informed consent for research use of the tissues, and the procedures were approved by the ethics committee of the University Medicine Göttingen, Georg-August-Universität Göttingen (application number 26/01/17). Preparation of murine liver was carried out in compliance with the German laws on animal welfare (§ 4 Absatz 3 TierSchG), and all animals used were reported to the Landesveterinär- und Lebensmitteluntersuchungsamt Mecklenburg-Vorpommern.

Human ($N = 12$, eight female and four male) and murine [$N = 18$, each $N = 6$ of C57BL/6N (two female and four male), C57BL/6J, and FVB mice (each three female and three male)] liver samples were mechanically crushed in a stainless-steel mortar system, precooled in liquid nitrogen. Approximately 100 mg tissue powder was used for isolation of native integral membrane using the ProteoExtract Native Membrane Extraction Kit according to the manufacturer's instructions (Merck). Tissues were additionally homogenized in a glass douncer during the step of cell lysis. Cell pellets of OCT1-overexpressing cells were directly used for extraction. Total protein content of the resulting membrane fraction was determined by bicinchoninic acid assay. If necessary, membrane fractions were adjusted to a maximum protein amount of 2 μg/μl. Subsequently, 100 μl of each membrane fraction was mixed with 10 μl dithiothreitol (200 mM), 40 μl ammonium bicarbonate buffer (50 mM, pH 7.8), and 10 μl ProteaseMAX (1%, m/v) and incubated for 30 minutes at 60°C (denaturation). After cooling down, 10 μl iodoacetamide (400 mM) was added, and the samples were incubated in a darkened water bath for 15 minutes at 37°C (alkylation). For protein digestion, 10 μl trypsin (trypsin/protein ratio of 1:40) was added, and samples were incubated in a water bath for 16 hours at 37°C . Digestion was stopped by addition of 20 μl formic acid (10%, v/v). All samples were stored at -80°C until further processing. Finally, 35 μl of the digested membrane fraction was mixed with 35 μl isotope-labeled internal standard (IS) peptide mix (10 nM of each IS; Thermo Fisher Scientific). All sample preparation and digestion steps were performed using Protein LoBind tubes (Eppendorf, Hamburg, Germany). Protein quantification was conducted on a 5500

QTRAP triple quadrupole mass spectrometer (AB Sciex) coupled to an Agilent Technologies 1260 Infinity system (Agilent Technologies) using validated LC-MS/MS methods as recently described (Drozdzik et al., 2019). Transporter proteins and the respective proteospecific peptides and the stable isotope-labeled internal standard peptides considered in our analysis are given in Supplementary Table 3. Protein abundance of human OCT1 was determined by using three peptides, whereas mouse OCT1 and Na^+/K^+ -ATPase were determined by using one peptide. For each peptide, two to three mass transitions have been monitored.

IVIVE to Estimate the Liver Partition Coefficient of Metformin. The uptake of metformin across the sinusoidal membrane into the liver was estimated based on in vitro uptake measurements in stably transfected HEK293 cells overexpressing human or mouse OCT1. In vitro clearance ($CL_{in\ vitro}$) was calculated as follows:

$$CL_{in\ vitro} = \frac{v_{max}}{K_M},$$

where v_{max} is the maximum transport rate (picomole \times minute $^{-1}$ \times milligram protein $^{-1}$) and K_M is the Michaelis constant (micromolar) determined in HEK293 cells. The obtained $CL_{in\ vitro}$ (microliter \times minute $^{-1}$ \times milligram protein $^{-1}$) was used for extrapolation toward total human or mouse liver clearance, which was mediated by active transport via human or mouse OCT1, respectively.

The active OCT1-mediated uptake into the liver ($CL_{in,act}$) was calculated as follows:

$$CL_{in,act} = CL_{in\ vitro} \times \frac{E_{in\ vivo}}{E_{in\ vitro}} \times LW \times \text{total protein per unit LW},$$

where E refers to the total OCT1 expression in human or mouse liver ($E_{in\ vivo}$) and in HEK293 cells overexpressing human or mouse OCT1 ($E_{in\ vitro}$) (picomole \times milligram protein $^{-1}$). In this case, we assumed that 100% of the OCT1 protein is localized in the plasma membrane both in the liver and in HEK293 cells. LW refers to the liver weight in human and mouse (gram), respectively. The total protein amount per unit LW is given as (milligram protein \times gram liver $^{-1}$) and was obtained from Sohlenius-Sternbeck (2006).

Passive diffusion into the liver [CL_{diff} (microliter \times minute $^{-1}$ \times milligram protein $^{-1}$)] was estimated by using HEK293 cells transfected with the empty pcDNA5/FRT vector and was calculated as follows:

$$CL_{diff} = CL_{in\ vitro} \times LW \times \text{total protein per unit LW}$$

The liver partition coefficient (K_p) was calculated based on the extended clearance concept according to Guo et al. (2018):

$$K_p = \frac{CL_{in,act} + CL_{in, diff}}{CL_{ef,act} + CL_{ef, diff} + CL_{bile} + CL_{met}}$$

The following simplifications were made based on the findings that metformin is neither metabolized nor significantly excreted by transporters or secreted into the bile (Pentikäinen et al., 1979; Tucker et al., 1981): the clearances for transporter-mediated efflux ($CL_{ef,act}$), for biliary excretion of unchanged drug (CL_{bile}), and for metabolism (CL_{met}) were set to zero. Furthermore, we assumed that the passive influx diffusion permeation ($CL_{in, diff}$) is equal to the passive efflux diffusion permeation ($CL_{ef, diff}$), here further designated simply as CL_{diff} . K_p was predicted using the following equation (Yabe et al., 2011; Shitara et al., 2013):

$$K_p = \frac{CL_{in,act} + CL_{diff}}{CL_{diff}}$$

As metformin had been shown to have negligible protein binding, the fraction unbound in plasma can be assumed as 1 (Tucker et al., 1981). The prediction of the liver partition coefficient for unbound drug concentration thereby was assumed as

$$K_{p,u} = K_p$$

Computational Modeling. Available templates for structural modeling were identified by using fold recognition methods offered by the pGenThreader server (available at <http://bioinf.cs.ucl.ac.uk/psipred/>) (Lobley et al., 2009). The human glucose 3 transporter (GLUT3; Protein Data Bank identifier (PDB ID) 4zw9) was selected as an optimal template for both human and mouse OCT1 modeling (Deng et al., 2015). This crystal structure of GLUT3 was selected based on multiple criteria, such as being resolved in high resolution (1.5 Å) and adopting an outward-occluded conformation, which is particularly useful for the investigation of substrate binding. Sequence-to-structure alignment between GLUT3 and human and mouse OCT1 sequence, respectively, was initially generated in PROMALS3D and subsequently revised and corrected by manual intervention (Pei et al., 2008). The large extracellular loop between TMH1 and TMH2 (88 residues) and one C-terminal intracellular loop (22 residues) were lacking structural templates and were therefore omitted for structural modeling purposes.

In total, 100 structural models were generated for both human and mouse OCT1 using Modeler 9.17 (Eswar et al., 2006). Energy minimization was performed to optimize the orientation of side chains. AMBER99SB-ILDN force field (Lindorff-Larsen et al., 2010) and GROMACS version 5.1.4 (Abraham et al., 2015) were used for steepest descent minimization. The convergence criterion was set to a maximum force <100.0 kJ/mol per nanometer. The final models for human and mouse OCT1 were selected on the basis of the MolProbity score ranking (<http://molprobity.biochem.duke.edu/>) and a proper orientation of the Asp474/475 residue, which is the main residue known to be implicated in ligand binding. Ramachandran outliers were visually inspected in Molecular Operating Environment 19 (Chemical Computing Group ULC, Montreal, QC, Canada). In silico models generated for human OCT1 and mouse OCT1 are available as supplement to this publication (Supplemental Files hOCT1.pdb and mOCT1.pdb).

Two independent algorithms—FTSite (Ngan et al., 2012) and SiteFinder in Molecular Operating Environment 19—were used to identify possible interaction sites in human and mouse OCT1 transporters. The FTSite program docks 16 small probe molecules to identify hot spots in the protein structure. Probe molecules are clustered, and the poses are ranked according to the empirical free energy function. The SiteFinder tool in MOE utilizes the alpha sphere method in which the protein cavities are explored by virtual spheres generated in the site. Every sphere can also differentiate with respect to potential hydrophobic or hydrophilic contacts. Predicted binding sites are ranked according to the number of alpha spheres located in every detected binding site. Standard settings of SiteFinder were applied. Hydrophobic interactions were analyzed by a helical wheel projection using DrawCoil 1.0 (available at <https://grigoryanlab.org/drawcoil/>).

Data Analyses. Kinetic parameters of metformin and thiamine transport (K_M and v_{max}) were determined by nonlinear regression to the Michaelis-Menten equation using GraphPad Prism version 5.01 (GraphPad Software Inc., La Jolla, CA). The kinetic parameters or uptake values were compared between human and mouse OCT1, human-mouse chimeric OCT1, or mutant OCT1 using ANOVA followed by Tukey's honestly significant difference post hoc comparisons in SPSS Statistics version 25 (SPSS Inc., IBM, Chicago, IL).

Results

Characterization of the Model System. In this study, we used HEK293 cells stably transfected to overexpress human and mouse OCT1 by targeted chromosomal integration. As a first step, we characterized our model system with respect to the levels of OCT1 protein expression. We used targeted proteomics to quantify OCT1 expression in the stably transfected HEK293 cells and to compare it with OCT1 expression in human and mouse liver (Fig. 1). In HEK293 cells, OCT1 expression was 36% higher in the cells stably overexpressing human compared with the cells stably overexpressing mouse OCT1 (Fig. 1A). The native OCT1 expression in human and mouse liver was similar but more than 10-fold lower than in the

model cell lines (16-fold for human OCT1, 11-fold for mouse OCT1, Fig. 1B). We also compared the OCT1 expression in three mouse strains: C57BL/6 substrains J and N and FVB. C57BL/6J mice showed more than 30% higher expression than the other two strains (Supplemental Fig. 1).

Differences in the Kinetics of Metformin Uptake between Human and Mouse OCT1. More importantly, we compared the uptake of metformin between human and mouse OCT1. To this end, we performed concentration-dependent uptake measurements in stably transfected HEK293 cells. The maximal transport rates (v_{max}) were 45% lower in human than in mouse OCT1 (v_{max} of 939 and 1353 pmol \times min⁻¹ \times pmol OCT1⁻¹, respectively). The differences in the apparent affinity (K_M) for metformin were much stronger (Fig. 2A). Mouse OCT1 showed a 4.9-fold higher affinity for metformin than human OCT1 (K_M of 491 and 2197 μ M, respectively, $P < 0.0001$). This resulted in a 6.5-fold higher intrinsic clearance of mouse compared with human OCT1 (2.86 and 0.50 μ l \times min⁻¹ \times pmol OCT1⁻¹, respectively). The strong differences in metformin kinetics were confirmed when using a transient transfection model (Supplemental Fig. 2; Supplementary Table 4). In transiently transfected HEK293 cells, we observed an 8.1-fold higher affinity of mouse OCT1 compared with human OCT1, resulting in a 12.5-fold difference in intrinsic clearance. Also, time-dependent analyses showed substantially higher uptake by mouse than by human OCT1 (Fig. 2B). The difference was strongest in the beginning of the incubation period (5.7-fold at 1 minute) but remained above 60% even after 30 minutes of incubation. There were no indications for different modes of transport between human and mouse OCT1 (Fig. 2C). After incubating with clinically relevant concentrations of 10 μ M metformin (Shu et al., 2007), we observed 4.4-fold higher intracellular concentrations in HEK293 cells overexpressing mouse than in those overexpressing human OCT1 (99.6 and 22.7 μ M, respectively, Fig. 2D).

We also determined the transport kinetics of rat OCT1 (Supplemental Fig. 3). With a K_M of 422 μ M and intrinsic clearance of 30.5 μ l \times min⁻¹ \times mg protein⁻¹, the rat ortholog did not differ significantly from mouse OCT1 (K_M of 491 μ M and intrinsic clearance of 37 μ l \times min⁻¹ \times mg protein⁻¹) but differed strongly compared with human OCT1 (K_M of 2197 μ M and intrinsic clearance of 7.85 μ l \times min⁻¹ \times mg protein⁻¹).

Estimation of the Differences in Partition Coefficients of Metformin between Human and Mouse Livers Using IVIVE. Assuming that OCT1 is the major determinant of metformin levels in the liver (Wang et al., 2002; Shu et al., 2007), it could be expected that

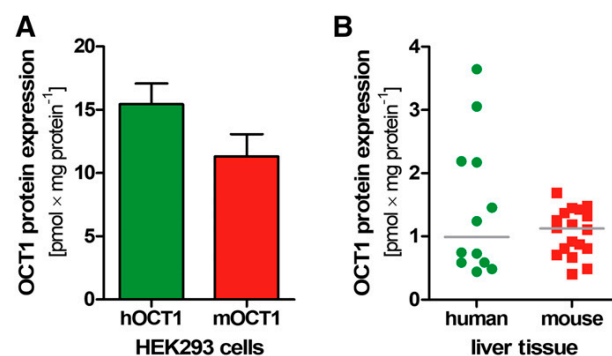


Fig. 1. OCT1 protein expression in (A) stably transfected HEK293 cells and (B) human and mouse liver. OCT1 expression in the membrane fraction of (A) HEK293 cells stably overexpressing human (green) or mouse (red) OCT1 or (B) human and mouse liver samples was measured by targeted LC-MS/MS. Please consider the difference in scaling of the y-axis. Shown are (A) means and S.E.M. of nine samples each and (B) concentrations of single samples (12 human and 20 mouse livers) and respective medians.

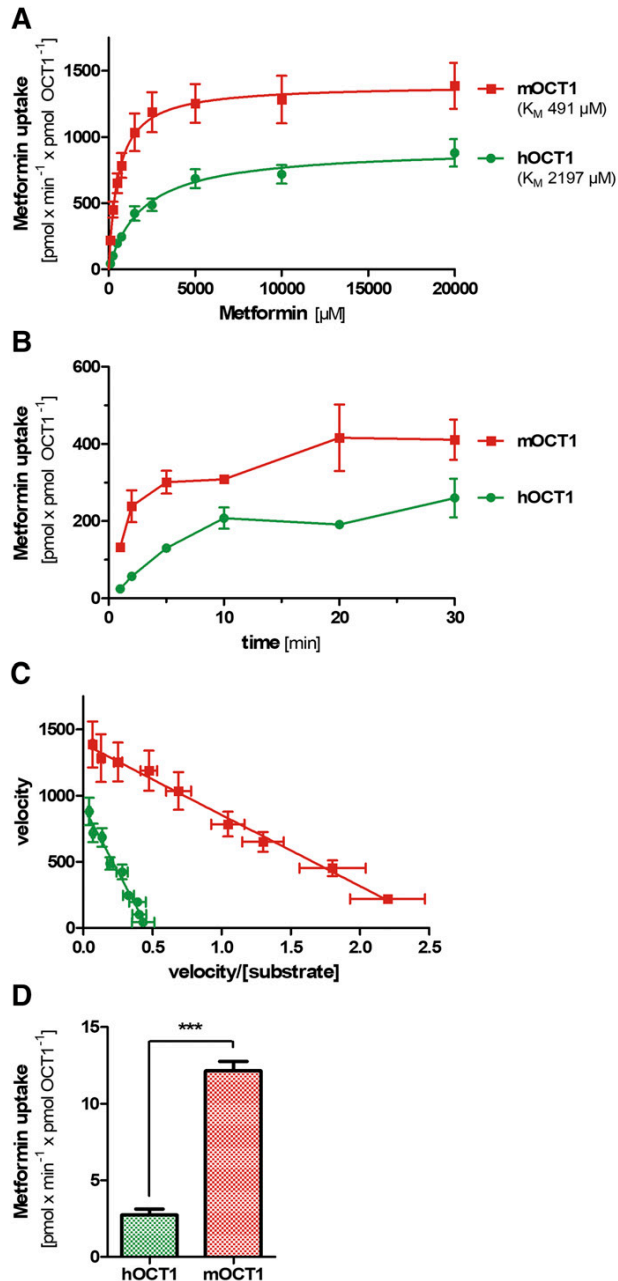


Fig. 2. Differences in metformin uptake between human and mouse OCT1. (A) Concentration-dependent uptake and (B) time-dependent uptake of metformin by human (green) and mouse (red) OCT1. OCT1-overexpressing HEK293 cells were incubated with (A) increasing concentrations of metformin for 2 minutes or (B) with 100 μM metformin for up to 30 minutes. The uptake values were normalized to the amount of OCT1 protein in the respective HEK293 cells, as determined by targeted proteomics (see Fig. 1A). (C) Eadie-Hofstee transformation of the data in (A). (D) Intracellular metformin concentrations in HEK293 cells stably transfected with human or mouse OCT1 after incubation with 10 μM metformin for 2 minutes. The intracellular concentrations were calculated assuming an intracellular volume of 1.2 μl for 1×10^6 HEK293 cells, following the estimations of Chien et al. (2016). All subfigures represent OCT1-mediated uptake that was calculated by subtracting the uptake of control cells (pcDNA5) from the uptake of cells overexpressing OCT1. Shown are means and S.E.M. of at least three independent experiments. *** $P < 0.001$ in a one-way ANOVA.

substantial differences in OCT1 clearance between human and mouse will result in substantial differences in the exposure to metformin in human and mouse liver. We used an IVIVE approach to estimate the K_p both in human and in mouse. Taking into account the differences in OCT1 expression (Fig. 1), we estimated the liver-to-blood K_p in human to be 3.34 compared with 14.4 in mouse (Table 2). Considering the estimated portal vein concentration of metformin in humans (Shu et al., 2007; Gormsen et al., 2016) and in mice (Wilcock and Bailey, 1994), we could expect 11-fold higher maximal intrahepatic concentrations in mice than in humans (66.9 μM in human and 746 μM in mouse, Table 2).

We compared the predicted values for mouse with the experimentally measured values (Wilcock and Bailey, 1994). The predicted K_p for mouse liver was almost 2-fold higher than the experimentally measured one (Table 2). In line with this, the predicted hepatic concentration in mouse was 98% higher. This suggests that our model overestimates the K_p in mice and that factors other than OCT1-mediated uptake may play a role.

To the best of our knowledge, there is no experimental data on hepatic concentrations of metformin in humans. However, using ^{11}C -labeled metformin in PET analyses, Gormsen et al. (2016) estimated a hepatic K_p of 2.5 in humans, which is 34% lower than our estimation (Table 2) and supports the differences in hepatic metformin concentrations in vivo between human and mouse that were suggested by the IVIVE model.

Identification of the Structural Causes for the Differences in Metformin Kinetics. Next, we looked for structural differences between human and mouse OCT1 that confer the differences in their affinity for metformin. To this end, we generated chimeric constructs of human and mouse OCT1 (hmoOCT1 and mmhOCT1) and characterized their metformin uptake. We separated the protein into three parts: from N terminus to the large intracellular loop, from TMH7 to TMH9, and from TMH10 to C terminus (Fig. 3A). Concentration-dependent measurements pointed to the first six TMHs of OCT1 to confer the differences in affinity for metformin between human and mouse OCT1 (Fig. 3, B and C). Immunofluorescence staining demonstrated the correct membrane localization of the wild types and chimeras and suggested a reduced total expression of the mmhOCT1 chimera (Fig. 3D) that correlates with its reduced v_{max} (Fig. 3B).

To further narrow down the region within the first six TMHs conferring these differences, we generated chimeric constructs with single TMH substitutions between human and mouse OCT1. Uptake experiments at single concentrations pointed to TMH2 and TMH3 as being primarily involved. Substituting TMH2 or TMH3 in human OCT1 with TMH2 or TMH3 of mouse OCT1 resulted in the only significant increase of metformin uptake (Fig. 4A). In line with this, substituting TMH2 or TMH3 in mouse OCT1 with TMH2 or TMH3 of human OCT1 resulted in the strongest decrease of metformin uptake (by 77% and 55%, respectively; Fig. 4B). A single concentration of 100 μM was chosen to optimally reflect the difference in the K_M based on the data from the wild-type constructs (Fig. 2). However, the effects (especially reduction of the uptake) may also be caused by a general reduction of activity (as observed for the mmhOCT1 chimera, Fig. 3B). To exclude this, we performed concentration-dependent measurements for TMH2- and TMH3-containing chimeras. We observed a strong increase in metformin affinity upon introduction of either mouse TMH2 or mouse TMH3 into human OCT1 (3-fold lower K_M compared with human OCT1, Fig. 4C). Vice versa, introduction of either human TMH2 or human TMH3 into mouse OCT1 did not significantly change affinity. However, simultaneous introduction of human TMH2 and TMH3 into mouse OCT1 resulted in a significantly decreased affinity (14-fold higher K_M compared with mouse OCT1, Fig. 4C), and the concentration-dependent uptake almost completely mimicked the uptake of human OCT1 (Fig. 4D). These experiments clearly

TABLE 2
Parameter of OCT1-mediated metformin pharmacokinetics in humans and mice measured experimentally or extrapolated using IVIVE

Parameter	Mouse					Human				
	Mean	n	S.D.	95% CI		Mean	n	S.D.	95% CI	
Maximal velocity, v_{\max} ($\text{pmol} \times \text{min}^{-1} \times \text{mg protein}^{-1}$)	17,496	11	7097	12,727	22,265	14,703	11	4346	11,783	17,623
Affinity for metformin uptake, K_M (μM)	491	11	155	387	595	2198	11	1154	1422	2973
Metformin in vitro clearance, $\text{CL}_{\text{in vitro}}$ ($\mu\text{l} \times \text{min}^{-1} \times \text{mg protein}^{-1}$)	37	11	16.1	26.2	47.9	7.85	11	3.9	5.23	10.5
Metformin passive diffusion, CL_{diff} ($\mu\text{l} \times \text{min}^{-1} \times \text{mg protein}^{-1}$) ^a	0.34	11	0.20	0.21	0.48	0.34	11	0.20	0.21	0.48
OCT1 expression in liver, $E_{\text{in vivo}}$ ($\text{pmol} \times \text{mg protein}^{-1}$)	1.27	20	0.72	0.93	1.61	1.44	12	1.09	0.75	2.13
OCT1 expression in vitro, $E_{\text{in vitro}}$ ($\text{pmol} \times \text{mg protein}^{-1}$)	11.3	9	5.23	7.3	15.3	15.4	9	4.92	11.7	19.2
Predicted metformin liver-to-blood partition coefficient, $K_{p,u}$ ^b	14.4	11	4.93	11.1	17.7	3.34	11	0.98	2.68	4.0
Predicted maximal hepatic metformin concentrations (μM) ^c	746	11	254	575	659	66.9	11	19.6	53.7	80.0
Observed metformin $K_{p,u}$ ^d	6.8					2.5				
Observed maximal hepatic metformin concentrations (μM) ^d	350									

S.D., standard deviation; n, number of independent measurements; CI, confidence interval.

^aPassive diffusion was estimated based on the uptake in control HEK293 cells transfected with the empty vector pcDNA5 (Supplemental Fig. 2A). Therefore, the values do not differ between mouse and human.

^b $K_{p,u}$ was calculated as described in *Materials and Methods*. Liver weights used for the calculations were 1500 g for human and 2.5 g for mouse (Rogers and Dintzis, 2018). The total amount of protein was 90 and 115 mg \times g liver⁻¹ for human and mouse liver, respectively (Sohlenius-Stenberg, 2006).

^cHepatic concentrations were calculated assuming portal vein concentrations of 51.7 μM for mouse (Wilcock and Bailey, 1994) and 20 μM for human [double the C_{max} observed in humans after 1 g of metformin (Shu et al., 2007; Gormsen et al., 2016)].

^dThe experimental data of mouse K_p and hepatic concentrations were obtained from Wilcock and Bailey (1994) 30 min after an oral dose of 50 mg/kg metformin. The concentrations were calculated assuming 2.5 g average weight and 1.3 ml average volume of mouse liver. The human K_p was obtained from Gormsen et al. (2016). No experimental data of human liver concentrations were available.

identify TMH2 and TMH3 to confer the differences between human and mouse OCT1 in their affinity for metformin.

Next, we generated homology models of human and of mouse OCT1 to identify single amino acids within TMH2 and TMH3 that may confer the differences in metformin uptake. There were no major differences in the tertiary structure between human and mouse OCT1 as visible by superposition of the two models (Fig. 5A). Two major binding cavities with the involvement of TMH2 or TMH3 were identified using two distinct algorithms. One of the proposed binding cavities is located in the middle of the translocation pore and is a highly populated site with probe molecules (143 alpha spheres for human OCT1 and 120 alpha spheres for mouse OCT1). This “classical” binding site has been reported in several previous studies (Chen et al., 2017; Boxberger et al., 2018; Gorboulev et al., 2018). The binding site is enframed by TMH1, TMH4, TMH5, TMH7, TMH8, TMH9, TMH11, and (more importantly) TMH2. None of the five nonconserved amino acids in TMH2 (Fig. 5B; Supplemental Fig. 4) could be suggested to be directly involved in substrate binding. However, our structural models show that Leu155 in human (hLeu155) that corresponds to Val156 in mouse OCT1 (mVal156) in TMH2 can form a hydrophobic core packing with Ile35 located in TMH1 (Fig. 6, A and B) that may have an impact on tertiary structure stability (the mouse OCT1 protein is longer than the human one, resulting in a one-count shift in amino acid position after number 84). The stronger hydrophobic interaction between Leu155 and Ile35 in human OCT1 (compared with Val156 and Ile35 in mouse OCT1) might aggravate the entrance of substrates and conformational changes in this region of the transporter. These observations might explain a generally lower affinity for metformin uptake in human OCT1 compared with mouse OCT1.

Indeed, simultaneous introduction of mouse TMH3 and mutation of Leu155Val increased the affinity of human OCT1 by more than 70% (of the difference between human and mouse OCT1, Fig. 6C), and vice versa, simultaneous introduction of human TMH3 and mutation of Val156Leu decreased metformin affinity in mouse OCT1 by 55% (Fig. 6, C and D). Without the simultaneous introduction of TMH3, the mutation of Leu155Val in human showed only limited effects, and Val156Leu in mouse OCT1 showed no significant effects. This points to the importance of the interaction with TMH3 for the effects of hLeu155/mVal156.

An alternative explanation involving both TMH2 and TMH3 may be provided by the second predicted binding site. This binding cavity

(although with a lower score, as indicated by 51 alpha spheres for human OCT1 and 25 alpha spheres for mouse OCT1; Fig. 7, A and B, left panel) is a membrane-exposed pocket with involvement of residues from both TMH2 and TMH3. Interestingly, this “outer” cavity is framed by two nonconserved residues—one from each TMH—lying just opposite of each other. Whereas in mouse OCT1, these residues are valines (Val166 and Val182), in human OCT1, these residues are glycines (Gly165 and Gly181; Fig. 7, A and B). The glycines in human OCT1 could lead to a higher conformational flexibility of the helices in that region, whereas the valines in mouse OCT1 introduce hydrophobicity and potentially stronger interactions with ligands at this position. Indeed, mutation of valines 166 and 182 in mouse OCT1 to glycines (Val166Gly, Val182Gly), either alone or in combination, significantly decreased metformin uptake (Fig. 7C). However, the decrease was maximally 26%, and the reverse mutation of glycines in human OCT1 to valines (Gly165Val, Gly181Val) neither alone nor in combination affected metformin uptake. This suggests that the glycine-to-valine differences at positions 165/166 and 181/182 alone cannot explain the differences in metformin uptake between human and mouse OCT1.

As an alternative approach, we took advantage of the observation that the affinity of human OCT2 for metformin is rather similar to the affinity of mouse OCT1 than to the affinity of human OCT1 (Supplemental Fig. 5A). We mutated amino acids that are identical in mouse OCT1 and human OCT2 but are different in human OCT1 (Supplemental Fig. 5B) and analyzed the effects on metformin uptake. In addition to hLeu155/mVal156, which we already analyzed (Fig. 6C), this affected hPhe169/mIle170 in TMH2. Interestingly, this variation is the only difference between human and mouse OCT1 within the conserved A-motif (Fig. 5B), which is suggested to be important for both structure and function of MFS transporters by interacting with residues from surrounding TMHs and supporting conformational changes during the transport cycle (Henderson and Maiden, 1990; Pao et al., 1998; Quistgaard et al., 2016). Mutation of Ile170 in mouse OCT1 to the corresponding amino acid in human OCT1 (Ile170Phe) decreased metformin uptake by 28% ($P = 4 \times 10^{-4}$, Supplemental Fig. 5C). However, mutation of both Val156Leu and Ile170Phe in mouse OCT1 did not lead to a stronger decrease in metformin uptake than mutation of Val156Leu alone (43%, Supplemental Fig. 5C). Mutation of these amino acids in human OCT1 (Leu155Val and Phe169Ile) neither alone nor in combination had an effect on metformin uptake (Supplemental Fig. 5C). Thus, hPhe169/mIle170 neither alone nor

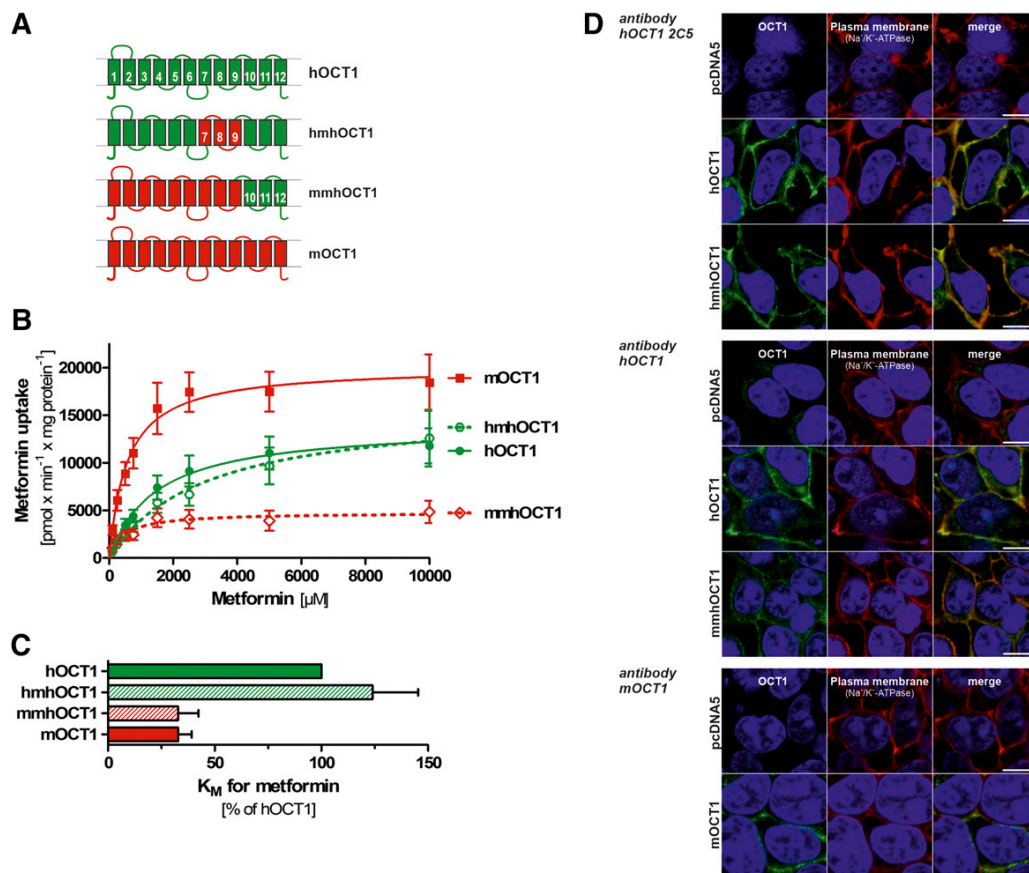


Fig. 3. Metformin uptake in human-mouse chimeric OCT1. (A) Schematic representation of human and mouse wild-type and human-mouse chimeric OCT1 constructs with numbering of the individual TMHs. Colors indicate the origin of the TMHs of either human (green) or mouse (red) OCT1. (B) Concentration-dependent uptake of metformin by human and mouse wild-type and human-mouse chimeric OCT1. OCT1-overexpressing HEK293 cells were incubated with increasing concentrations of metformin for 2 minutes. OCT1-mediated uptake was calculated by subtracting the uptake of control cells (pcDNA5) from the uptake of cells overexpressing OCT1. (C) Affinity for metformin (K_M) of human and mouse wild-type OCT1 compared with human-mouse chimeric OCT1. Represented are K_M values of the data shown in (B) as percentage of human OCT1. Shown are means and S.E.M. of at least three independent experiments. (D) Membrane localization of OCT1 as assessed by immunofluorescence staining. To enable staining of human-mouse chimeric OCT1, two different antibodies against human OCT1 were used, binding in the large intracellular loop (2C5, top panel) or in the C terminus (middle panel). The antibody against mouse OCT1 binds in the C terminus and could therefore not be used for staining chimeric OCT1. Cells were costained with an antibody against Na⁺/K⁺-ATPase as a marker for the plasma membrane. Cell nuclei were stained with DAPI. Scale bar, 10 μm .

in combination with hLeu155/mVal156 can explain more than 30% of the observed differences in metformin affinity between the species. In addition, this suggests that independent mechanisms confer the higher affinity for metformin in mouse OCT1 and in human OCT2.

Differences in the Kinetics of Thiamine Uptake between Human and Mouse OCT1. Similar to metformin, we observed strong differences in the affinity for thiamine between human and mouse OCT1 (Fig. 8). Mouse OCT1 had a 9.5-fold higher apparent affinity for thiamine than human OCT1 (K_M of 143 and 1057 μM , respectively; Fig. 8A). In contrast to metformin, the lower affinity for thiamine of human OCT1 resulted in 80% higher maximal transport rates (v_{max} of 528 and 287 $\text{pmol} \times \text{min}^{-1} \times \text{pmol OCT1}^{-1}$ for human and mouse OCT1, respectively). Nevertheless, this resulted in a 5.1-fold higher intrinsic clearance of thiamine by mouse OCT1 compared with human OCT1 (Fig. 8C).

Concentration-dependent measurements using human-mouse chimeric OCT1 also pointed to the first six TMHs of OCT1 to confer the differences in affinity for thiamine between human and mouse OCT1 (Fig. 9, A and B). Considering the observed key role of TMH2 and TMH3 in the transport of metformin, we analyzed thiamine uptake by human-mouse chimeric OCT1 carrying the simultaneous substitution of

both helices. Similar to metformin, simultaneous introduction of mouse TMH2 and TMH3 into human OCT1 resulted in a significant increase of affinity for thiamine (K_M of 456 μM compared with 1517 μM of human OCT1, Fig. 9, C and D). However, in contrast to metformin, introduction of human TMH2 and TMH3 into mouse OCT1 did not result in a significant decrease of affinity (Fig. 9C). Therefore, it could be concluded that TMH2 and TMH3 are also involved in the mechanisms conferring differences in thiamine uptake between human and mouse OCT1, but the mechanisms are not identical to the ones for metformin.

Discussion

In this study, we report strong differences between human and mouse OCT1 in the transport of metformin and thiamine. The most pronounced difference was the substantially higher apparent affinity of mouse compared with human OCT1, which results in a much higher intrinsic uptake clearance. This was observed for both metformin and thiamine. As a consequence, higher concentrations of metformin may be reached in mouse than in human liver. Furthermore, differences in the affinity for metformin between human and mouse OCT1 could be attributed to

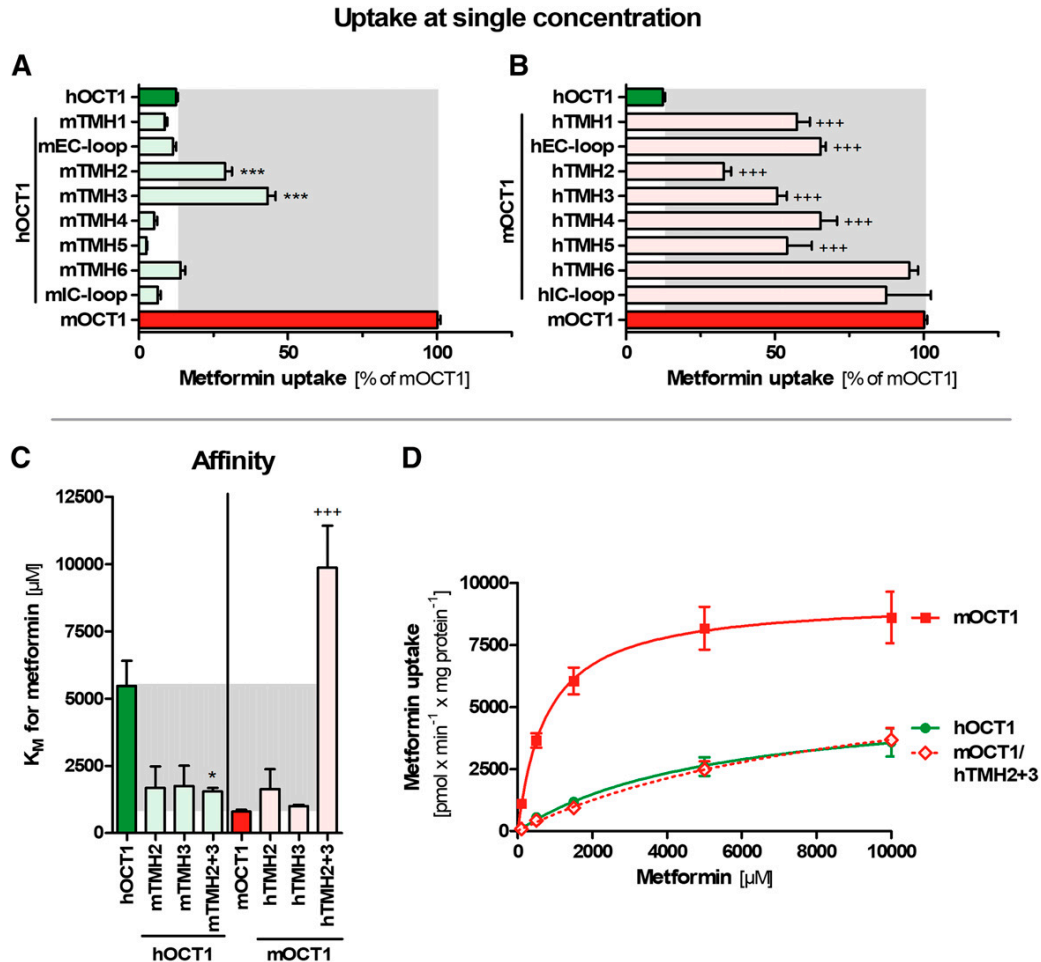


Fig. 4. Identification of TMH2 and TMH3 as major determinants of the differences in metformin uptake between human and mouse OCT1. (A and B) Human-mouse chimeric constructs with single TMH substitutions of each of the first six TMHs, the large extracellular (EC), or the large intracellular (IC) loop (human background, light green; mouse background, light red). Metformin uptake was measured at single concentrations of 100 μM metformin and related to the uptake by human (green) and mouse (red) wild-type OCT1. (C) Effects on metformin affinity (K_M) after substituting TMH2 and TMH3 alone or in combination between human and mouse OCT1. (D) Concentration-dependent metformin uptake of human (green) and mouse (red) wild-type OCT1 and mouse OCT1 with TMH2 and TMH3 of human OCT1 (red dotted line). In all cases, HEK293 cells transiently overexpressing OCT1 were incubated with metformin for 2 minutes. OCT1-mediated uptake was calculated by subtracting the uptake of control cells (pcDNA5) from the uptake of cells overexpressing OCT1. Shown are means and S.E.M. of at least three independent experiments. * $P < 0.05$; ***/ $P < 0.001$ compared with (*) human or (**) mouse OCT1 in a Tukey's post hoc analysis following one-way ANOVA.

differences in TMH2 and 3 of the transporter, revealing new insights into the transport mechanism of metformin by OCT1.

We observed a 4.9-fold higher affinity for metformin by mouse than by human OCT1 and a 45% higher transport capacity (Fig. 2). The affinity of human OCT1 for metformin observed here (K_M of 2197 μM , Fig. 2) is similar to previous reports (Shu et al., 2007; Umehara et al., 2007; Nies et al., 2009). Despite mouse being a commonly used model organism to study metformin pharmacokinetics and effects, to the best of our knowledge, this is the first study reporting metformin uptake kinetics via mouse OCT1. Moreover, we characterized metformin uptake by human and mouse OCT1 in parallel. Therefore, the obtained data are highly comparable, especially since we used a model system that is well characterized in terms of protein expression, enabling us to normalize uptake data to the amount of OCT1 protein. There are no previous reports on metformin uptake kinetics in mouse hepatocytes. However, comparison of human and rat hepatocytes showed a 27-fold higher metformin clearance in rats than in humans (Umehara et al., 2007), and

we observed highly similar uptake kinetics between rat and mouse OCT1 (Supplemental Fig. 3).

Based on our *in vitro* data on OCT1 affinity and on the differences in the portal vein concentrations, metformin concentrations can be expected to be about 11-fold higher in mouse than in human liver (Table 2). This could result in differences in the hepatic actions of metformin between human and mouse. Low metformin concentrations were suggested to activate AMPK (Zhou et al., 2001) and to suppress gluconeogenic gene expression and glucose production (Cao et al., 2014), whereas high metformin concentrations inhibit mitochondrial complex I (El-Mir et al., 2000) and lead to an AMPK-independent suppression of gluconeogenesis (Foretz et al., 2010). Especially as OCT1-overexpression has recently been shown to substantially increase the mitochondrial accumulation of the drug (Chien et al., 2016), mitochondrial effects of metformin may be more likely in mouse than in human liver. In general, our data warrants attention when using mouse data to extrapolate the hepatic effects of metformin in humans.

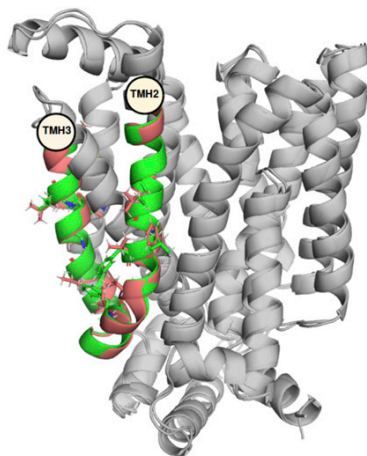
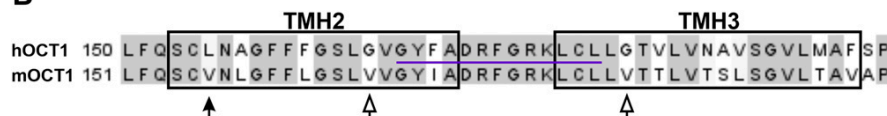
A**B**

Fig. 5. Structural differences between human and mouse OCT1 with a focus on TMH2 and TMH3. (A) Superposition of human and mouse OCT1 structural models with TMH2 and TMH3 highlighted in green (human OCT1) and red (mouse OCT1). (B) Protein sequence alignment of human and mouse OCT1 using EMBOSS Needle (Madeira et al., 2019) with TMH2 and TMH3 highlighted and the conserved A-motif of MFS transporters underlined in violet. Coloring is based on amino acid identity. Arrows indicate the positions of amino acids hLeu155/mVal156 (closed arrow) and hGly165/mVal166 and hGly181/mGly182 (open arrows), which were of particular interest.

Because of the better uptake by mouse OCT1, the liver-mediated effects may be more pronounced in mice than in humans.

The effects of OCT1 deficiency are more pronounced in mice than in humans (Wang et al., 2002, 2003; Shu et al., 2007; Zhou et al., 2009; Dujic et al., 2017; Sundelin et al., 2017). A 30-fold decrease in hepatic metformin concentrations was shown in OCT1 knockout mice (Wang et al., 2002). Precise measurements in humans are difficult, but a study using PET imaging showed about 2-fold lower hepatic metformin concentrations in carriers of loss-of-function OCT1 variants (Sundelin et al., 2017). These numbers generally fit to the tendency observed here that OCT1-mediated uptake is about 14-fold higher compared with diffusion in mice and only 3-fold higher compared with diffusion in humans (Table 2). One explanation, supported by our data, is that because of the different efficacy of human and mouse OCT1 in transporting metformin, knockout in mice and loss-of-function genetic variants in humans do not have comparable effects on hepatic metformin concentrations. However, the fact that some of the human OCT1 genetic variants do not lead to a complete loss of metformin uptake should also be considered (Kerb et al., 2002; Shu et al., 2003; Seitz et al., 2015).

Our IVIVE calculations based on the uptake kinetics of mouse OCT1 overestimated the hepatic K_p and hepatic concentrations of metformin experimentally measured in mice (Table 2). The most probable reason for this is that the major reflection of the higher uptake clearance via mouse OCT1 is observed in the first minutes of the uptake (Fig. 2B), and the available experimental data are obtained after 30 minutes or more (Wilcock and Bailey, 1994). In the longer incubation, other factors, like reaching steady state of intracellular versus extracellular metformin concentrations, may play a role. Therefore, short-term differences in the concentrations of metformin between human and mouse liver may be more pronounced than the long-term differences. Nevertheless, accounting for the PET-based estimation of K_p of 2.5 in humans and a maximal portal vein concentration of 20 μM (Shu et al., 2007; Gormsen et al., 2016), an intrahepatic concentration of about 50 μM could be estimated for the human liver. This is 7-fold lower than the intrahepatic concentrations of metformin measured in mice (Wilcock and Bailey, 1994).

The involvement of alternative transporters like OCT3 that are not reflected in our IVIVE calculations is less probable. In the mouse liver,

OCT1 has much higher expression levels than OCT3 [OCT1-to-OCT3 mRNA ratio of about 30 (Chen et al., 2015)]. Consistently, up to 30-fold lower hepatic metformin concentrations were measured in OCT1 knockout mice (Wang et al., 2002), but there were no significant changes in OCT3 knockout mice (Chen et al., 2015; Lee et al., 2018). Also, in the human liver, OCT1 is expressed much more strongly than OCT3. The ratio of OCT1 to OCT3 in human liver is 22 based on protein quantification (Drozdik et al., 2019) and 32 based on mRNA quantification (Nies et al., 2009). The intrinsic clearance of metformin can thus be estimated to be at least 11-fold lower by OCT3 than by OCT1, indicating that OCT1 is the predominant uptake transporter of metformin both in the mouse and in the human liver.

Similar to metformin, the affinity for thiamine was much higher by mouse than by human OCT1 (9.5-fold lower K_M , Fig. 8), which is supported by a previous study (Chen et al., 2014). Also, clear effects of OCT1 deficiency on thiamine levels were reported in mice (Chen et al., 2014; Liang et al., 2018) but not in humans (Jensen et al., 2020). One explanation may be that in humans, at “physiological” low concentrations, thiamine is predominantly transported by thiamine transporters THTR-1 and THTR-2, which have a substantially higher affinity (>1600 -fold lower K_M) but also a substantially lower capacity (>130 -fold lower v_{\max}) than OCT1 (Jensen et al., 2020). In contrast, in mice, the differences in affinity and capacity of thiamine uptake between OCT1 and THTR-1 are much smaller (2.9-fold lower K_M and 7.6-fold lower v_{\max}), and mouse OCT1 showed more than 5-fold higher thiamine uptake than mouse THTR-1 at low concentrations [100 nM; (Chen et al., 2014)]. This, together with the much higher affinity of mouse compared with human OCT1, suggests that OCT1 may play a more important role in thiamine uptake at low concentrations in mice than in humans and thereby may contribute to the different effects of OCT1 deficiency on thiamine plasma levels between these species. Alternatively, the strong renal OCT1 expression in mouse but not in human (Gorboulev et al., 1997; Zhang et al., 1997; Green et al., 1999; Schmitt et al., 2003) may contribute to the different effects of OCT1 deficiency on systemic thiamine concentrations in the two species. However, compensatory effects of OCT2, which is strongly expressed in the kidney, are probable (Chen et al., 2014).

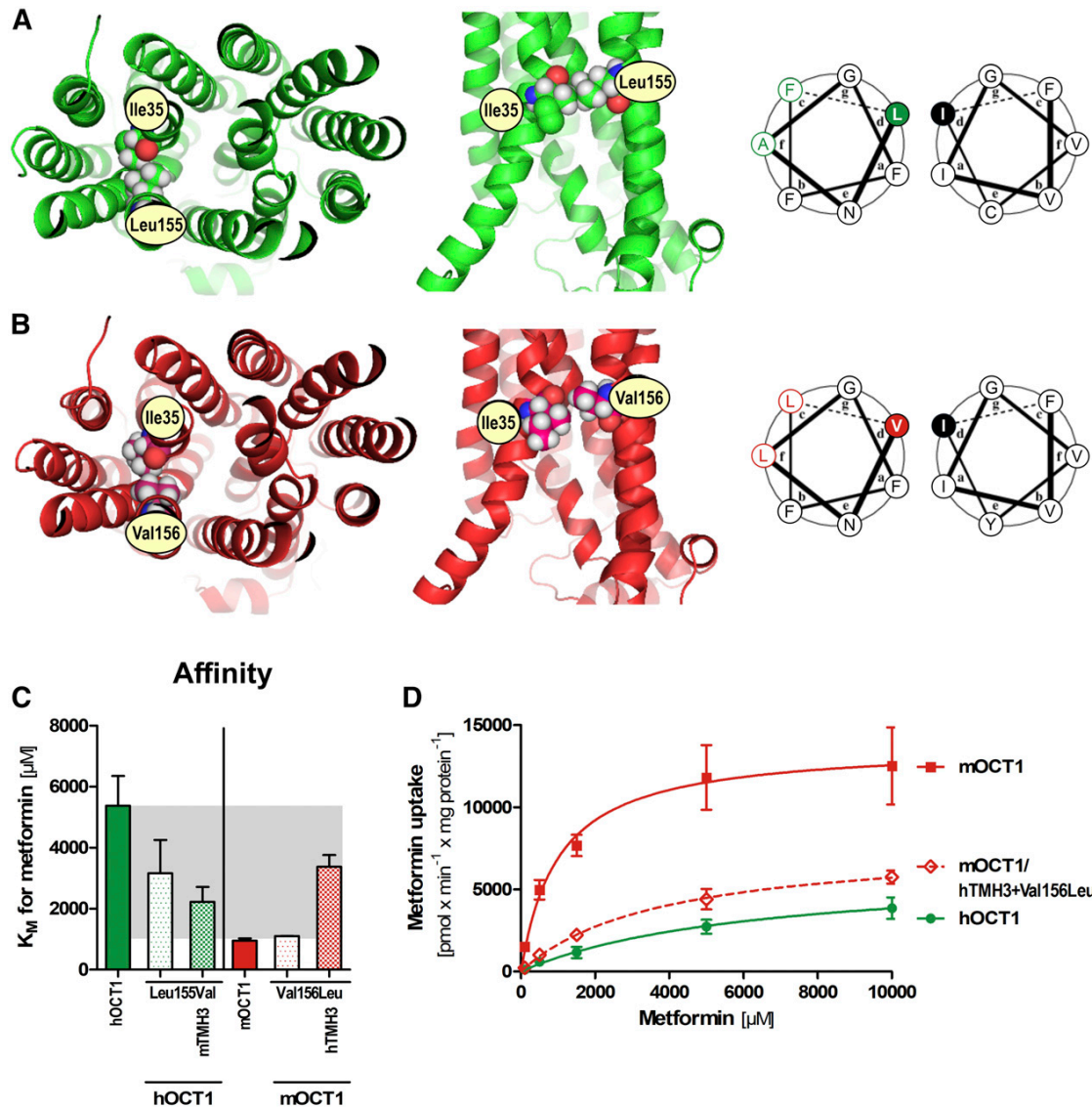


Fig. 6. Potential involvement of hLeu155/mVal156 in TMH2 of human/mouse OCT1 in conferring the differences in metformin affinity. Hydrophobic interactions between hLeu155/mVal156 and Ile35 in (A) human and (B) mouse OCT1 in (left panel) top view and (middle panel) side view. (A and B, right panel) Helical wheel projection of TMH2 and TMH1 showing the positioning of hLeu155/mVal156 (TMH2) and Ile35 (TMH1) in position “d” of the helical wheels, respectively. Nonconserved amino acids are highlighted in color, and Ile35 is highlighted in black. (C) Effects on metformin affinity (K_M) after simultaneous substitution of Leu155Val and mouse TMH3 in human OCT1 and Val156Leu and human TMH3 in mouse OCT1. (D) Concentration-dependent metformin uptake of human (green) and mouse (red) wild-type OCT1 and mouse OCT1 with Val156Leu mutation and human TMH3 (red dotted line). In all cases, HEK293 cells transiently overexpressing OCT1 were incubated with metformin for 2 minutes. OCT1-mediated uptake was calculated by subtracting the uptake of control cells (pcDNA5) from the uptake of cells overexpressing OCT1. Shown are means and S.E.M. of four independent experiments. * $P < 0.05$, ***/**** $P < 0.001$ compared with (*) human or (*) mouse OCT1 in a Tukey’s post hoc analysis following one-way ANOVA.

Transmembrane helices TMH2 and TMH3 were experimentally identified to confer the differences in metformin affinity between human and mouse OCT1. Three independent hypotheses were generated to identify the single amino acids responsible: 1) differences in the hydrophobic interaction of hLeu155/mVal156 (TMH2) with Ile35 (TMH1), 2) higher flexibility by Gly165 and Gly181 in human as opposed to higher hydrophobicity and ligand interaction by Val166 and Val182 in mouse OCT1, and 3) similar affinities of mouse OCT1 and human OCT2 caused by Val156 and/or Ile170 that differ in human OCT1 (Ile170 being located within the conserved

A-motif of the MFS transporters). Experimentally, the strongest effects were observed by mutating Val156 to Leu together with exchanging TMH3 (Fig. 6), supporting the first hypothesis the most. The mechanisms may be expected to be similar for rat OCT1, since the potentially involved amino acids are identical between rat and mouse OCT1 (Supplemental Fig. 3D).

Another interesting observation is that the substitution of a single TMH (either 2 or 3) is sufficient to increase the affinity of human OCT1, but both human TMHs are needed to decrease the affinity of mouse OCT1 (Fig. 4C). This may suggest that a decreased affinity requires an

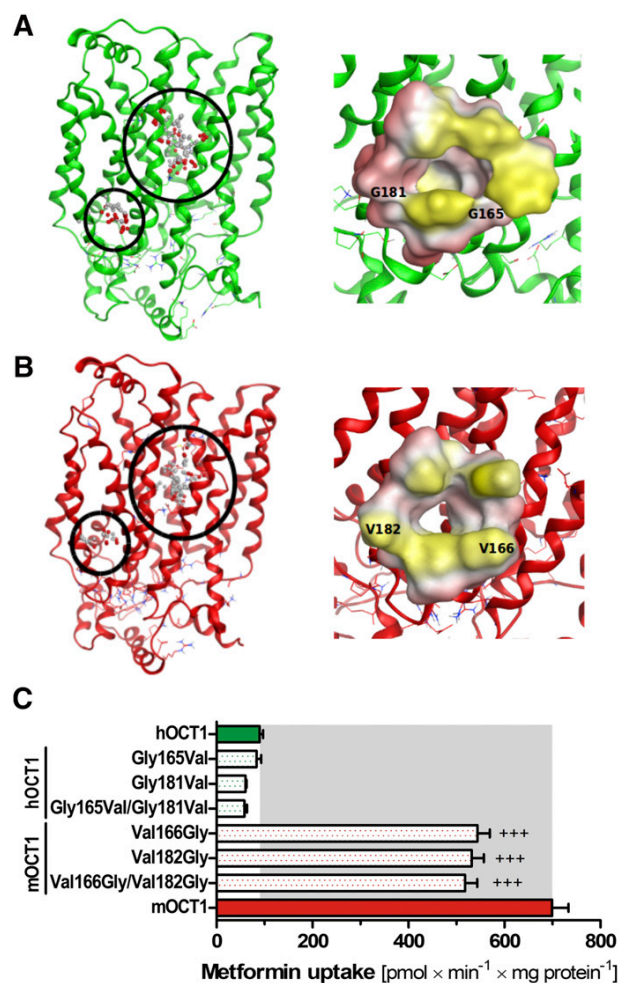


Fig. 7. Potential involvement of TMH2 and TMH3 in conferring the differences in metformin affinity between human and mouse OCT1. (A and B, left panel) Two predicted binding cavities within the TMH2-TMH3 region in (A) human and (B) mouse OCT1 structural models. (A and B, right panel) Top view of the “inner” binding cavity with protein surface colored according to lipophilicity (red, hydrophilic; yellow, hydrophobic; white, neutral surface) with (A) glycine residues 165 and 181 in human OCT1 and (B) valine residues 166 and 182 in mouse OCT1 highlighted. (C) Effect of mutations of Gly165Val and Gly181Val in human and Val166Gly and Val182Gly in mouse OCT1 on metformin uptake. HEK293 cells transiently overexpressing OCT1 were incubated with 100 μ M metformin for 2 minutes. OCT1-mediated uptake was calculated by subtracting the uptake of control cells (pcDNA5) from the uptake of cells overexpressing OCT1. Shown are means and S.E.M. of two to four independent experiments performed in duplicates. *** $P < 0.001$ compared with mouse OCT1 in a Tukey’s post hoc analysis following one-way ANOVA.

interaction between the two TMHs and removing one of the TMHs is enough to destroy this interaction.

From the structural perspective, the “knob-into-hole” motif of the hydrophobic interaction between hLeu155/mVal156 in TMH2 with Ile35 in TMH1 (Fig. 6) is somewhat analogous to coiled-coil structures (Liu et al., 2006). Interestingly, hLeu155/mVal156 are located at position “d” when depicting the helix as a helical wheel projection (Fig. 6). Since this position is suggested to be more vulnerable to amino acid substitution (Zhu et al., 1993), hLeu155/mVal156 may have a huge impact on tertiary structure stability. The stronger hydrophobic interaction between Leu155 and Ile35 in human OCT1 (compared with Val156 and Ile35 in mouse OCT1) might obstruct substrate entry and conformational

changes in this region, thereby potentially explaining the lower affinity for metformin.

Another aspect to be considered is the substrate-specific effects of OCT1. Recently, Morse et al. (2020) reported substantial differences in the uptake kinetics between human and mouse primary hepatocytes for ondansetron and tropisetron but not for sumatriptan and fenoterol. In our study, the differences in the affinity between human and mouse OCT1 were comparable between metformin and thiamine (Figs. 2 and 8). This is consistent with previous reports suggesting similar binding sites of metformin and thiamine in OCT1 (Chen et al., 2014). However, our data show that the affinities for these two compounds are conferred by similar, but not identical, structures in OCT1 (Figs. 4 and 9). This underlines the polyspecificity of OCT1 and points out that structure-to-function relations of OCT1 need to be established separately for each substrate and, based on the present results, also for each species.

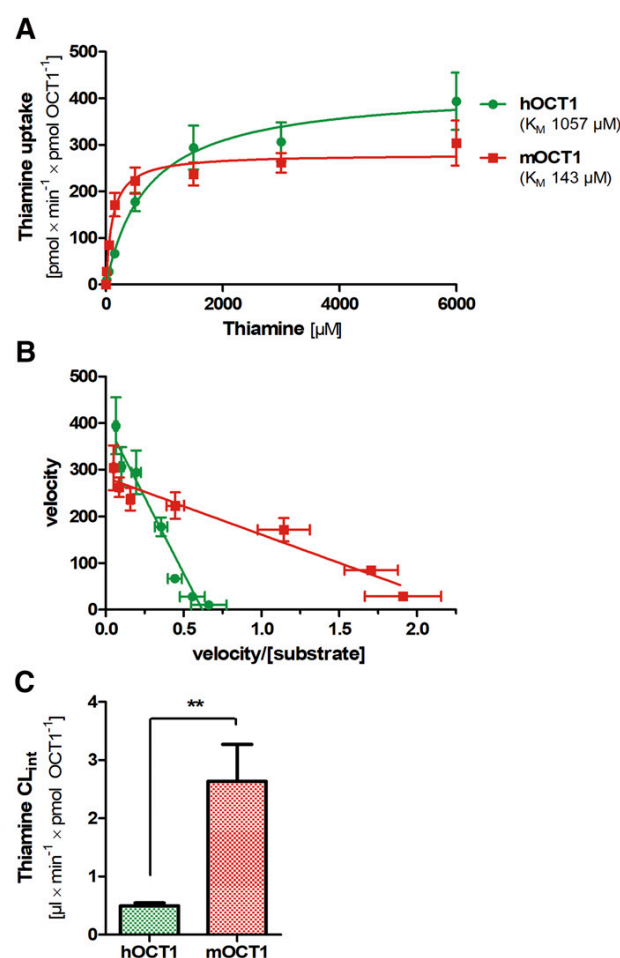


Fig. 8. Differences in thiamine uptake between human and mouse OCT1. (A) Concentration-dependent uptake of thiamine by human (green) and mouse (red) OCT1. OCT1-overexpressing HEK293 cells were incubated with increasing concentrations of thiamine for 2 minutes. OCT1-mediated uptake was calculated by subtracting the uptake of control cells (pcDNA5) from the uptake of cells overexpressing OCT1. Background thiamine levels were subtracted from all values to exclude influence of endogenous thiamine on the measurement. The uptake values were normalized to the amount of OCT1 protein in the respective HEK293 cells, as determined by targeted proteomics (see Fig. 1A). (B) Eadie-Hofstee transformation of the data in (A). (C) Comparison of the intrinsic clearance (CL_{int}) between human and mouse OCT1 calculated using v_{\max} and K_M of the data in (A). Shown are means and S.E.M. of at least three independent experiments. ** $P < 0.01$ in a one-way ANOVA.

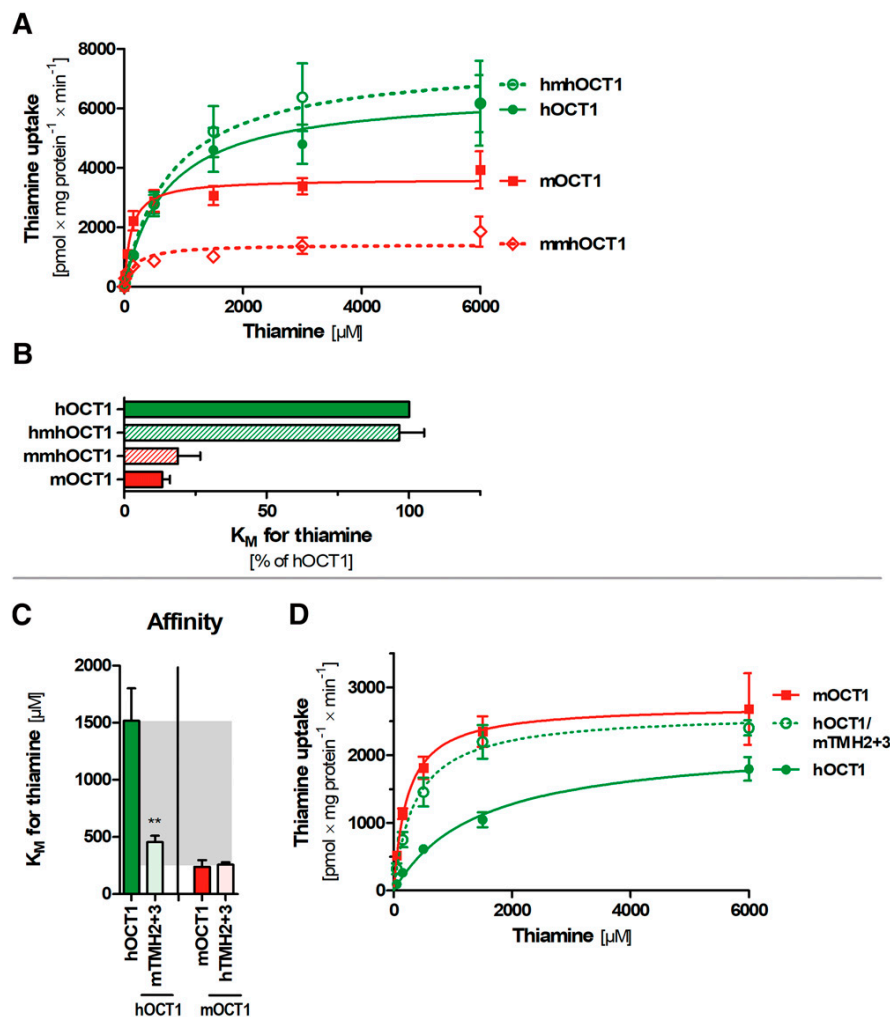


Fig. 9. Thiamine uptake in human-mouse chimeric OCT1 constructs. (A) Concentration-dependent uptake of thiamine by human and mouse wild-type and human-mouse chimeric OCT1. For detailed description of chimeric constructs see figure 3A. (B) Affinity for thiamine (K_M) of human and mouse wild-type compared with human-mouse chimeric OCT1. Represented are K_M values of the data shown in (A) as percentage of human OCT1. (C) Affinity for thiamine (K_M) of human and mouse wild-type OCT1 compared with human-mouse chimeric OCT1 carrying the simultaneous substitution of TMH2 and TMH3. (D) Concentration-dependent thiamine uptake of human (green) and mouse (red) wild-type OCT1 and human OCT1 with TMH2 and TMH3 of mouse OCT1 (green dotted line). HEK293 cells (A and B) stably or (C and D) transiently overexpressing OCT1 were incubated with increasing concentrations of thiamine for 2 minutes. OCT1-mediated uptake was calculated by subtracting the uptake of control cells (pcDNA5) from the uptake of cells overexpressing OCT1. Background thiamine levels were subtracted from all values to exclude the influence of endogenous thiamine on the measurement. Shown are means and S.E.M. of at least three independent experiments. $**P < 0.01$ in a Tukey's post hoc analysis following one-way ANOVA.

In conclusion, mouse OCT1 has a much higher affinity for metformin and thiamine than human OCT1. This may be an important factor contributing to the substantially higher metformin concentrations measured in the mouse than in the human liver in vivo and should be considered when interpreting findings about the hepatic mechanism of action of metformin that are obtained in mouse models. The determinants of the differences in metformin affinity between human and mouse OCT1 are clearly located in TMH2 and TMH3 and comprise hLeu155/mVal156 (TMH2) and amino acid(s) in TMH3. The underlying mechanism is probably complex, and the identification of the precise amino acids in TMH3 and additional protein structures in that region involved needs further investigation.

Acknowledgments

The authors would like to acknowledge Helen Massy for her contribution to the initial cloning of mouse and rat OCT1 orthologs and the generation of the first chimeric constructs. We also highly acknowledge the technical support of Kerstin Schmidt and Tina Sonnenberger (Greifswald) in the uptake measurements and cloning and Cornelia Willnow (Göttingen) for support in collecting the human liver samples.

Authorship Contributions

Participated in research design: Meyer, Brockmöller, Zdravil, Tzvetkov.

Conducted experiments: Meyer, Tuerkova, Wenzel.

Contributed new reagents or analytic tools: Tuerkova, Römer, Seitz, Gaedcke, Zdravil.

Performed data analysis: Meyer, Tuerkova, Römer, Wenzel, Oswald, Zdravil, Tzvetkov.

Wrote or contributed to the writing of the manuscript: Meyer, Tuerkova, Römer, Oswald, Brockmöller, Zdravil, Tzvetkov.

References

- Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, Hess B, and Lindahl E (2015) GROMACS: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* 1–2:19–25 DOI: 10.1016/j.softx.2015.06.001.
- 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 DOI: 10.1016/j.bcp.2018.08.028.
- Cao J, Meng S, Chang E, Beckwith-Fickas K, Xiong L, Cole RN, Radovick S, Wondisford FE, and He L (2014) Low concentrations of metformin suppress glucose production in hepatocytes through AMP-activated protein kinase (AMPK). *J Biol Chem* 289:20435–20446 DOI: 10.1074/jbc.M114.567271.
- Chen EC, Khuri N, Liang X, Stecula A, Chien H-C, Yee SW, Huang Y, Sali A, and Giacomini KM (2017) Discovery of competitive and noncompetitive ligands of the organic cation transporter 1 (OCT1; SLC22A1). *J Med Chem* 60:2685–2696 DOI: 10.1021/acs.jmedchem.6b01317.
- Chen EC, Liang X, Yee SW, Geier EG, Stocker SL, Chen L, and Giacomini KM (2015) Targeted disruption of organic cation transporter 3 attenuates the pharmacologic response to metformin. *Mol Pharmacol* 88:75–83 DOI: 10.1124/mol.114.096776.
- Chen L, Shu Y, Liang X, Chen EC, Yee SW, Zur AA, Li S, Xu L, Keshari KR, Lin MJ, et al. (2014) OCT1 is a high-capacity thiamine transporter that regulates hepatic steatosis and is a target of metformin. *Proc Natl Acad Sci USA* 111:9983–9988 DOI: 10.1073/pnas.1314939111.

- Chien H-C, Zur AA, Maurer TS, Yee SW, Tolsma J, Jasper P, Scott DO, and Giacomini KM (2016) Rapid method to determine intracellular drug concentrations in cellular uptake assays: application to metformin in organic cation transporter 1-transfected human embryonic kidney 293 cells. *Drug Metab Dispos* **44**:356–364 DOI: 10.1124/dmd.115.066647.
- Deng D, Sun P, Yan C, Ke M, Jiang X, Xiong L, Ren W, Hirata K, Yamamoto M, Fan S, et al. (2015) Molecular basis of ligand recognition and transport by glucose transporters. *Nature* **526**: 391–396 DOI: 10.1038/nature14655.
- Drozdzik M, Busch D, Lapczuk J, Müller J, Ostrowski M, Kurzawski M, and Oswald S (2019) Protein abundance of clinically relevant drug transporters in the human liver and intestine: a comparative analysis in paired tissue specimens. *Clin Pharmacol Ther* **105**:1204–1212 DOI: 10.1002/cpt.1301.
- Duijic T, Zhou K, Yee SW, van Leeuwen N, de Keyser CE, Javorský M, Goswami S, Zaharenko L, Hougaard Christensen MM, Out M, et al. (2017) Variants in pharmacokinetic transporters and glycemic response to metformin: a metagen meta-analysis. *Clin Pharmacol Ther* **101**:763–772 DOI: 10.1002/cpt.567.
- El-Mir MY, Nogueira V, Fontaine E, Avéret N, Rigoulet M, and Leverve X (2000) Dimethylbiguanide inhibits cell respiration via an indirect effect targeted on the respiratory chain complex I. *J Biol Chem* **275**:223–228 DOI: 10.1074/jbc.275.1.223.
- Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, Shen M-Y, Pieper U, and Sali A (2006) Comparative protein structure modeling using modeller. *Curr Protoc Bioinformatics* **Chapter 5**:Unit-5.6 DOI: 10.1002/0471250953.b0506615.
- Foretz M, Hébrard S, Leclerc J, Zarrinpashneh E, Soty M, Mithieux G, Sakamoto K, Andreelli F, and Viollet B (2010) Metformin inhibits hepatic gluconeogenesis in mice independently of the LKB1/AMPK pathway via a decrease in hepatic energy state. *J Clin Invest* **120**:2355–2369 DOI: 10.1172/JCI40671.
- Gorboulev V, Rehman S, Albert CM, Roth U, Meyer MJ, Tzvetkov MV, Mueller TD, and Koepsell H (2018) Assay conditions influence affinities of rat organic cation transporter 1: analysis of mutagenesis in the modeled outward-facing cleft by measuring effects of substrates and inhibitors on initial uptake. *Mol Pharmacol* **93**:402–415 DOI: 10.1124/mol.117.110767.
- Gorboulev V, Ulzheimer JC, Akhondova A, Ulzheimer-Teuber I, Karbach U, Queser S, Baumann C, Lang F, Busch AE, and Koepsell H (1997) Cloning and characterization of two human polyspecific organic cation transporters. *DNA Cell Biol* **16**:871–881 DOI: 10.1089/dna.1997.16.871.
- Gormsen LC, Sundelin EI, Jensen JB, Vendelbo MH, Jakobsen S, Munk OL, Hougaard Christensen MM, Brøsen K, Frøkiær J, and Jessen N (2016) In vivo imaging of human 11C-metformin in peripheral organs: dosimetry, biodistribution, and kinetic analyses. *J Nucl Med* **57**: 1920–1926 DOI: 10.2967/jnumed.116.177774.
- Green RM, Lo K, Sterritt C, and Beier DR (1999) Cloning and functional expression of a mouse liver organic cation transporter. *Hepatology* **29**:1556–1562 DOI: 10.1002/hep.510290530.
- Guo Y, Chu X, Parrott NJ, Brouwer KLR, Hsu V, Nagar S, Matsson P, Sharma P, Snoeyns J, Sugiyama Y, et al.; International Transporter Consortium (2018) Advancing predictions of tissue and intracellular drug concentrations using in vitro, imaging and physiologically based pharmacokinetic modeling approaches. *Clin Pharmacol Ther* **104**:865–889 DOI: 10.1002/cpt.1183.
- Henderson PJ and Maiden MC (1990) Homologous sugar transport proteins in *Escherichia coli* and their relatives in both prokaryotes and eukaryotes. *Philos Trans R Soc Lond B Biol Sci* **326**: 391–410 DOI: 10.1098/rstb.1990.0020.
- Horton RM, Hunt HD, Ho SN, Pullen JK, and Pease LR (1989) Engineering hybrid genes without the use of restriction enzymes: gene splicing by overlap extension. *Gene* **77**:61–68 DOI: 10.1016/0378-1119(89)90359-4.
- Jensen O, Matthaei J, Blome F, Schwab M, Tzvetkov MV, and Brockmöller J (2020) Variability and heritability of thiamine pharmacokinetics with focus on OCT1 effects on membrane transport and pharmacokinetics in humans. *Clin Pharmacol Ther* **107**:628–638 DOI: 10.1002/cpt.1666.
- Kerb R, Brinkmann U, Chatskaya N, Gorbunov D, Gorboulev V, Mornhinweg E, Keil A, Eichelbaum M, and Koepsell H (2002) Identification of genetic variations of the human organic cation transporter hOCT1 and their functional consequences. *Pharmacogenetics* **12**:591–595 DOI: 10.1097/00008571-200211000-00002.
- Lee N, Hebert MF, Wagner DJ, Easterling TR, Liang CJ, Rice K, and Wang J (2018) Organic cation transporter 3 facilitates fetal exposure to metformin during pregnancy. *Mol Pharmacol* **94**: 1125–1131 DOI: 10.1124/mol.118.112482.
- Liang X, Yee SW, Chien H-C, Chen EC, Luo Q, Zou L, Piao M, Mifune A, Chen L, Calvert ME, et al. (2018) Organic cation transporter 1 (OCT1) modulates multiple cardiometabolic traits through effects on hepatic thiamine content. *PLoS Biol* **16**:e2002907 DOI: 10.1371/journal.pbio.2002907.
- Lindorff-Larsen K, Piana S, Palmo K, Maragakis P, Klepeis JL, Dror RO, and Shaw DE (2010) Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **78**: 1950–1958 DOI: 10.1002/prot.22711.
- Liu J, Zheng Q, Deng Y, Cheng C-S, Kallenbach NR, and Lu M (2006) A seven-helix coiled coil. *Proc Natl Acad Sci USA* **103**:15457–15462 DOI: 10.1073/pnas.0604871103.
- Lobley A, Sadowski ML, and Jones DT (2009) pGenTHREADER and pDomTHREADER: new methods for improved protein fold recognition and superfamily discrimination. *Bioinformatics* **25**:1761–1767 DOI: 10.1093/bioinformatics/btp302.
- Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, Basutkar P, Tivey ARN, Potter SC, Finn RD, et al. (2019) The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res* **47**:W636–W641 DOI: 10.1093/nar/gkz268.
- Morse BL, Kolar A, Hudson LR, Hogan AT, Chen LH, Brackman RM, Sawada GA, Fallon JK, Smith PC, and Hillgren KM (2020) Pharmacokinetics of organic cation transporter 1 (OCT1) substrates in Oct1/2 knockout mice and species difference in hepatic OCT1-mediated uptake. *Drug Metab Dispos* **48**:93–105 DOI: 10.1124/dmd.119.088781.
- Ngan C-H, Hall DR, Zerbe B, Grove LE, Kozakov D, and Vajda S (2012) FTSite: high accuracy detection of ligand binding sites on unbound protein structures. *Bioinformatics* **28**:286–287 DOI: 10.1093/bioinformatics/btr651.
- Nies AT, Koepsell H, Winter S, Burk O, Klein K, Kerb R, Zanger UM, Keppler D, Schwab M, and Schaeffeler E (2009) Expression of organic cation transporters OCT1 (SLC22A1) and OCT3 (SLC22A3) is affected by genetic factors and cholestasis in human liver. *Hepatology* **50**: 1227–1240 DOI: 10.1002/hep.23103.
- Pao SS, Paulsen IT, and Saier MH Jr (1998) Major facilitator superfamily. *Microbiol Mol Biol Rev* **62**:1–34.
- Pei J, Tang M, and Grishin NV (2008) PROMALS3D web server for accurate multiple protein sequence and structure alignments. *Nucleic Acids Res* **36**:W30–W34 DOI: 10.1093/nar/gkn322.
- Pentikäinen PJ, Neuvonen PJ, and Penttilä A (1979) Pharmacokinetics of metformin after intravenous and oral administration to man. *Eur J Clin Pharmacol* **16**:195–202 DOI: 10.1007/bf00562061.
- Quistgaard EM, Löw C, Guettou F, and Nordlund P (2016) Understanding transport by the major facilitator superfamily (MFS): structures pave the way. *Nat Rev Mol Cell Biol* **17**:123–132 DOI: 10.1038/nrm.2015.25.
- Rena G, Hardie DG, and Pearson ER (2017) The mechanisms of action of metformin. *Diabetologia* **60**:1577–1585 DOI: 10.1007/s00125-017-4342-z.
- Rogers AB and Dintzis RZ (2018) Hepatobiliary system, in *Comparative Anatomy and Histology: A Mouse, Rat, and Human Atlas*, 2nd ed (Treuting PM, Dintzis SM, and Montine KS eds) pp 229–239, Academic Press, London.
- Rueden CT, Schindelin J, Hiner MC, DeZonia BE, Walter AE, Arena ET, and Eliceiri KW (2017) ImageJ2: ImageJ for the next generation of scientific image data. *BMC Bioinformatics* **18**:529 DOI: 10.1186/s12859-017-1934-z.
- Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, Preibisch S, Rueden C, Saalfeld S, Schmid B, et al. (2012) Fiji: an open-source platform for biological-image analysis. *Nat Methods* **9**:676–682 DOI: 10.1038/nmeth.2019.
- Schmitt A, Mössner R, Gossmann A, Fischer IG, Gorboulev V, Murphy DL, Koepsell H, and Lesch KP (2003) Organic cation transporter capable of transporting serotonin is up-regulated in serotonin transporter-deficient mice. *J Neurosci Res* **71**:701–709 DOI: 10.1002/jnr.10521.
- Seitz T, Stalman MK, Dalila N, Chen J, Pojar S, Dos Santos Pereira JN, Krätzner R, Brockmöller J, and Tzvetkov MV (2015) Global genetic analyses reveal strong inter-ethnic variability in the loss of activity of the organic cation transporter OCT1. *Genome Med* **7**:56 DOI: 10.1186/s13073-015-0172-0.
- Shitara Y, Maeda K, Ikejiri K, Yoshida K, Horie T, and Sugiyama Y (2013) Clinical significance of organic anion transporting polypeptides (OATPs) in drug disposition: their roles in hepatic clearance and intestinal absorption. *Biopharm Drug Dispos* **34**:45–78 DOI: 10.1002/bdd.1823.
- Shu Y, Leabman MK, Feng B, Mangravite LM, Huang CC, Stryke D, Kawamoto M, Johns SJ, DeYoung J, Carlson E, et al.; Pharmacogenetics Of Membrane Transporters Investigators (2003) Evolutionary conservation predicts function of variants of the human organic cation transporter, OCT1. *Proc Natl Acad Sci USA* **100**:5902–5907 DOI: 10.1073/pnas.0730858100.
- Shu Y, Sheardown SA, Brown C, Owen RP, Zhang S, Castro RA, Ianculescu AG, Yue L, Lo JC, Burchard EG, et al. (2007) Effect of genetic variation in the organic cation transporter 1 (OCT1) on metformin action. *J Clin Invest* **117**:1422–1431 DOI: 10.1172/JCI30558.
- Smith PK, Krohn RI, Hermanson GT, Mallia AK, Gartner FH, Provenzano MD, Fujimoto EK, Goeke NM, Olson BJ, and Klenk DC (1985) Measurement of protein using bicinchoninic acid. *Anal Biochem* **150**:76–85 DOI: 10.1016/0003-2697(85)90442-7.
- Sohlenius-Sternbeck A-K (2006) Determination of the hepatocellularity number for human, dog, rabbit, rat and mouse livers from protein concentration measurements. *Toxicol In Vitro* **20**: 1582–1586 DOI: 10.1016/j.tiv.2006.06.003.
- Sundelin E, Gormsen LC, Jensen JB, Vendelbo MH, Jakobsen S, Munk OL, Christensen M, Brøsen K, Frøkiær J, and Jessen N (2017) Genetic polymorphisms in organic cation transporter 1 attenuates hepatic metformin exposure in humans. *Clin Pharmacol Ther* **102**:841–848 DOI: 10.1002/cpt.701.
- Tucker GT, Casey C, Phillips PJ, Connor H, Ward JD, and Woods HF (1981) Metformin kinetics in healthy subjects and in patients with diabetes mellitus. *Br J Clin Pharmacol* **12**:235–246 DOI: 10.1111/j.1365-2125.1981.tb01206.x.
- Tzvetkov MV, Saadatmand AR, Bokelmann K, Meineke I, Kaiser R, and Brockmöller J (2012) Effects of OCT1 polymorphisms on the cellular uptake, plasma concentrations and efficacy of the 5-HT(3) antagonists tropisetron and ondansetron. *Pharmacogenomics J* **12**:22–29 DOI: 10.1038/tpj.2010.75.
- Umehara K-I, Iwatsubo T, Noguchi K, and Kamimura H (2007) Functional involvement of organic cation transporter1 (OCT1/Oct1) in the hepatic uptake of organic cations in humans and rats. *Xenobiotica* **37**:818–831 DOI: 10.1080/00498250701546012.
- Wang D-S, Jonker JW, Kato Y, Kusuhara H, Schinkel AH, and Sugiyama Y (2002) Involvement of organic cation transporter 1 in hepatic and intestinal distribution of metformin. *J Pharmacol Exp Ther* **302**:510–515 DOI: 10.1124/jpet.102.034140.
- Wang D-S, Kusuhara H, Kato Y, Jonker JW, Schinkel AH, and Sugiyama Y (2003) Involvement of organic cation transporter 1 in the lactic acidosis caused by metformin. *Mol Pharmacol* **63**: 844–848 DOI: 10.1124/mol.63.4.844.
- Wilcock C and Bailey CJ (1994) Accumulation of metformin by tissues of the normal and diabetic mouse. *Xenobiotica* **24**:49–57 DOI: 10.3109/00498259409043220.
- Yabe Y, Galetin A, and Houston JB (2011) Kinetic characterization of rat hepatic uptake of 16 actively transported drugs. *Drug Metab Dispos* **39**:1808–1814 DOI: 10.1124/dmd.111.040477.
- Zhang L, Dresser MJ, Gray AT, Yost SC, Terashita S, and Giacomini KM (1997) Cloning and functional expression of a human liver organic cation transporter. *Mol Pharmacol* **51**:913–921 DOI: 10.1124/mol.51.6.913.
- Zhou G, Myers R, Li Y, Chen Y, Shen X, Fenyl-Melody J, Wu M, Ventre J, Doebber T, Fujii N, et al. (2001) Role of AMP-activated protein kinase in mechanism of metformin action. *J Clin Invest* **108**:1167–1174 DOI: 10.1172/JCI13505.
- Zhou K, Donnelly LA, Kimber CH, Donnan PT, Doney ASF, Leese G, Hattersley AT, McCarthy MI, Morris AD, Palmer CNA, et al. (2009) Reduced-function SLC22A1 polymorphisms encoding organic cation transporter 1 and glycemic response to metformin: a GoDARTS study. *Diabetes* **58**:1434–1439 DOI: 10.2337/db08-0896.
- Zhu BY, Zhou NE, Kay CM, and Hodges RS (1993) Packing and hydrophobicity effects on protein folding and stability: effects of beta-branched amino acids, valine and isoleucine, on the formation and stability of two-stranded alpha-helical coiled coils/leucine zippers. *Protein Sci* **2**: 383–394 DOI: 10.1002/pro.5560020310.

Address correspondence to: Dr. Mladen V. Tzvetkov, Institute of Pharmacology, Center of Drug Absorption and Transport, University Medicine Greifswald, Felix-Hausdorff-Str. 3, 17489 Greifswald, Germany. E-mail: mladen.tzvetkov@med.uni-greifswald.de

3.5 Data-driven Ensemble Docking to Unravel Interactions of Steroid Analogs with Hepatic Organic Anion Transporting Polypeptides

TUERKOVA, Alzbeta; UNGVÁRI, Orsolya; MERNYÁK, Erzsébet; SZAKÁCS, Gergely; ÖZVEGY-LACZKA, Csilla; ZDRAZIL, Barbara. **2020** *In preparation*

**Corresponding author: barbara.zdrazil@univie.ac.at*

A. Tuerkova created the structure-based modeling pipeline, generated structural models, performed molecular docking, established binding mode hypotheses, and wrote original draft. C. Laczka, O. Ungvári, and E. Mernyák performed transport inhibition experiments with 13-epiestrones and determined IC_{50} values. B. Zdrazil conceived the study and provided supervision for the structure-based modeling. The manuscript was written by the contribution of all authors.

The Supplementary Information can be found in Part V.

Research Article

Data-driven Ensemble Docking to Unravel Interactions of Steroid Analogs with Hepatic Organic Anion Transporting Polypeptides

Alzbeta Tuerkova^a, Orsolya Ungvári^b, Erzsébet Mernyák^b, Gergely Szakács^c, Csilla Özvegy-Laczka^b, Barbara Zdrazil^{a,*}

a University of Vienna, Department of Pharmaceutical Chemistry, Division of Drug Design and Medicinal Chemistry, Althanstraße 14, A-1090 Vienna, Austria

b Membrane Protein Research Group, Institute of Enzymology, RCNS, H-1117, Budapest, Magyar tudósok krt. 2, Hungary

c Department of Medicine I, Institute of Cancer Research, Comprehensive Cancer Center, Medical University of Vienna, Vienna, Austria

* Corresponding author.

E-mail address: barbara.zdrazil@univie.ac.at

Abstract

Hepatic Organic Anion Transporting Polypeptides - OATP1B1, OATP1B3, and OATP2B1 - are collectively expressed at the basolateral membrane of the hepatocytes, being responsible for uptake of a wide range of natural substrates and structurally unrelated pharmaceuticals. Impaired function of hepatic OATPs can be linked to clinically relevant drug-drug interactions, which can lead to the development of drug-induced liver injury. Therefore, understanding the commonalities and differences across the three transporters represents useful knowledge to guide the drug discovery process at an early stage. Unfortunately, such efforts are challenging due to the lack of experimentally-resolved protein structures for any OATP member. In this study, we established a rigorous computational protocol to generate and validate structural models for hepatic OATPs. The pipeline represents a multistep

procedure, involving the systematic exploration of available protein structures with shared protein fold using normal mode analysis, calculation of multiple template backbones from elastic network models, utilization of multiple template conformers to generate OATP structural models with various degrees of conformational flexibility, and prioritization of models on the basis of enrichment docking. In this study, the final OATP models for OATP1B1, OATP1B3, and OATP2B1 have been used to elucidate binding modes of steroid analogs in the three transporters. Both, data sources from the open domain as well as an in house compound data set with measured bioactivities for the three hepatic OATPs have been investigated in this study. Important structural determinants conferring shared and distinct binding patterns of steroid analogs in the three transporters have been identified. Overall, this comparative study provides novel insights into hepatic OATP-ligand interactions and selectivity. Furthermore, the integrative computational workflow for structure-based modeling can be leveraged for other pharmaceutical targets of interest.

List of Abbreviations

ANM, Anisotropic Network Model; DHEA, Dehydroepiandrosterone; DSSP, Define Secondary Structure of Proteins; E-3-S, Estrone-3-sulfate; FucP, Fucose transporter; GNM, Gaussian Network Model; MFS, Major Facilitator Superfamily; NMA, Normal Mode Analysis; OATP, Organic Anion Transporting Polypeptide; PLIF, Protein-Ligand-Interaction Fingerprint; SLC, SoLute Carrier; TMH, TransMembrane Helix

1. Introduction

Solute carriers (SLC) are being increasingly recognized for their pivotal role in compound pharmacokinetics, given their involvement in drug absorption, disposition, metabolism, elimination, clinically relevant drug-drug interactions, and related organ toxicities. (1,2) Here, we focus on a triad of Organic Anion Transporting Polypeptides - OATP1B1 (*SLCO1B1* gene), OATP1B3 (*SLCO1B3* gene), and OATP2B1 (*SLCO2B1* gene) - which are belonging to the SLCO (SLC21) superfamily. (3,4) These proteins are collectively expressed at the basolateral membrane of hepatocytes, mediating cellular uptake of a broad spectrum of endogenous substrates and xenobiotics. (5) Endogenous compounds include bilirubin, bile acids, steroid conjugates, and hormones. Drugs transported by hepatic OATPs are structurally and functionally quite heterogeneous, such as statins (pitavastatin, rosuvastatin, fluvastatin) (6), antihistamines (fexofenadine) (7), anticancer agents (SN-38, paclitaxel, imatinib) (8), antibiotics (rifampicin, clarithromycin, benzylpenicillin) (9), or anti-inflammatory drugs (ibuprofen, diclofenac, lumiracoxib) (10). OATP-mediated drug-drug interactions represent a challenge for drug development. Therefore, U.S. Food and Drug Administration recommends to test novel drug candidates for their potential interaction with hepatic OATPs. Computational prediction of whether a certain drug might interact with hepatic OATPs is a promising approach at the early stage in the drug discovery pipeline to minimize the risk of attrition. However, these efforts become challenging, mainly due to the lack of experimental protein structures for any OATP member. In general, the sequence identity of hepatic OATPs to the closest structural analogs does not exceed 16%, which makes the generation of the high-quality structural models a non-trivial task. Performing homology modeling as a function of sequence gets error-prone below 30% sequence identity (referred to as ‘twilight zone’ protein modeling). (11) Therefore, the majority of computational studies done for hepatic OATPs were ligand-based, including different QSAR models (12–17), proteochemometric models (12,18), ligand-based pharmacophore models (19), and substructure analyses (20).

OATPs are glycoproteins with 643-722 amino acids. Hydropathy analysis shows twelve transmembrane helices (TMHs) interconnected by intracellular and extracellular loops. (21) They were reported to possess Major Facilitator Superfamily (MFS) fold. (22) MFS proteins contain multiple binding sites that are capable of recognizing structurally unrelated compounds. A large extracellular loop between TMH9 and TMH10 contains eleven cysteine

residues which form disulfide bonds, resembling the kazal-type domain of serine protease inhibitors. (23) Other important structural features are the N-glycosylation sites in the extracellular loops 2 and 5,(4) phosphorylation sites at the N- and C-terminus, (24) and the consensus sequence region spanning extracellular loop 3 and TMH6 region. (25) The first structural models for OATP1B3 and OATP2B1 date back to 2005. (22) Based on the comparison of OATP1B3 and OATP2B1 structures, ARG181 was suggested to contribute to OATP1B substrate specificity, while HIS579 was suggested as a structural determinant conferring OATP2B specificity. Mandery et al published newer structural models for OATP1B3 and OATP2B1 in 2011. (26) Comparative analysis revealed that LYS361 and LYS399 are highly conserved across the OATP family, where LYS361 is pointing towards the translocation pore. In another study, mutagenesis experiments supplemented by structural model generation showed that several residues at THM2 of OATP1B1 (ASP70, PHE73, GLU74, GLY76, ASN77) are implicated in transport function.(27) Moreover, two distinct binding sites (low- and high-affinity binding site) for estrone-3-sulfate (E-3-S) were identified in that study. Another structure-based modeling study combined with alanine-scanning experiments identified the importance of TMH11 for OATP1B1 ligand uptake.(28) Further, Glaeser et al used structure-based modeling supported by experimental validation to show the importance of a positive charge at position 41 and 580 for OATP1B3 transport function.(29) Gui and Hagenbuch created a 3D structure of OATP1B3 and identified several crucial residues at TMH10 (TYR537, SER545, and THR550), which were subsequently confirmed via site-directed mutagenesis. In a recent study by Khuri et al., a combination of structure-based modeling and machine learning approaches was used to identify novel OATP2B1 inhibitors. (30)

In this paper, we present an integrative computational pipeline for retrieving high quality structural models for OATP1B1, OATP1B3, and OATP2B1. The models have subsequently been used to elucidate binding modes for steroid analogs highlighting commonalities and differences between the three transporters. To the best of our knowledge, this is the first comparative structure-based modeling study including all three hepatic OATPs.

Molecular determinants contributing to OATP-steroid interactions (and selectivity) have been analyzed in detail in our previous paper. (20) Here, we are focusing on structural

determinants as revealed by molecular docking into the developed transporter protein models, revealing distinct binding sites and ligand-transporter interactions. The computational findings delivered here have been prospectively validated by comparison to published mutagenesis data and known single nucleotide polymorphisms. Structural models generated herein can be leveraged for, e.g., virtual screening purposes to identify novel OATP tool compounds. An in house data set of new 13-epiestrones with measured IC_{50} values on OATP1B1, OATP1B3, and OATP2B1 has been used to augment the compound data set from public sources with compounds showing a pronounced activity for OATP2B1 (which is to date the least studied transporter of the three hepatic OATPs studied herein). (31) A special focus of the developed pipeline is put on ensemble docking. Inclusion of protein conformational flexibility shall increase confidence about the usability of the structural models for docking the respective highly active compounds in the data set. Applying ensemble docking in this study was motivated by the assumption that selectivity of SLC transporters might not be exclusively modulated by sequence variability, but also by conformational flexibility. (32) Therefore, the final structural models were prioritized according to their ability to enrich known actives among a pool of known inactives and decoys. In order to create a representative subset of OATP conformations, normal mode analysis (NMA) was applied for available MFS structures to detect soft modes of motion which cover conformational diversity of MFS transporters (so called “signature dynamics”). (33) The pipeline presented here is versatile enough to be reproduced for other protein targets of interest and was exclusively built upon freely available tools and software (pGenThreader, Modeller, ProDy, Phenix, GROMACS, AutoDock Vina, PyMol, KNIME) which enables full adaptability and reusability.

2. Materials & Methods

2.1. Comparative modeling of hepatic OATPs

Structural templates were detected by using the fold-recognition tool(34) pGenThreader (default settings). (35)

The fucose transporter in an outward-open conformation (FucP, PDB ID 3o7q, 3.14 Å

resolution) (36) possessing Major Facilitator Superfamily (MFS) fold was identified as a high-quality template for OATP1B1 (p-value ≤ 0.0001 , prediction score 75, 14.5% sequence identity), OATP1B3 (p-value ≤ 0.0001 , prediction score 75, 15.6% sequence identity), and OATP2B1 (p-value ≤ 0.0001 , prediction score 93, 15.2% sequence identity), respectively (see Table 1). Except for the high prediction score for all the three transporters, FucP template was selected due to the reasonable crystal resolution (3.14 Å), and an outward-open conformation state, which appears advantageous for studying ligand recognition. PROMALS3D was used to generate multiple structure-to-sequence alignments between FucP and all human OATPs(37). The generated alignment was subjected to manual adjustments. A pairwise template-to-sequence alignment is available as Supplementary File S1 (OATP1B1-FucP), Supplementary File S2 (OATP1B3-FucP), and Supplementary File S3 (OATP2B1-FucP). Due to the lack of structural templates for extra- and intra-cellular domains, respectively, our models cover the transmembrane region only. Amino acid residue numbers in the FucP template used for comparative modeling are the following: 22-56, 60-114, 115-177, 200-239, 242-290, 292-409, 412-434. Corresponding regions in OATP1B1, OATP1B3, and OATP2B1, respectively, are listed in the Supplementary Table S7.

Table 1

Five top ranked templates predicted for (A) OATP1B1, (B) OATP1B3, (C) OATP2B1.

(A)

Net Score	p-value	Alignment length	Template length	Target length	PDB ID
82.116	3e-07	402	409	691	6e9n
79.538	5e-07	414	453	691	3wdo
78.253	6e-07	426	465	691	6e8j
75.504	1e-06	430	434	691	3o7q
73.959	2e-06	404	414	691	4av3

(B)

Net Score	p-value	Alignment length	Template length	Target length	PDB ID
74.704	1e-06	400	409	702	6e9n
74.551	2e-06	403	414	702	3o7q
70.557	4e-06	428	434	702	1pw4
69.415	5e-06	202.0	404	702	6e8j
69.372	5e-06	236.0	432	702	1wa5

(C)

Net Score	p-value	Alignment length	Template length	Target length	PDB ID
93.148	2e-08	401	414	709	3o7q
86.484	9e-08	419	453	709	3wdo
84.431	2e-07	434	465	709	6e8j
83.100	2e-07	399	409	709	6e9n
83.039	2e-07	433	434	709	1pw4

Enrichment docking into an ensemble of OATP conformations was conducted to prioritize the best model per transporter. Multiple OATP structures with various degrees of global (i.e.,

backbone conformer) and local (i.e., side-chain rotamer) flexibility were modeled. A similar strategy to the one by Carlsson et al was adopted by performing NMA on the template structure. (38) The modeling protocol introduced by Carlsson et al has been expanded to perform more rigorous sampling of the protein conformational space. First, anisotropic network models (ANM) were calculated for all the available experimental structures possessing MFS fold to identify dominant motions within the whole protein family (so called “signature dynamics”, for details see Subsection 2.1.1.). Second, alternate conformations for FucP were calculated by including the implicit membrane model (see Subsection 2.1.2.). NMA calculations done here were performed by using the ProDy software (freely available at <http://prody.csb.pitt.edu/>). (39)

2.1.1 Signature dynamics of Major Facilitator Superfamily proteins

The FucP template (PDB ID: 3o7q) was selected as a reference query for the retrieval of structurally analogous proteins by using the Dali server. (40) In total, 92 protein structures with shared fold were identified in the Protein Data Bank (PDB). Sequence identity between the structural analogs and the FucP structure was set to 10% to reduce the large pool of detected protein structures to a manageable amount. After data reduction, 45 PDB structures were retained in the structural ensemble. 20 Gaussian Network Models (GNMs) modes per every protein structure were calculated at C α carbon resolution. Calculated GNM were analyzed with respect to commonalities and differences in the mode shapes, shared covariance between residues, cross-correlations, as well as mean square fluctuations. In addition, the similarity of protein structures in the structural ensemble was evaluated on the basis of their sequence (Hamming distance), 3D structure (RMSD), and their intrinsic dynamics (arccosine function of the covariance overlap). Specifically, the lowest frequency mode for each protein structure was used for this comparative analysis.

2.1.2. Conformational sampling of FucP template

In this study, an implicit membrane model was incorporated into the ANM calculations. (41) The restoring force for any protein displacement was set to be 16-times greater in x- or

y-direction than in the z-direction. The scaling factor is applied here to preferentially restrain radial motions. Such defined restraints aim to mimic the constraints imposed by the membrane on the conformational dynamics of membrane proteins. Boundaries for the implicit membrane effect have been set to a distance of ± 15.35 Å from the membrane core, as predicted by the OPM server. (42)

In order to prevent non-physical distortions and/or bond stretching, individual residues were coarse-grained into pre-defined rigid blocks. (43) Here, an assignment of residues into rigid blocks was done on the basis of the hydrogen-bond estimation (DSSP) algorithm. (44) Rigid block decomposition led to 135 blocks. 1,000 alternate conformations were sampled along the two lowest frequency modes. In the next step, sampled conformers were refined by performing energy minimization in GROMACS 5.1.4. (45) using the steepest descent algorithm in the AMBER99SB-ILDN force field. (46) The convergence criterion was set to a maximum force < 100.0 kJ/mol/nm. In order to reduce the large pool of conformational ensembles while preserving variance, only conformations with a cut-off distance of 3 Å from the average were kept resulting in 20 distinct conformers used for OATP structural modeling.

2.1.3. Construction of OATP structural models in different conformations

On the basis of the 20 FucP template conformers, 60 distinct models per transporter were calculated, following these consecutive steps:

(1) 20 different template conformers with an average RMSD of 3 Å were selected from the template conformational ensemble (counting 1,000 conformers in total, see previous Section).

(2) 100 comparative models per distinct template conformer were generated resulting in 2,000 different models.

(3) N- and C- termini in helix breaks were acetylated (shortcut 'ACE' in Modeller 9.17) and methylamidated (shortcut 'CT3' in Modeller 9.17). Energy minimization of the comparative models was performed in GROMACS using the same settings as described in section 2.1.2.

(4) Minimized models were ranked on the basis of the MolProbity score calculated by using the Phenix software.(47,48) Only three top-ranked rotamers per distinct conformer were retained for enrichment docking, resulting in 60 distinct models per transporter that were used for enrichment docking.

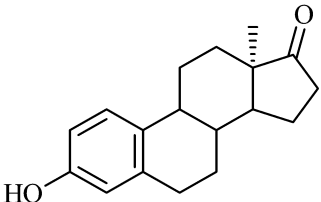
2.2. Retrieving ligand-protein interactions by molecular docking

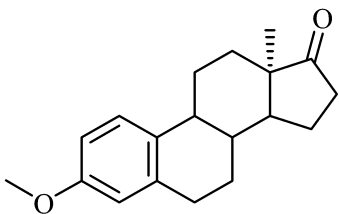
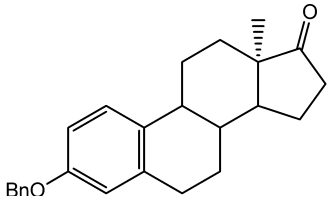
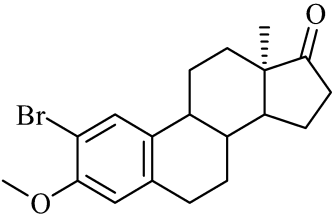
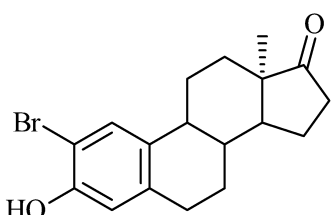
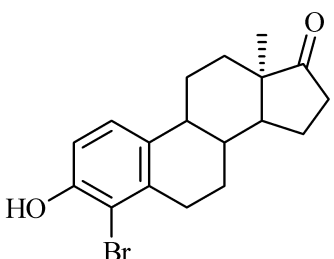
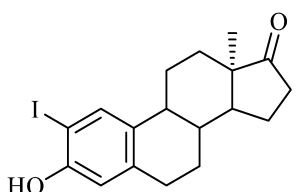
2.2.1 Preparation of the docking library

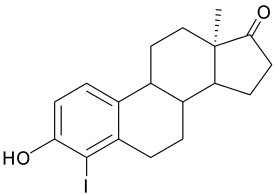
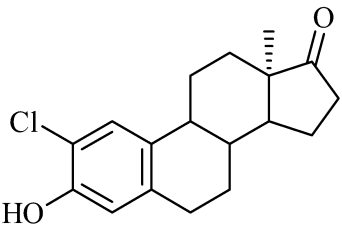
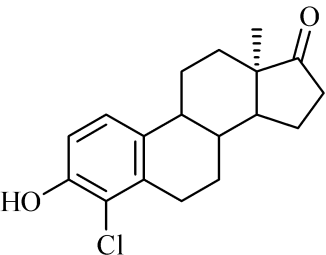
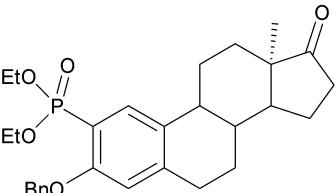
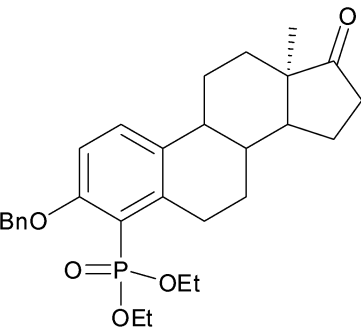
Data sets for OATP1B1, OATP1B3, and OATP2B1 were collected from the open domain as described in Tuerkova et al.(20), and from our recently published paper reporting about 13-epiestrones.(49) Experimental measurements for sixteen 13-epiestrones have been extended in this study in order to report activity on all three transporters. 13-epiestrones used in this study do possess two major variations: (1) phosphonation and (2) halogenation at either the R-2 or R-4 position. The R-3 position is composed of a hydroxyl, methoxy, or benzyloxy moiety (see Table 2).

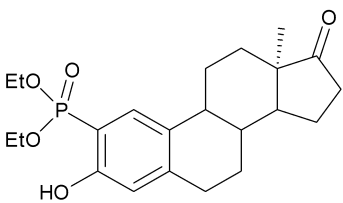
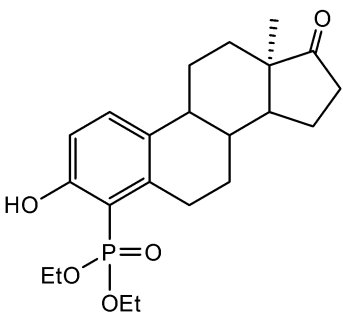
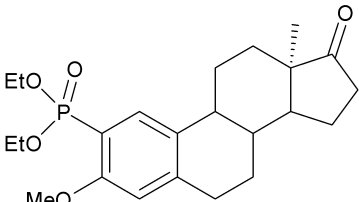
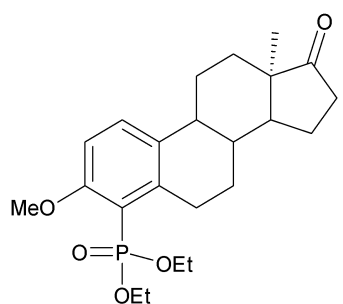
Table 2

In house data set of steroid analogs used to characterize binding sites of OATP1B1, OATP1B3, and OATP2B1. Chemical structures, and bioactivity values (IC₅₀ in MicroM measured for the respective transporter are given.

Code	Structure	IC ₅₀ (μM)		
		OATP1B1	OATP1B3	OATP2B1
1		>50	>50	>50

2		>50	>50	5.41
3		>50	>50	7.17
4		>50	>50	8.39
5		>50	>50	1.19
6		4.07	9.20	2.97
7		22.52	18.16	0.6

8		9.28	8.28	3.58
9		>50	>50	0.90
10		24.99	>50	10.45
11		0.76	2.18	0.18
12		3.22	2.49	0.75

13		3.79	5.24	3.18
14		>50	>50	10.24
15		10.12	9.5	2.96
16		32.39	11.85	2.74

For the first two rounds of docking which served for extracting the best protein model per transporter, all compounds with bioactivity measurements for the three transporters were used - independent of their core molecular scaffold. The only exception was the in house data set of 13-epiestrones, which was held out for enrichment docking, because it was used as a validation set later in the study. By including compounds with diverse core scaffolds an unbiased selection of the “best” model (according to ligand enrichment) can be guaranteed.

In order to obtain the recommended ratio of actives:inactives(/decoys) of 1:36 the data sets were enriched by decoys from DUD-E.(50) A compound was defined as active if the bioactivity value was $\leq 1 \mu\text{M}$ in case of OATP1B1 and OATP1B3 and $\leq 5 \mu\text{M}$ in case of OATP2B1 (due to a smaller amount of data), and inactive if the activity value was $> 10 \mu\text{M}$. If a compound occurred in different protonation states at pH 7.0 (± 2.0), each protomer was considered as a separate compound for docking. The resulting docking library consisted of 57 actives, 917 inactives, and 1,223 decoys for OATP1B1, 25 actives, 900 inactives, and no decoys for OATP1B3, and 12 actives, 153 inactives, and 279 decoys in case of OATP2B1. Ligand conformers were generated by the LigPrep tool in Maestro (version 19-1; OPLS3e force field). Ionization states were generated at target pH 7.0 ± 2.0 (Epik algorithm in Maestro).

2.2.2. Enrichment docking for model prioritization

Compounds were docked into the top 60 models per transporter by using the program Autodock Vina 1.1.2. (51) Exhaustiveness of the global search was set to 10. Ligand enrichment was calculated in R-3.4.2 (available at <https://www.r-project.org/>). The area under the curve (AUC) and the enrichment factor (EF) at the top 1% of the data set was calculated as a metric for ligand enrichment. In the first round of docking calculations, the entire transmembrane region (encompassing all 12 TMHs) was defined as a putative binding site. Models were ranked on the basis of AUC values. Five top-ranked models per transporter were retained for further inspection. AUC values for the pre-selected models can be found in Supplementary Table S3. For performing the second round of docking calculations, the putative protein binding site was further restricted. Specifically, the contact surface area between the active compounds docked in the first round was calculated and the calculated region was used as search space in the second round of enrichment docking calculations. The procedure is visualized in Supplementary Figure S3. The top final model was selected on the basis of the ranking of both AUC and EF 1% of the data set.

The stepwise computational procedure for structural model generation is visually depicted in Figure 1.

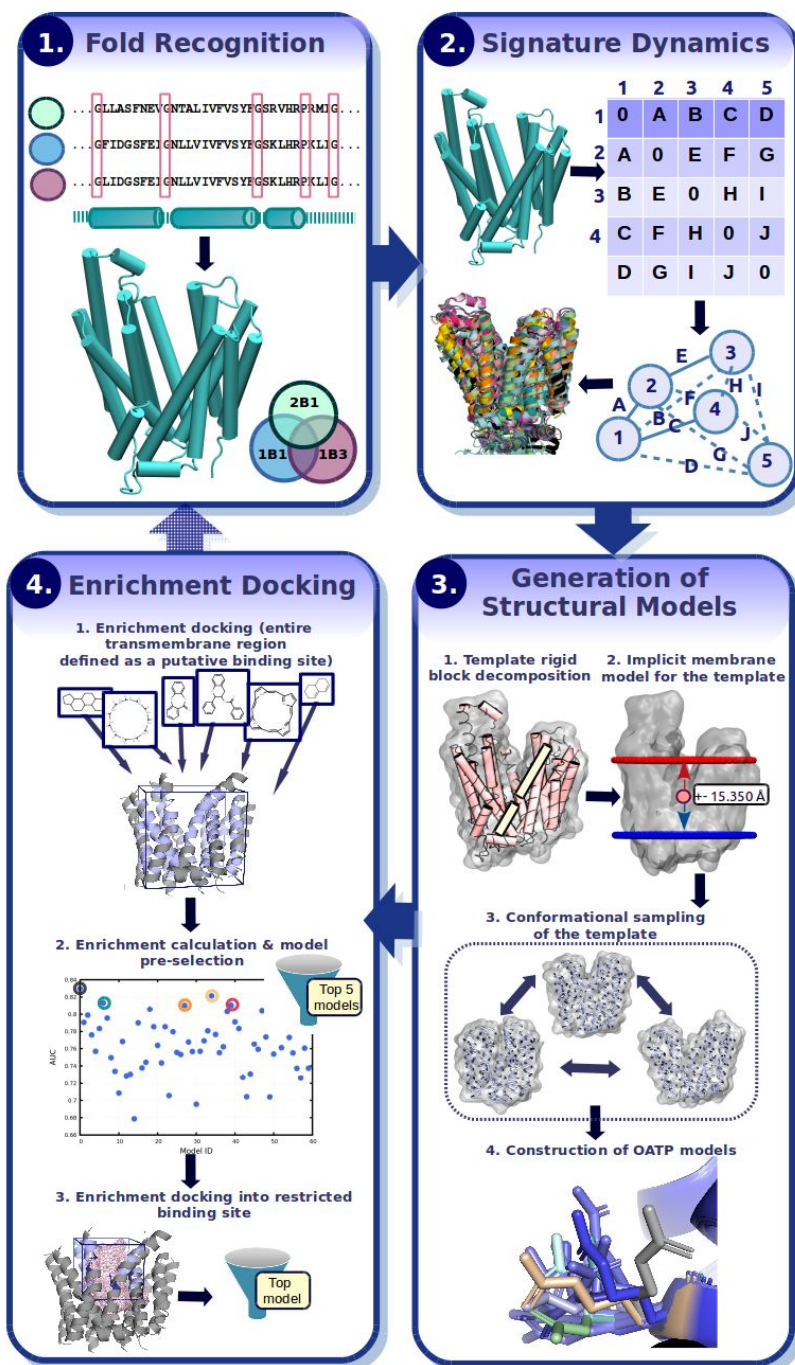


Figure 1

Major steps applied to retrieve structural models. In Step (1) a fold recognition algorithm (here: pGenThreader)

is used to identify the protein fold from the target sequences. Step (2) involves structural alignment of available protein structures possessing the desired fold. NMA is performed to identify conserved motions shared across the proteins with the same fold (here called ‘signature dynamics’). In Step (3) structural models (here: OATP1B1/OATP1B3/OATP2B1) are generated via a multistep procedure; First, the template structure (here: Fucose transporter) is decomposed into individual rigid blocks via the hydrogen bond estimation algorithm (DSSP). Next, a scaling factor is applied to prioritize motions in the radial direction, which mimics the membrane environment (hence ‘implicit membrane model’). Alternate conformers of the template structure are sampled via NMA. Structural models (here: OATP1B1/OATP1B3/OATP2B1) are subsequently built on the basis of different template conformations. For each template conformer, 20 structural models possessing 20 different side-chain rotamers are calculated. Enrichment docking (Step (4)) is performed in two consecutive steps; First, an entire transmembrane region is defined as a putative binding site (as indicated by the cube). Known active ligands and inactives/decoys are docked and the models are subsequently ranked on basis of their AUC values. The top five models are retained. The contact surface area accommodating known actives from the first round of docking is calculated (visualized as pink mesh) and further used to restrict the search space for the second round of enrichment docking calculations. After the enrichment docking is repeated into the top five models, the best model is prioritized on the basis of its AUC value. In case no significant difference between AUC values was observed, the EF 1% was used as an additional metric to prioritize models.

2.2.3. Molecular docking of steroid analogs

The list of sixteen 13-epiestrones with measured activity on OATP1B1, OATP1B3, and OATP2B1 used in this study is given in Table 2. Supplementary File S4 shows steroidal compounds retrieved from public data sources along with their measured bioactivities. Correct stereochemistry of the steroidal nucleus was checked by comparing to experimentally resolved steroids in the Protein Data Bank (PDB) which was fetched via RESTful web services in KNIME (analogous to previously published work by our group).⁽⁵²⁾ Retrieved steroidal structures are visually depicted in Supplementary Table S1.

Autodock Vina 1.1.2. was used to dock steroid analogs. 10 binding modes were sampled. Exhaustiveness of the search was set to 20. To map possible interaction sites for steroid analogs, the entire transmembrane region was defined as a putative binding site.

Docked poses were analyzed via hierarchical pose clustering as follows: (1) A docked ligand structure was reduced to its core scaffold (saved in pdbqt format), (53) (2) a pdbqt file with a core scaffold was converted into a mol file format using OpenBabel 2.4.1., (54) (3) the maximum common substructure (MCS, here: [*]1-,*[*]-,*[*]-,*[*]2-,*[*](-,*[*]-,*[*]1)-[*]1-[*]-,*[*]-[*]3(-[*](-[*]-1-[*]-[*]-2)-[*]-[*]-[*]-3)-[*] in SMARTS, see Figure 5) for all retrieved core scaffolds was calculated (using the FindMCS functionality in RDKit; bond order kept flexible) by using an in-house script (available in the Supplementary File 4), (4) output coordinates were saved in xyz format and converted back to pdbqt format using OpenBabel 2.4.1. (5) MCSs were loaded into PyMOL and (7) the agglomerative hierarchical clustering algorithm within the PyDRA plugin was used to calculate average distances (distance cut-off set to 2 Å). (55)

Cluster analysis was performed in KNIME 4.1.2.(56) Compounds were analyzed by calculating protein-ligand interaction fingerprints (PLIFs) in MOE.(57) H-donor (cut-off 0.5-1.5 [kcal/mol]), H-acceptor (cut-off 0.5-1.5 [kcal/mol]), ionic attraction (cut-off 0.5-3.5 [kcal/mol]), metal ligation (cut-off 0.5-3.5[kcal/mol]), and arene attraction (cut-off 0.5-1.0 [kcal/mol]), were defined as distinct interaction types used in the calculation. The pocket volume was calculated via the open-source POVME binding pocket analysis software. (58) Radius of gyration was calculated by using the `gyradius` functionality within the Psico module (a PyMOL extension).

2.3. Transport inhibition experiments for 13-epiestrones

13-epiestrones were synthesized previously, as described in (59) . 20 mM stocks were prepared from each compound that were stored for further usage at -20°C. A431 cells overexpressing OATPs, OATP1B1, OATP1B3 or OATP2B1, or mock transfected controls were generated previously.(59) A431 cells were maintained in DMEM medium (Thermo Fischer Scientific, Waltham, MA, US) supplemented with 10% fetal bovine serum, 2 mM L-glutamine, 100 units/mL penicillin and 100 µg/mL streptomycin, at 37°C with 5% CO₂. Interaction of 13-epiestrones with OATPs, 1B1, 1B3 and 2B1 was measured in A431 cells

overexpressing the given OATP using the previously identified OATP1B and OATP2B1 substrate pyranine (8-Hydroxypyrene-1,3,6-trisulfonic acid trisodium salt, Sigma, Merck, Hungary).(59,60) Uptake of pyranine was measured on microplates based on the method developed by us previously.(59,60)

Briefly, one day prior to the uptake measurement cells (8×10^4 cells /well in 200 μ l DMEM) were seeded on 96-well plates. On the following day, the medium was removed and the cells were washed three times with 200 μ l phosphate-buffered saline (PBS, pH 7.4) and pre-incubated with 50 μ l uptake buffer (125 mM NaCl, 4.8 mM KCl, 1.2 mM CaCl_2 , 1.2 mM KH_2PO_4 , 12 mM MgSO_4 , 25 mM MES (2-(N-morpholino)ethanesulfonic acid, and 5.6 mM glucose, pH 5.5) with or without increasing concentrations of the tested compounds. The reaction was started by the addition of 50 μ l uptake buffer containing pyranine in a final concentration of 10 μ M (OATP1B1) or 20 μ M (OATP1B3 and OATP2B1). Cells were incubated with the dye at 37°C for 15 min (OATP1B1 and OATP2B1) or 30 min (OATP1B3), after which the supernatant was removed, and the cells were washed three times with 200 μ l ice-cold PBS. Fluorescence (in 200 μ l PBS/well) was determined in an Enspire plate reader (Perkin Elmer, Waltham, MA) ex/em: 403/517 nm. OATP-dependent transport was calculated by extracting fluorescence measured in mock transfected cells. Transport activity was calculated based on the fluorescence signal in the absence (100%) of the tested compounds. Experiments were repeated at least three times on cells deriving from different passages.

2.3.1. IC_{50} value determination

IC_{50} values were calculated by Hill1 fit, using the Origin Pro8.6 software (GraphPad, La Jolla, CA, USA).

3. Results and Discussion

3.1. Insights from conformational sampling of experimentally-resolved MFS structures & ensemble docking into OATP structural models

Biologically relevant motions like protein conformational changes happen at time scales in the range of micro- to milliseconds or even seconds and therefore generally cannot be studied by classical Molecular Dynamics (MD) simulations. By using normal mode analysis (NMA), functional protein motions can be captured by global (“soft”) modes which represent collective motions of entire protein (sub)domains. In this study, the motivation for inclusion of NMA was twofold: (1) To compare normal modes for available structures of the Major Facilitator Superfamily (MFS) members (45 structures) from the Protein Data Bank (PDB) and thus identify functionally important protein motions, and (2) to sample alternate template conformers.

In the former case, our intention was to explore how intrinsic protein dynamics might diverge across a protein family with a shared fold (so called “signature dynamics”). Global fluctuations shared across the MFS proteins might deliver useful insights into the transport mechanism. In the latter case, we incorporated the knowledge about MFS dynamics from step (1) into our ensemble docking strategy in such a way that anisotropic Network Models (ANMs) for the selected template (here: FucP, PDB ID: 3o7q) were calculated.

Signature dynamics of MFS transporters. In order to assess feasibility of the elastic network models for exploring intrinsic dynamics of MFS proteins, square fluctuations derived from NMA can be compared with crystallographic B-factors for experimentally-determined structures. In this study, a crystal structure of FucP transporter (PDB ID: 3o7q) was taken as a reference to compare NMA-based fluctuations of C α carbons with the B-factors from X-ray crystallography (Figure 2). A mean square fluctuation profile of the five softest modes of 45 proteins with MFS fold from PDB (Figure 2A and Supplementary Figure S2) exhibited significant fluctuations in distinct regions (based on residue numbering): 45 - 68, 182 -195, 240 - 260, 280 - 302, and 392 – 420 (see colored regions in Figure 2A).

A high variance in the region 182 - 195 is caused due to the fact that some transporters (such as human GluT3 transporter, PDB IDs: 4zwb, 4zwc, 4zw9, 5c65), (61) are possessing a significantly extended TMH1, reaching to the extracellular region. Fluctuations located in this region show discrepancy between GNM-based values and B-factors for FucP template. The sharp increase in the theoretical fluctuations derived from GNM are caused by a lower number of inter-residue contacts in the elastic network, which leads to the high flexibility of the respective loop region. These findings are consistent with other network models for MFS transporters available in the literature. (62) Large fluctuation values were also observed for the region 240 - 260 which is located in the cytoplasmic part of MFS transporters. Specifically, variance in the cytoplasmic region is likely caused by the presence or absence of specific intracellular domains, such as the YAM domain of E. coli Transporter YajR (PDB ID: 3wdo) (63,64) or the intracellular helical (ICH) domain consisting of three to four helices in sugar transporters. (65–67)

Due to the structural ambiguity in extra- and intra-cellular regions identified herein, we have decided to model OATPs in such a way that the final structural models do cover transmembrane regions only. Specifically, our primary aim was to unravel the binding modes for steroid analogs which are chemically closely related to endogenous substrates of these transporters (such as DHEA). The general hypothesis to date is that substrates are binding to the inner transmembrane region. (68) Figure 3 shows a schematic overview of the transmembrane regions in OATPs and indicates residues identified within the given regions to be important for transport function.

Interestingly, mode 1 and mode 2 show mutual (“out-of-plane”) shifts in the upper part of TMH1 and TMH2 (region 45 - 68). Fluctuation values of this region are in a good agreement with experimental fluctuations for FucP transporter (Figure 2A). As this motion occurs inside of the transmembrane core, we assume that the fluctuations at the TMH1/TMH2 interface could have an impact on the ligand accessibility and binding. The other (albeit less pronounced) motions inside of the transmembrane core are located in the upper part of TMH7 (280 - 302) and TMH11 (392 - 420). Again, the fluctuations of these regions share corresponding mode shapes with B-factors from experiments (Figure 2A). These findings prompted us to generate alternate conformations of the FucP template along mode 1 and mode 2, so the final conformational ensemble reflects the fluctuations TMH1, TMH2, TMH7, and TMH11.

By reordering sequence- and structure-based matrices according to the “dynamics-based” similarities, a cluster of analogous proteins (n=15, Supplementary Table S2, Supplementary Figure S2) was identified. Interestingly, some of those dynamically-related proteins (PDB ID: 4m64, 3wdo, 4gc0, Supplementary Table S2) were predicted by the pGenThreader algorithm as suitable templates to model hepatic OATPs. These findings show that the structural templates individually predicted by fold recognition tools are related not only sequentially and structurally, but also dynamically, which increases the confidence in fold recognition methods for detecting valuable templates.

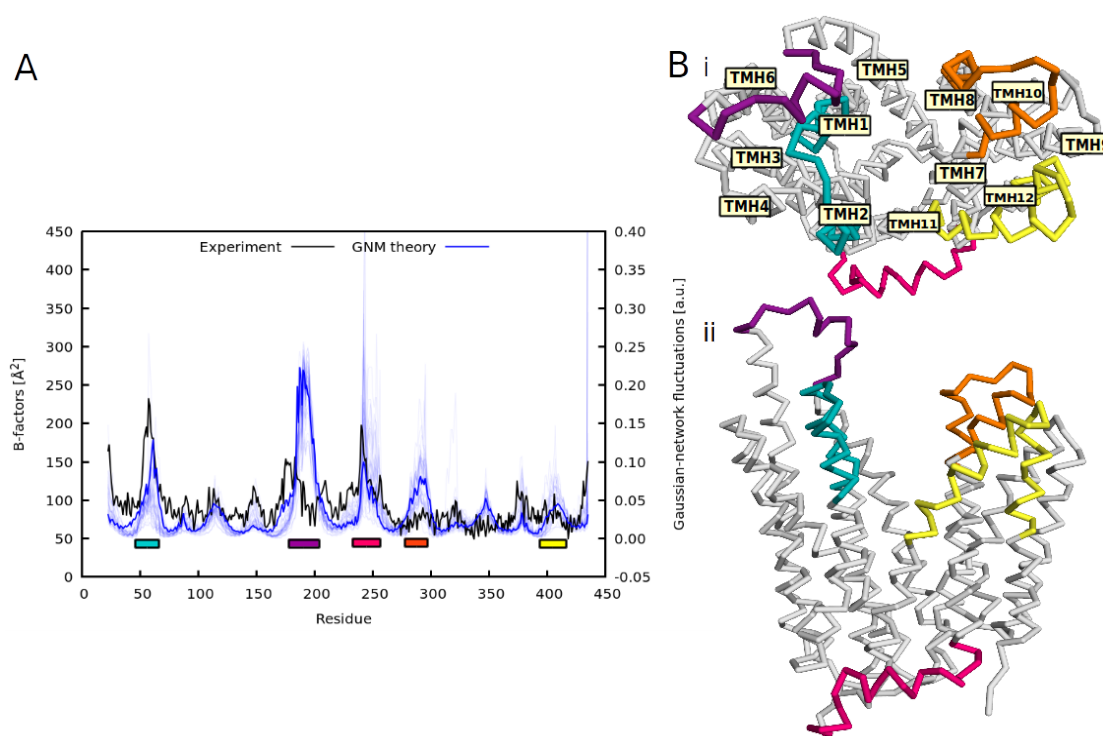


Figure 2

(A) Mean square fluctuations of the selected MFS proteins ($n=45$) derived from GNMs (blue curves) and from X-ray structure of FucP transporter (PDB ID: 3o7q, black curve). The experimental mean square fluctuations are indicated in Å² units, while the theoretical calculations are given in arbitrary units. The regions of fluctuations discussed in the text are marked by a respective color. The coloring corresponds to the concrete regions in MFS structure, as shown in (B) (i) top view with transmembrane helix numbering, (ii) side view.

Ensemble docking into OATP1B1, OATP1B3, and OATP2B1 structural models.

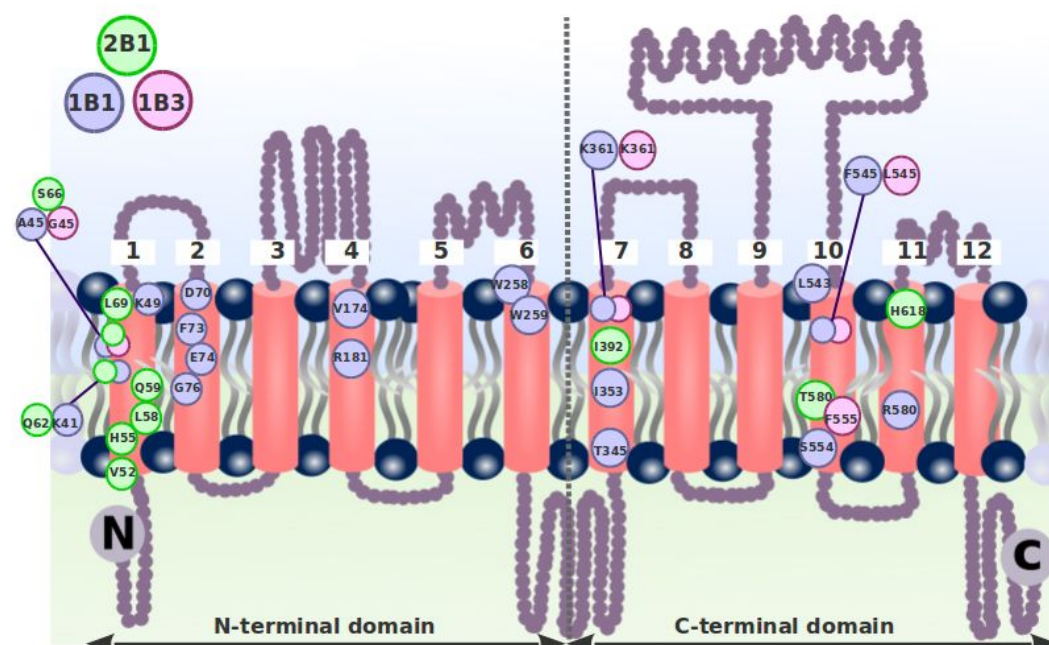
Signature dynamics of the MFS ensemble revealed the most dominant protein motions for TMH1, TMH2, TMH7, and TMH11 (normal mode 1 and 2). The FucP template (PDB ID: 3o7q) was therefore sampled in such a way that the final conformational ensemble includes variance along normal mode 1 and 2. Specific settings for conformational sampling of the template (including implicit membrane model) can be found in the methodological section (subsection 2.1.2.).

Studying the ROC curves at EF1% for the top five models (Figure 4Bi-iii) for OATP1B1 (AUC=0.68-0.83, EF1%=0.0%-5.0%), OATP1B3 (AUC=0.75-0.94, EF1%=18.5%-23.8%), and OATP2B1 (AUC=0.50-0.70, EF1%=0.0%-6.04%), differences in the model performances are becoming obvious. The models' abilities to separate highly actives from inactives/decoys performed best for the OATP1B3 models, despite a smaller data set of actives when compared to OATP1B1 (25 actives vs. 57 actives). This phenomenon can be explained by the fact that in case of OATP1B1 docking 1,223 decoys from DUD-E have been added to the set of measured inactives, whereas for OATP1B3 docking only measured inactives (n=900) were used for ligand enrichment calculations. The ROC curve of the top OATP1B1 models is flatter since obviously the decoys are more often falsely classified as actives (false positives) than the measured inactives.

The comparably least performing models were those for OATP2B1 (Figure 4Biii), which can be explained by the small overall compound set for docking (only 12 highly actives, 153 inactives, 279 decoys) and a (relatively) weaker cut-off for defining activity that was used in this case (5 μ M).

Interestingly, a certain trend between AUC values and the radius of gyration of the OATP structural models was observed (Figure 4A). An example is given in Figure 4A for OATP1B1 models. Specifically, the initial template structure (black structure in Figure 4A) and models with increased accessibility of the translocation pore (orange structure in Figure 4A) generally exhibit higher AUC values (0.76-0.82), compared to structures with smaller radius of gyration and thus a narrower translocation pore (pink structure in Figure 4A, AUC values ranging from 0.67 to 0.78). Similarly, an increased pocket volume of the translocation pore is related to an increase in AUC values. The five prioritized models for OATP1B1, OATP1B3, and OATP2B1, differ from the initial template conformation (average RMSD of 2,8 Å for OATP1B1, 3,1 Å for OATP1B3, and 3,2 Å for OATP2B1, Figure 4Ci and Figure 4Cii). Specifically, correlated movements of TMH1 and TMH2 (out-of-plane motion, Figure 4Cii), as well as the fluctuations in the upper part of TMH7 and TMH11 (opening the central cavity, Figure 4Cii), are possessing the highest deviation from the initial template. These observations indicate that the models might benefit not solely from the increased ligand accessibility (as evidenced by the increase of the radius of gyration and pocket volume compared to the initial template structure), but also from the specific directionality of TMH1

The final structural models used in this study are available from the Supplementary Material (Supplementary Files S1-S3 and are visually depicted in Supplementary Figure S4). An overview of amino acid residues spanning the different transmembrane regions for the three transporters is given in Supplementary S7.



Schematic overview of regions which are known to carry residues important for OATP1B1 (the blue coloring), OATP1B3 (the magenta coloring), and OATP2B1 (the green coloring) for transport function. Colored circles indicate specific amino acid residues identified within the given region to be important for transport function.

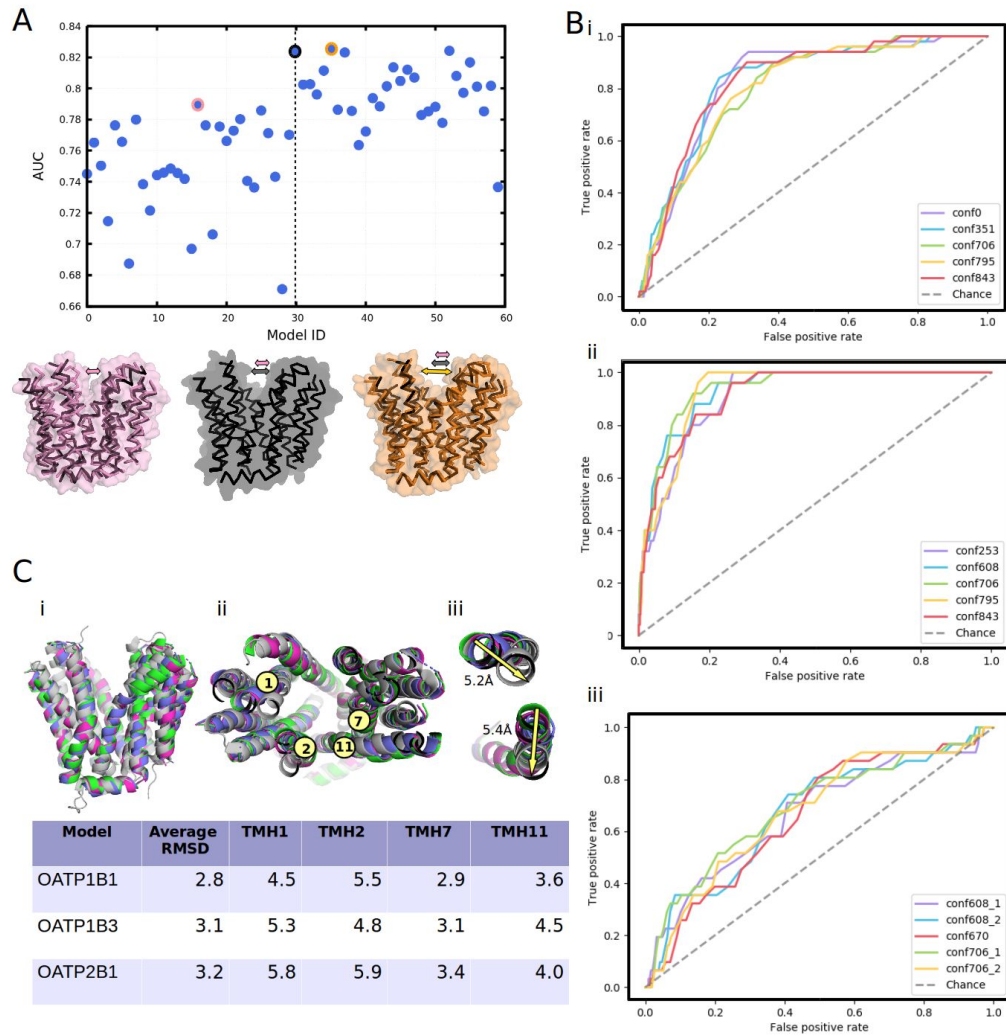


Figure 4

Enrichment docking into OATP1B1/OATP1B3/OATP2B1 structural models. (A) Distribution of AUC values for OATP1B1 models from the first docking round. The models are ordered according to the ascending value of their radius of gyration. The highlighted points in the plot correspond to the models depicted below the plot. The arrows indicate the degree of openness of the individual transporter structures. (B) ROC curves of the first docking run for the top five (i) OATP1B1, (ii) OATP1B3, and (iii) OATP2B1 structural models, respectively. (C) (i) Prioritized models for OATP1B1 (blue), OATP1B3 (magenta), and OATP2B1 (green) compared to the initial template (gray). Figure (ii) highlights TMHs with the highest variability. Figure (iii) depicts an out-of-plane motion of TMH1 and TMH2 compared to the initial template, as indicated by yellow arrows. The table lists distances [in Å] for TMH1 (C(55/76)---C(52) distance), TMH2 (C(66/87)---C(62) distance), TMH7 (C(361/396)---C(283) distance), and TMH11 (C(596/623)---C(407) distance) between the initial template and

the final models.

3.2. *General trends from molecular docking of steroid analogs: Orientational versatility of the steroidal core*

Before individual ligand-transporter interactions were studied in more detail, general trends with respect to the position of side chains on the steroidal core, being responsible for transporter interactions, were investigated. Interestingly, a significant portion of the docked steroids (55% for OATP1B1, 41% for OATP1B3, and 44% for OATP2B1) accomplish binding via their substituents located at position 17. Next, substituents at the position 3 (31% for OATP1B1, 21% for OATP1B3, and 27% for OATP2B1) and 2 (9% for OATP1B1, 18% for OATP1B3, and 28% for OATP2B1) also form distinct interactions. Frequencies of variations at different R-group positions in our data set (Table 2) are depicted in Figure 5. These findings are in accordance with our previously published study on exploring OATP ligand profiles.⁽²⁰⁾ It has to be noticed, however, that these trends might be influenced by the availability of different substituents at the respective steroid core positions (e.g., for the OATP2B1 data set more variability in position 2 is present).

Given the bolaamphiphilic nature of steroidal compounds, orientational versatility which allows a 180 degrees flip of the docked poses appears possible, similarly to previously reported studies on steroid receptors.⁽⁶⁹⁾ Different orientations of the steroidal core (standard/head-to-tail reversed) are enabled, as long as the key interactions are at least partially preserved. Such interactions are mainly mediated by R-3 and R-17 substituents for our data set 95% of OATP1B1, 75% of OATP1B3, and 66% of OATP2B1 actives were able to flip within the binding site to adopt the alternate orientation. To give an example, E-3-S adopts two distinct binding modes in OATP1B1 (Figure 6A). GLU185 appears as a key residue in both modes, as it is capable of forming a hydrogen bond with either the R-3 or R-17 substituent. A similar scenario has been observed for estrone-3-sulfate docking into OATP1B3: ASN213 contributes to hydrogen bond formation with either the R-3 or R-17 substituent (Figure 6B). In OATP2B1, E-3-S binds to the N-terminal cavity, where the

steroidal core can be flipped both up and down. GLN196 stabilizes the compound in either orientation by providing a nitrogen atom which acts as a H-bond donor for either the R-3 or R-17 substituent (Figure 6C). Orientational flexibility was also observed for other steroid analogs, such as chenodeoxycholic acid. In case of OATP1B1, chenodeoxycholic acid binds to its N-terminal site with two possible orientations. The compound interacts with GLU74 in both poses. In addition, head-to-tail reversed orientation reveals interaction of the R-17 substituent with ASP70 and the R-3 substituent with ASN77. Both Glu74 and Asp70 have been identified as important residues involved in E-3-S uptake in alanine scanning experiments on OATP1B1. (27) For some steroids, however, the flip was disabled. These findings were observed e.g., for 2- and 4-phosphonated 13-epiestrones docked into OATP2B1, likely due to their bulky substituents which form distinct interactions at the TMH1/2 interface (see Section 3.4 for further details). Similarly, halogenated epiestrones showed a certain orientational preference due to preferred halogen bond formation in the N-terminal domain of OATP2B1 (see Section 3.4).

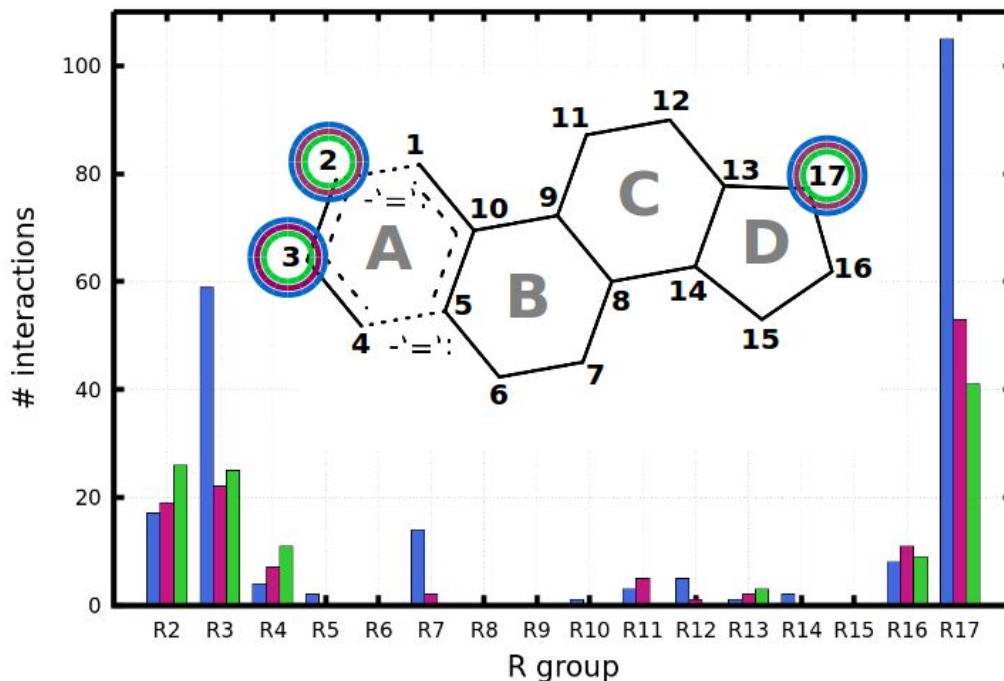


Figure 5

Counts of protein-ligand interactions per R-group position (considering all poses). Color code: OATP1B1 substituents....blue, OATP1B3 substituents....magenta, OATP2B1 substituents....green.

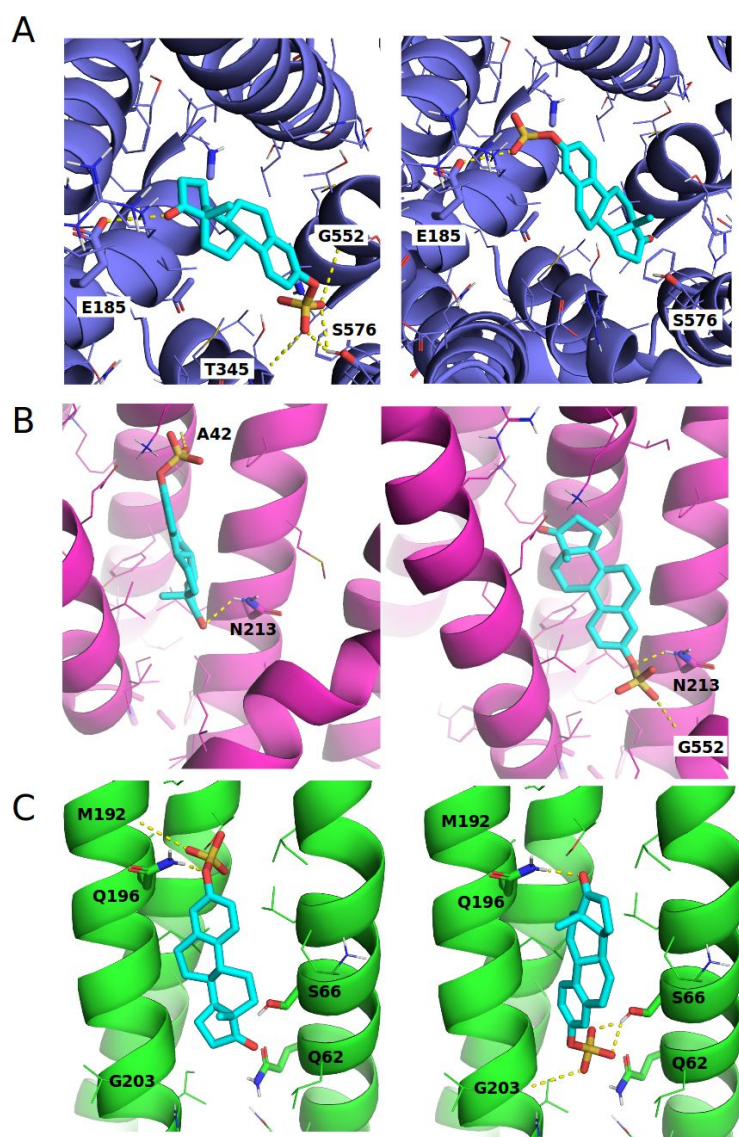


Figure 6

Orientational flexibility of steroidal compounds exemplified by estrone-3-sulfate docked into (A) OATP1B1, (B) OATP1B3, and (C) OATP2B1, respectively. Possible polar contacts are depicted with dashed yellow lines. Color code: docked compounds....cyan (carbon), red (oxygen), yellow (sulfur); OATP1B1....blue; OATP1B3....magenta; OATP2B1....green.

3.3. Shared and distinct interactions of steroid analogs with hepatic OATPs

Analyzing common and distinct binding modes of the three related transporters can be carried out systematically by pose clustering and frequency analysis of transporter-ligand interactions. Cluster analysis yielded 15 distinct clusters for OATP1B1, 9 distinct clusters for OATP1B3, and 9 distinct clusters for OATP2B1, respectively (Supplementary Tables S4-S6). Clusters per transporter were prioritized on the basis of both the number of poses per cluster (compounds may appear more than once in a single cluster) and the number of unique compounds per cluster. Filtering for clusters that possess more than 50% of the actives per respective transporter, three distinct clusters for OATP1B1 (85%, 65%, and 55% of unique compounds), two distinct clusters for OATP1B3 (100% and 83% of unique compounds), and one single cluster for OATP2B1 (94% of unique compounds) were retained for further investigations (Supplementary Figures S5-S7).

Next, protein-ligand interaction fingerprint (PLIF) analysis retrieved shared and distinct interactions across the three transporters (Figure 7, Table 3, and Supplementary Figures S7-S9).

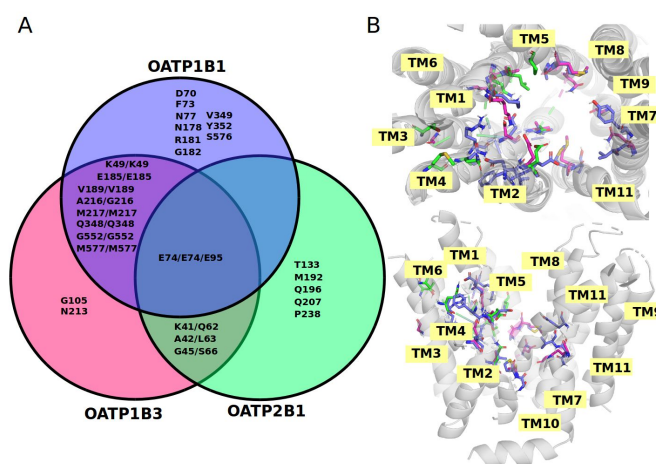


Figure 7

Key amino acid residues in OATP1B1/OATP1B3/OATP2B1 interacting with steroid analogs. (A) Shared and

distinct key protein ligand interactions. (B) Visualization of the key interacting residues within the transmembrane region (OATP1B1 residues shown in blue, OATP1B3 in magenta, OATP2B1 in green).

Table 3

Key amino acid residues involved in ligand binding considering prioritized clusters only (highlighted in bold). ○ = 1st, ● = 2nd, ● = 3rd enriched OATP1B1 cluster, ● = 1st, ● = 2nd, enriched OATP1B3 cluster, ● = top enriched OATP2B1 cluster. Residues that are reported in the literature to be important for transport activity are marked in color (blue for OATP1B1, pink for OATP1B3, and green for OATP2B1, respectively).

TMH	OATP1B1		OATP1B3		OATP2B1
1	PHE38		TYR38	● ●	GLN59
1	LYS41		LYS41	● ●	GLN62 ●
1	THR42		ALA42	● ●	LEU63 ●
1	ALA45		GLY45	● ●	SER66 ●
1	LYS49	● ●	LYS49	● ●	LYS70
2	ASP70	● ●	ASP70		ALA91 ●
2	PHE73	● ●	PHE73		ASN94 ●
2	GLU74	● ●	GLU74	●	GLU95 ●
2	ASN77	● ● ●	ASN77		ASN98
3	GLY105		GLY105	●	ALA126
3	ALA112		SER112		THR133 ●
4	TYR173		TYR173		MET192 ●
4	GLY177		GLY177		GLN196 ●
4	ASN178	● ● ●	ASN178		THR197
4	ARG181	● ● ●	ARG181		LEU199
4	GLY182	● ●	GLY182		VAL201
4	GLY184		GLY184		GLY203 ●

4	GLU185		GLU185		VAL204
4	ILE188		ILE188		GLN207
4	VAL189		VAL189		PRO208
5	ASN213		ASN213		ALA232
5	ALA216		GLY216		MET235
5	MET217		MET217		MET236
5	GLY219		GLY219		PRO238
7	THR345		THR345		GLN380
7	GLN348		GLN348		LEU384
7	VAL349		SER349		SER385
7	TYR352		ILE352		ALA388
10	GLY552		GLY552		HIS579
10	VAL556		ILE556		PHE583
11	SER576		SER576		MET604
11	MET577		MET577		PHE605
11	ARG580		ARG580		ARG607

Table 4

Frequency of interactions formed by steroid analogs with OATP1B1, OATP1B3, and OATP2B1 calculated by PLIFs. The percentage of all possible interactions, as well as the percentage of all poses (indicated in brackets) is listed in the table.

TMH	OATP1B1		OATP1B3		OATP2B1	
1	PHE38	0% (0% poses)	TYR38	8%(6% poses)	GLN59	0%(0% poses)
1	LYS41	0% (0% poses)	LYS41	16%(13% poses)	GLN62	34%(30% poses)
1	THR42	0% (0% poses)	ALA42	23%(16% poses)	LEU63	10%(10% poses)
1	ALA45	0% (0% poses)	GLY45	14.5%(15% poses)	SER66	32%(30% poses)

1	LYS49	5%(5% poses)	LYS49	16%(16% poses)	LYS70	0%(0% poses)
2	ASP70	3%(3% poses)	ASP70	0% (0% poses)	ALA91	32%(24% poses)
2	PHE73	10%(8% poses)	PHE73	0% (0% poses)	ASN94	2%(2% poses)
2	GLU74	5%(5% poses)	GLU74	2%(2% poses)	GLU95	27%(27% poses)
2	ASN77	25%(14% poses)	ASN77	0%(0% poses)	ASN98	0%(0% poses)
3	GLY105	0% (0% poses)	GLY105	2%(2% poses)	ALA126	0%(0% poses)
3	ALA112	0% (0% poses)	SER112	0%(0% poses)	THR133	12%(12% poses)
4	TYR173	0% (0% poses)	TYR173	0%(0% poses)	MET192	5%(5% poses)
4	GLY177	0% (0% poses)	GLY177	0%(0% poses)	GLN196	41%(29% poses)
4	ASN178	10%(10% poses)	ASN178	0%(0% poses)	THR197	0%(0% poses)
4	ARG181	22%(12% poses)	ARG181	0%(0% poses)	LEU199	0%(0% poses)
4	GLY182	7%(7% poses)	GLY182	0%(0% poses)	VAL201	0%(0% poses)
4	GLY184	0% (0% poses)	GLY184	0%(0% poses)	GLY203	20%(12% poses)
4	GLU185	17%(10% poses)	GLU185	35%(33% poses)	VAL204	0%(0% poses)
4	ILE188	0% (0% poses)	ILE188	0%(0% poses)	GLN207	7%(5% poses)
4	VAL189	8%(8% poses)	VAL189	13%(10% poses)	PRO208	0%(0% poses)
5	ASN213	0% (0% poses)	ASN213	15%(8% poses)	ALA232	0%(0% poses)
5	ALA216	15%(3% poses)	GLY216	6%(6% poses)	MET235	0%(0% poses)
5	MET217	8%(8% poses)	MET217	4%(4% poses)	MET236	0%(0% poses)
5	GLY219	0% (0% poses)	GLY219	0%(0% poses)	PRO238	2%(2% poses)
7	THR345	0% (0% poses)	THR345	0%(0% poses)	GLN380	0%(0% poses)
7	GLN348	5%(5% poses)	GLN348	6%(4% poses)	LEU384	0%(0% poses)
7	VAL349	3%(3% poses)	SER349	0%(0% poses)	SER385	0%(0% poses)
7	TYR352	7%(7% poses)	ILE352	0%(0% poses)	ALA388	0%(0% poses)
10	GLY552	19%(17% poses)	GLY552	15%(15% poses)	HIS579	0%(0% poses)
10	VAL556	0% (0% poses)	ILE556	0%(0% poses)	PHE583	0%(0% poses)
11	SER576	8%(8% poses)	SER576	0%(0% poses)	MET604	0%(0% poses)
11	MET577	5%(8% poses)	MET577	21%(19% poses)	PHE605	0%(0% poses)
11	ARG580	0% (0% poses)	ARG580	31%(13% poses)	ARG607	0%(0% poses)

Studying the top three ligand clusters in OATP1B1, it becomes obvious that many amino acids are shared (or overlapping) between the clusters (e.g., 8 out of 19 residues are shared among all three); see Table 3). The two top-ranked clusters (accommodating 85% and 65% of the docked compounds, respectively) are reaching into the central cavity of the transporter, being enframed by TMH5, TMH7-8, TMH10-11. The third cluster (55% of docked compounds) is located closer to the N-terminal domain of the transporter and is lined by TMH1-5. With respect to frequency of ligand interactions, ASN77 at TMH2 (25% interactions), GLU185 at TMH4 (17% interactions), ALA216 at TMH5 (15.2% interactions), and GLY552 at TMH10 (19% interactions), are the main contributing residues.

In OATP1B3, the top-ranked cluster interacts with residues from TMH1-5, (and partly TMH7 and TMH11), similar to the third cluster in OATP1B1. The second prioritized ligand cluster for OATP1B3, however, is located closer to the central cavity, being lined by TMH1, TMH2, TMH7, and TMH10-11. Here, even more residue interactions are shared among the two ligand clusters (14 out of 17), making a strict separation of the observed binding modes even more difficult for this transporter. In general, the most prominent ligand interactions are occurring with LYS49 (16% interactions) and GLY45 (14.5% interactions) at TMH1, GLU185 at TMH4 (35% interactions), and MET577 at TMH7 (21% interactions).

In contrast to OATP1B1 and OATP1B3, OATP2B1 ligand cluster analysis did only prioritize a single binding site near the N-terminus with 94% of active ligands docked into this site (Supplementary Figure S7). Polar interactions with R-3/R-17 substituents are mainly accomplished via GLN62 at TMH1 (34% interactions), SER66 at TMH1 (32% interactions), GLU95 at TMH2 (27% interactions) and/or GLN196 at TMH4 (27% interactions).

Comparing the highlighted binding regions across the three transporters, there is a tendency that steroidal compounds dock with a larger prevalence to the inner binding cavity in OATP1B1 (mainly TMH 5, 7, 8, 10, 11 involved), whereas for docking into OATP1B3 both, poses within the inner cavity, as well as poses closer to the N-terminal region were prioritized (mainly TMH 1-5 involved). OATP2B1, which is least similar by sequence to the other two transporters, did not accommodate a relevant amount of poses in the inner cavity

but clearly showed favorable steroid interactions with residues at the N-terminal region.

The role of GLU74/95 in hepatic OATPs

Across all three transporters, there is only a unique shared amino acid residue that was prioritized during steroid docking: GLU74 (in OATP1B1 and OATP1B3)/GLU95 (in OATP2B1) at TMH2 (Table 3, Figure 7A). By comparing the frequency of the interactions for OATP1B1 (5% interactions), OATP1B3 (2% interactions), and OATP2B1 (27% interactions), we reveal a preferred involvement of GLU95 in OATP2B1-ligand interactions compared to the other two transporters. In OATP2B1, several phosphonated 13-epiestrones are capable of forming polar contacts with GLU95. In OATP1B1, on the contrary, GLU74 interacts with mometasone furoate (median bioactivity value [μ M]=0.070). Specifically, the hydroxyl group at the R-17 position forms a H-bond with the carboxyl group of GLU74. In OATP1B3, GLU74 forms a H-bond with, beclomethasone (median bioactivity value [μ M]=6.174) in such that the hydroxyl group at the R-11 position interacts with the deprotonated oxygen in the side chain of GLU74.

In contrast with different frequency of ligand interactions with GLU74/95, a formation of intramolecular salt bridges by GLU74/95 with the neighboring residues was consistently found in all the three transporters (Supplementary Figure S12). In OATP1B1 and OATP2B1, GLU74 forms a salt bridge with ARG580/607 at TMH11 (OATP1B1/OATP2B1). In OATP1B3, however, GLU74 forms a salt bridge with LYS49 (TMH1). These findings suggest multiple scenarios of how the salt bridges can appear in hepatic OATPs (see Supplementary Figure S12).

The importance of GLU74 and other residues at TMH2 (ASP70, PHE73, GLY76) for E-3-S transport by OATP1B1 has previously been confirmed by mutagenesis studies. Mutation of these residues to alanine led to a significant loss of E-3-S uptake activity. (27) Li et al. have postulated the role of GLU74 for transport function to be mainly acting as a stabilizing factor for the binding site by formation of a salt bridge with a nearby positively charged amino acid. (27) As a result of our docking study, we observed GLU74/95 to be involved in direct

formation of H-bonds as well as stabilizing the protein through the intramolecular salt bridges.

Shared and distinct steroid interactions with OATP1B1 and OATP1B3

As seen from Table 3 and Figure 7, many of the direct protein-ligand interactions for the prioritized docking poses (8 out of 14) are shared among OATP1B1 and OATP1B3. These are shared interactions with multiple different TMHs (TMH1, TMH2, TMH4, TMH5, TMH7, TMH10, and TMH11, respectively) and are belonging to different clusters of poses (as discussed in the previous section). The interaction region is extended from the N-terminal domain to the inner cavity of OATP1B1 and OATP1B3 transporter, as already indicated by the cluster analysis.

In the central binding site, GLN348 (5% interactions for OATP1B1 and 6% interactions for OATP1B3), as well as GLY552 (19% interactions for OATP1B1 and 15% interactions for OATP1B3), are shared residues belonging to this region. An example is illustrated in the case of steroid derivatives possessing valeric acid R-17 substituent, which exhibit similar sorts of interactions in OATP1B1 and OATP1B3. These are deoxycholic acid (median bioactivity value $[\mu\text{M}] = 3.162$ for OATP1B1), lithocholate (median bioactivity value $[\mu\text{M}] = 1.000$ for OATP1B1 and 6.309 for OATP1B3), and chenodeoxycholic acid (median bioactivity value $[\mu\text{M}] = 3.162$ for OATP1B1). Hydroxyl group at the R-3 position was found to form a H-bond with GLN348, while the R-17 substituent formed an H-bond with ASN178. As observed for other classes of compounds, a reversed orientation of the steroidal nucleus could have been observed. In such mode, the R-17 substituents have exerted both electrostatic attraction and H-bond formation with ARG580 at TMH11. Interestingly, bile acids possessing a long flexible R-17 substituents were observed to interact with both hotspots in N- and C-terminal sites in OATP1B1. This behavior was observed for the R-17 valeric acid

derivatives (deoxycholic acid, lithocholate as shown in Figure 8Aii for OATP1B1, chenodeoxycholic acid), as well as for taurochenodeoxycholate (median bioactivity value $[\mu\text{M}] = 7.943$ for OATP1B1, Figure 8Aiii for OATP1B1)) and glycochenodeoxycholate (median bioactivity value $[\mu\text{M}] = 5.011$ for OATP1B1). Polar interactions for R-17 substituents predominantly come from ASN77 at TMH2, whereas the R-3 substituents are accommodated via the interactions with GLN348 at TMH7.

In the N-terminal domain of OATP1B1, polar interactions are mostly mediated by ASN77 at TMH2 (25% interactions), and/or GLU185 at TMH4 (33% interactions), ARG181 (22% interactions), ASN178 (10% interactions) at TMH4. As already mentioned, several residues at TMH2 (ASP70 adopting 3% of interactions, PHE73 adopting 10% of interactions, GLU74 adopting 5% of interactions, and GLY76 showing no polar interactions) were identified as important determinants for both high and low affinity binding of estrone-3-sulphate in the mutagenesis experiments.(27) Moreover, PHE73 is a site of the known single nucleotide polymorphism for OATP1B1.(70) In another experimental study, ARG181 at TMH4 of OATP1B1 was shown to mediate uptake of estradiol-17 β -glucuronide.(71) Interestingly, a six-membered ring at the R-17 position in estradiol-17 β -glucuronide (median bioactivity value $[\mu\text{M}] = 7.943$) has shown H-bond formation with backbone of ASP70, and sidechain interactions with ASN77 and ASN178 (Supplementary Figure S13Ai). Alternate pose identified an electrostatic attraction between the carboxylic group on the ring moiety of estradiol-17 β -glucuronide and ARG181. Next, ARG181 has appeared to adopt multiple interactions with the R-3 glycone moiety of digoxin (median bioactivity value $[\mu\text{M}] = 7.943$). In addition, lactone moiety at the R-17 position, as well as six-membered ring at the R-3 position in ouabain (median bioactivity value $[\mu\text{M}] = 0.597$) forms polar interactions ASN178. These observations point us to the preferred interaction of hydrophilic ring moieties with the residues from the N-terminal region (Supplementary Figure S13). In addition, the N-terminal region is highly positively charged and thus can account for the electrostatic attraction of the charged or hydrophilic substituents. ARG181 involvement was also found for the R-17 substituent of cholic acid methyl ester (median bioactivity value $[\mu\text{M}] = 0.200$).

In contrast to the OATP1B1, OATP1B3 ligands tend to be bound more deeply in the N-terminal region (up to TMH3). Interestingly, OATP1B3 prioritized model is captured in a state with a different network of salt bridges compared to OATP1B1 model (see Supplementary Figure S12). In the case of OATP1B1, there is a salt bridge formed in between LYS41 and GLU185. The equivalent residues in OATP1B3, however, do not form a salt bridge and therefore might contribute to the ligand binding to a larger extent. Also, the pocket accessibility is changed due to the different networks of the salt bridges. A prominent contribution of LYS41 was observed (16% interactions), e.g., for cholic acid methyl ester (median bioactivity value [μ M]=0.125), where the R-17 substituent is extended to the positively charged cavity in such that it interacts with LYS41, whereas the R-3 substituent interacts with ARG580 (31% interactions). Similarly, taurocholic acid (median activity value [μ M]=6.309) adopts a corresponding binding mode, with the R-17 substituent buried into the N-terminal half. Interestingly, the role of LYS41 in OATP1B3 transport has already been validated via experimental approaches.(29)

Another reason for different binding mechanisms between OATP1B1 and OATP1B3 could be the replacement of ALA45 in OATP1B1 to GLY45 in OATP1B3 at TMH1. A loss of methyl group increases the pocket volume (551 \AA^3 for OATP1B3 compared to 510 \AA^3 for OATP1B1). An integrated data mining approach applied in our previous study(20) delivered a full activity profile for digoxin. Comparing the median bioactivity value [μ M]s for OATP1B1 (7.943 μ M), OATP1B3 (0.794 μ M), and OATP2B1 (316.227 μ M), a selective binding to OATP1B3 was identified. The top ranked pose for OATP1B3 shows that the glycone moiety at the position R-3 is buried deeply in the N-terminal domain (Figure 10Aii). Such a pose is enabled due to the GLY45 which lacks the side chain; Therefore, the glycone moiety seems to be more flexible to adopt favorable interactions within a pocket (e.g., with LYS41, as shown in Figure 10Aii). In Figure 10Ai, we show a putative equivalent mode in OATP1B1, where GLY45 is replaced by ALA45. We visualize the van der Waals radius of methyl substituent as a transparent sphere. It becomes evident that ALA45 would lead to the steric clash with the glycone substituent. This amino acid replacement has already been experimentally confirmed as an important determinant for OATP1B1/OATP1B3 regiospecificity. (72) Although we have observed similar modes in OATP1B1 docking poses, we might conclude that the adoption of such a mode would be less entropically favored, since

the pocket volume is a bit more restricted compared to OATP1B3. To fully support this hypothesis, however, free energy calculations would need to be performed.

We might therefore conclude that the distinction in OATP1B1/1B3 steroids binding happens in the N-terminal domain and is predominantly driven by the ligand accessibility and volume of the interaction site.

Interactions of steroid analogs with OATP2B1

Among the three hepatic OATPs studied herein, OATP2B1 shares the smallest amount of amino acid residues that seem to be involved in binding of steroidal ligands with the other two OATPs (only half of the interactions are shared). This behavior was to be expected due to the remarkable difference in amino acid sequence identities between the OATP1B subfamily and OATP2B1 (identity only around 31%). Interestingly, among the shared interactions, those with amino acids in TMH1 (mainly Gln62, Leu63, and Ser66) are shared exclusively with OATP1B3, whereas some of the main interactions with residues at TMH2 (ALA91 and ASN94) are shared with OATP1B1 only. Glu95 (Glu74 in OATP1B1 and OATP1B3) is the only shared interacting residue among all three transporters (as discussed in the previous chapter).

The data set of 13-epiestrones (16 compounds, see Table 2), was intentionally not included into the data set for the enrichment docking procedure. For these novel compounds with measured bioactivities for all three transporters we therefore can guarantee an extremely unbiased selection of the final protein model for docking of those compounds. Since data for OATP2B1 is to date very sparse in the open domain (the only public steroidal compound in our data set with an active measurement (bioactivity values for E-3-S span a broad range from 93 μ M to 5 nM in case of OATP2B1), these new bioactivity measurements are a valuable source of information in order to get closer towards understanding OATP2B1-ligand interactions and potential drivers for selectivity. The latter is in particular supported by this new data set, since 5 of the 16 compounds are showing selectivity towards OATP2B1

(between at least 6-fold difference and 55-fold difference in activities towards OATP2B1 vs. the other two transporters), and 3 additional compounds are showing preferential activity towards OATP2B1. Six additional 13-epiestrones are pan-inhibitors, two are showing complete inactivity towards OATP2B1.

Phosphonated 13-epiestrones were previously reported as strong OATP2B1 inhibitors. (73) The most potent pan-inhibitor of the six phosphonated 13-epiestrones (and the whole group of 13-epiestrones discussed herein) is carrying a diethyl phosphono group in position R-2 and possesses a benzoyloxy moiety at position R-3 (compound 11; $IC_{50} = 0.18 \mu M$). It is at the same time the most potent OATP2B1 inhibitor reported in this study ($IC_{50} = 0.18 \mu M$). We observed two possible binding modes for this compound (Figures 8Ci and 8Cii). In the first one (Figure 8Ci), the diethyl phospho group is oriented towards the inner site of the N-terminal region and forms a H-bond with SER66 (TMH1) and GLY203 (TMH4). GLN196 (TMH4) interacts with the R-17 carbonyl at the other end of the steroid structure. At the same time the benzoyloxy moiety is buried into the lower transporter region in close proximity to PHE231 at TMH5. (Figure 8Ci) Both, SER66 and GLN96 are among the most frequent interactions for OATP2B1-steroid binding. Whereas SER66 was reported in literature to be important for transport of endogenous substrates (REF), GLN96 was already discussed in a previous chapter to be involved in E-3-S binding and orientational versatility of the poses. In the second possible orientation (Figure 8Cii), the diethyl phosphono group is pointing towards the translocation pore (H-bond formation with GLN207), while the R-3 benzyl-ether moiety is localized deeply in the hydrophobic region, surrounded by LEU63 at TMH2, ILE206 at TMH4, and LEU230 at TMH5. (Figure 8Cii) A similar mode of action has been found for R-4 phosphonated counterpart ($IC_{50} [\mu M]=0.794$). The R-4 phosphonated 13-epiestrones with the R-3 methoxy group ($IC_{50} [\mu M]=2.511$) did show an additional interaction between phosphonated moiety and GLN62 at TMH1 (Figure 8Ciii). On the other hand, methoxy groups at the R-3 position do not show any significant interaction. These observations help explain almost 10-fold drop in the bioactivity compared to the R-3 benzyl-ether derivatives. Comparable behavior was observed for the R-3 hydroxy derivatives possessing the phosphonated moiety at either R-2 ($IC_{50} [\mu M]=3.162$) or R-4 position ($IC_{50} [\mu M]=10.000$). Interestingly, estrone-3-sulphate ($IC_{50} [\mu M]=10.000$) adopts an analogous 'V-shaped' binding mode (due to the sp^3 -hybridization of the sulphur atom in sulfate group)

by forming polar interactions between sulfate at the R-3 position and SER66/GLN62 residues.

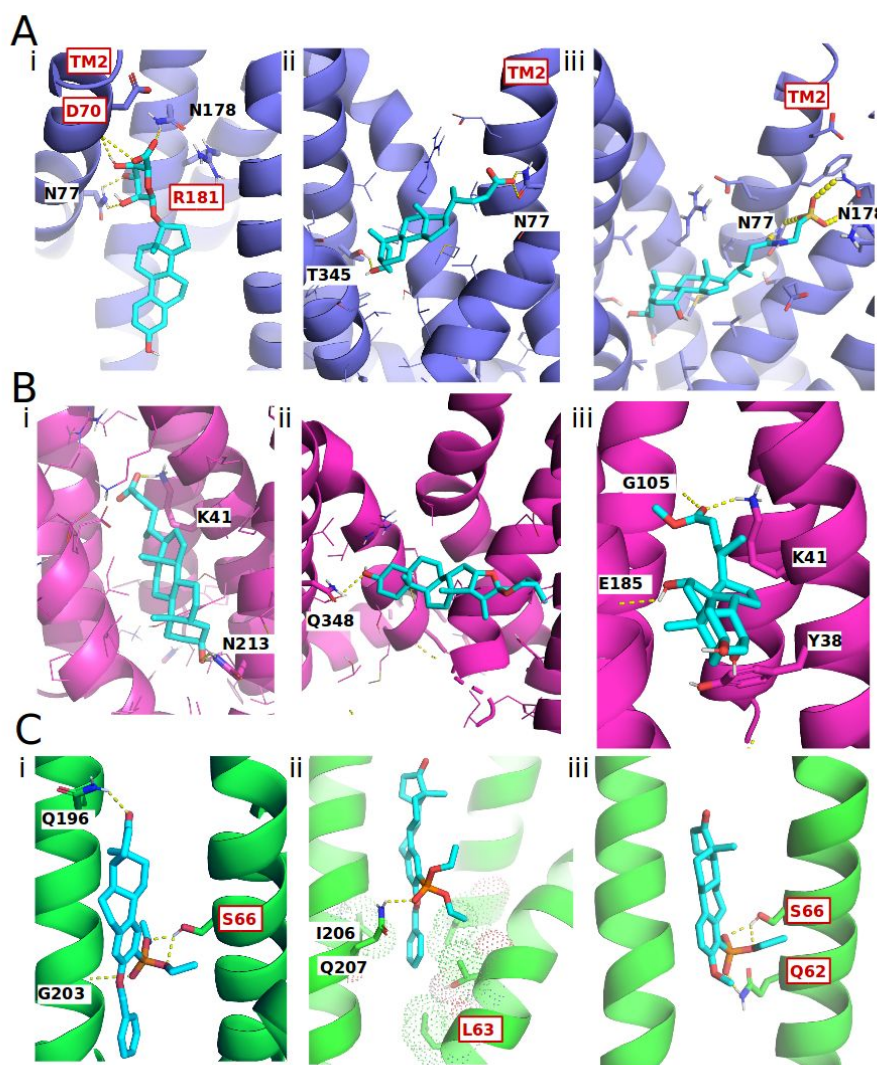


Figure 8

Examples of distinct steroid interactions in (A) OATP1B1, (B) OATP1B3, and (C) OATP2B1 transporter models. Regions/residues which are confirmed by mutational experiments, are highlighted in red. Color coding: docked compounds = cyan (carbon), red (oxygen), yellow (sulfur), OATP1B1 = blue, OATP1B3=magenta, OATP2B1 = green.

As shown in our previous study, halogenated 13-epiestrones act as potent inhibitors of OATP2B1.(49) IC_{50} values for the seven halogenated compounds included in this study are

ranging from 0.6 to 10.45 μM (Table 2). Strikingly, for 13-epiestrones the position of the halogen substituent seems to play a crucial role since halogen substituents at the R-2 position are more active compared to their R-4 counterparts. Comparing the structural analogs, respectively, revealed a 2.5-fold to almost 12-fold difference of the respective bioactivities.(49) Therefore, our intention was to decipher the role of halogen substituents in OATP2B1 ligand recognition. Interestingly, two distinct interaction sites for 13-epiestrones possessing a halogen at the R-2 (here: 'site 1') and the R-4 (here: 'site 2') position were identified. 'Site 1' is located in the upper half of the N-terminal domain, lined by TMH1, TMH2, TMH3, and TMH4.

In 'site 1', halogenated steroid analogs are located in such a way that the R-3 substituent (hydroxyl or methoxy group) is pointing towards the extracellular part of the transporter (Figure 9) and THR133 (TMH3) appears in close vicinity to act as an interaction partner for halogen bond formation. In general, threonine is a known residue capable of forming halogen bonds via its side chain hydroxyl group.(74) In this study, the likelihood for the halogen bond formation has been evaluated on the basis of X---O(THR133) distances and C20-X---O bond angle, where X corresponds to Cl (Figure 9A), Br (Figure 9B), or I (Figure 9C), respectively. Geometric parameters are listed in Table 5. Bond distances are falling into the range of 3.2 to 3.9 Å, while bond angles are ranging from 131° to 156°. Comparable geometric parameters were found for available PDB complexes adopting halogen bonds, with typical interaction distance ranges from 2.5 Å to 6.0 Å, while the interaction angle ranges from 120° to 180°. (75,76) Therefore, we suggest halogen bond formation as the major driver for OATP2B1-ligand interactions at 'site 1'. Moreover, comparing measured $\text{IC}_{50}[\mu\text{M}]$ values 2-iodo-13-epiestrone are more active than 2-bromo- and 2-chloro-13-epiestrones. It seems that the strength of the halogen bond increases with the increasing atomic radius, thus further supporting the existence of the halogen bond. (77) In addition, GLN196 has been found to form a H-bond with the hydroxyl group at the R-3 position. The methoxy substituent (R-3) in compound 4 disables the formation of the H-bond with GLN196 leading to a 7-fold difference in bioactivities observed for compound 4 compared to its hydroxyl derivative (compound 5)..

In 'site 2' (observed for steroids halogenated at position R-4), the steroidal core is flipped in such a way that the R-3 substituent is pointing towards the intracellular part (Supplementary

Figure S11). Here, GLN196 and GLN207 are acting as H-bond donor/acceptors, interacting with the hydroxyl group at R-3 and/or carboxyl oxygen at R-17. SER66 (TMH1) was found as a potential interaction partner for the R-4 halogenated substituents. However, the bond distances (ranging from 3.5 to 4.7 Å), as well as bond angles (ranging from 80° to 110°) do not appear geometrically favorable for the halogen bond formation.

These observations point us to the potential importance of GLN62 and SER66 in OATP2B1 transport function. These findings are in accordance with the alanine scanning experiments done on TMH1 of OATP2B1. (78) Specifically, GLN62ALA and SER66ALA mutations decreased binding affinity of estrone-3-sulfate and taurocholate, thus showing an involvement of these two residues in the OATP2B1 transport function. Moreover, GLN196 and THR133 have been identified as potentially important residues for direct interactions and halogen bond formation with OATP2B1, which could not be experimentally confirmed yet. Since these residues are non-conserved (compared to the other two hepatic OATPs) they could be interesting amino acids for further experimental studies.

Table 5

Geometric parameters for the halogen-bond formation. Values were measured in PyMOL (distance and get_angle in-built functionalities).

Compound ID	Interaction partner	Distance [Å]	Angle [atom selection]	Angle [degree]	Bioactivity value [μM]
Compound 9	THR133	3.2	C20-Cl1---OG1(T133)	131	0.900
Compound 7	THR133	3.9	C20-I1---OG1(T133)	156	0.600
Compound 5	THR133	3.8	C20-Br1---OG1(T133)	155	1.190
Compound 4	THR133	3.7	C2-Br3---OG1(T133)	154	8.394
Compound 10	SER66	3.3	C19-Cl1---OG(S66)	110	10.500
Compound 6	SER66	4.6	C19-Br1---OG(S66)	80	2.971

Compound 8	SER66	4.6	C19-I1---OG(S66)	80	3,580
------------	-------	-----	------------------	----	-------

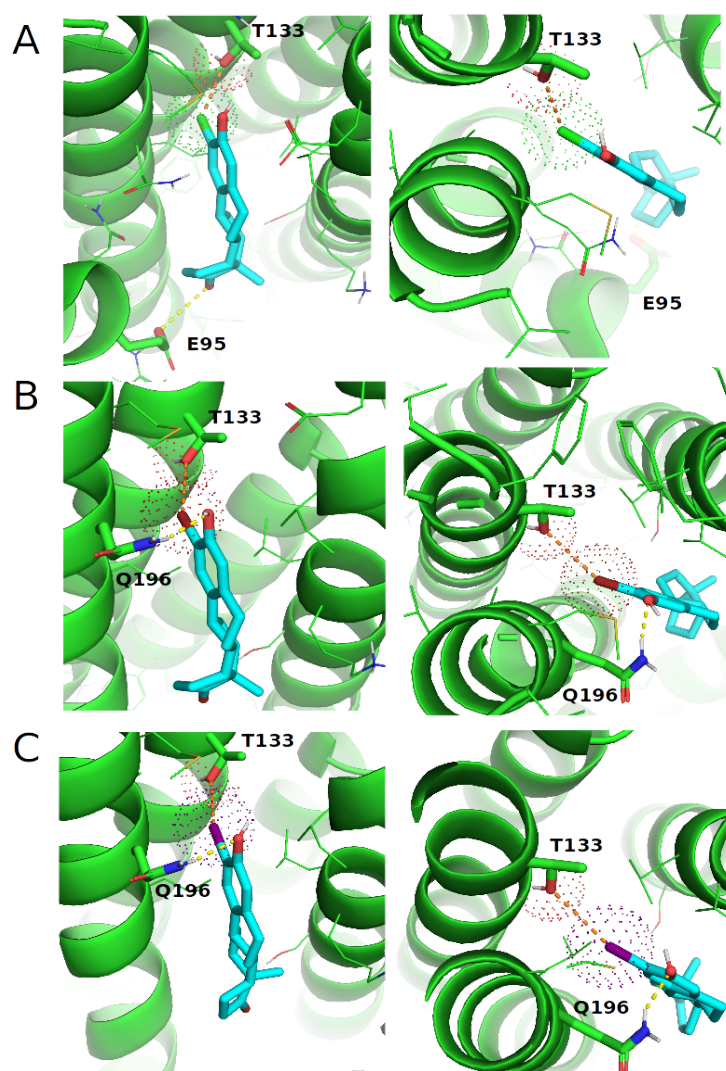


Figure 9

Halogen substituents at R-2 position interact with 'site 1'. Van der Waals radius of a halogen and carboxyl oxygen from THR133 is visualized as a dotted sphere. A halogen bond is visualized as a dashed orange line. H-bonds are visualized as a dashed yellow line. (A) Chlorine (compound 9) (B) bromine (compound 5) (C) iodine (compound 7). Color code: docked compounds = cyan (carbon), red (oxygen), green (chlorine), brown

(bromine), purple (iodine) OATP1B1 = blue, OATP1B3=magenta, OATP2B1 = green. Figures on the left hand side depict side view, while the figures on the right hand side depict top view on the OATP2B1-ligand complexes. Both views are visualized from the extracellular side.

3.4. *Non-conserved residues in the N-terminal binding site help explain OATP1B/OATP2B1 selectivity*

Full bioactivity profiles measured for 13-epiestrones allow the suggestion of structural determinants for OATP1B/OATP2B1 selectivity of steroid analogs. In general, phosphonated 13-epiestrones in OATP1B1/OATP1B3 tend to adopt the same vertical binding mode at the TMH1/TMH2 interface, as seen for OATP2B1 complexes. However, due to the presence of bulky and/or charged residues (such as ARG181 which is OATP1B-specific), the compounds are shifted more towards the region of the translocation pore and thus cannot fully adopt an equivalent mode as seen in the OATP2B1. In OATP2B1, SER66 at the TMH1 has been identified as a key residue accomplishing the polar interactions with sp³-hybridized phosphonated group at the position R-2 (Figure 10Biii). The corresponding residues in OATP1B1 (ALA45, Figure 10Bi) and OATP1B3 (GLY45, Figure 10Bii) are disabled to form the desired interactions. In OATP1B1, phosphonated 13-epiestrones interact with ASN178 (with the R-17 substituent) and VAL189 (providing hydrophobic contact for the R-3 substituent). However, the phosphonated group at position R-2 does not form any distinct interaction. Similarly, OATP1B3 provides a stabilizing hydrophobic contact to the R-3 substituents with aromatic moiety. GLU185 forms a polar contact with the R-17 substituents. The phosphonated group at the R-2 position, however, lacks any interaction partners within the OATP1B3 cavity.

In contrast to phosphonated 13-epiestrones which mostly do show decent activity on OATP1B1 and OATP1B3 as well, the halogenated derivatives are showing a tendency to be more active on OATP2B1 and sometimes show selectivity for OATP2B1 (compounds 4, 5, 7, 9). Interestingly, these selective compounds are all halogenated in position R-2, whereas the R-4 halogenated compounds tend to be still active on OATP1B1 and OATP1B3 with a 2-

to 3-fold weaker affinity (compound 6 and 8). One explanation could be the lower likelihood of R-4 halogenated substituents to adopt bond formation, as judged by the geometric parameters for typical halogen bonds (Table 5). In OATP1B1 and OATP1B3 modes, a certain preference of R-4 halogenated substituents to bind into the inner cavity was observed. In none of the calculated poses, however, the halogenated substituent at R-4 position appeared to be implicated in halogen bond formation. These findings could rationalize comparable bioactivity values for compound 6 and 8 against OATP1B1, OATP1B3, and OATP2B1.

In contrast to the binding modes observed for OATP2B1, calculated poses in OATP1B1 and OATP1B3 show a preferential binding of those compounds in the central cavity rather than in the N-terminal domain. In OATP1B1, estradiol-17 β -glucuronide (median bioactivity value [μ M]=6.309) forms an interaction via the R-17 substituent (carbonyl oxygen) with ASN178 and ARG181, and backbone interaction of SER576 with the R-3 substituent. However, no distinct interaction nor effect of chlorine atom at the R-4 position was identified.

In OATP1B3, binding of halogenated 13-epiestrones is accomplished via ASN213 (the R-3 substituent), GLU185 (the R-17 position).

However, halogenated substituents do not impose any distinct interactions. It appears interesting to compare the identified halogen binding site in OATP2B1 ('site 1') to the putative sites in OATP1B1/OATP1B3 (see Figure 11). Mapping an electrostatic surface onto the transporter structure reveals that the N-terminal binding pocket is highly positively charged in OATP1B1/OATP1B3, compared to the OATP2B1. The increased electrostatic surface in OATP1B1 and OATP1B3 is caused by the presence of positively charged residues (such as LYS41 and LYS49), which are replaced by the non-charged residues in OATP2B1. As halogens prefer to bind to a hydrophobic environment [REF], the preference for OATP2B1 becomes evident. Moreover, the non-conserved THR133 seems to be another reason for the preferential halogen bond formation, compared to the OATP1B1 (ALA112 at the corresponding position) and OATP1B3 (SER112).

As already stated in the previous section, OATP1B3-selective binding of digoxin was

explained by the increased pocket volume in OATP1B3 (551 Å³) compared to OATP1B1 (510 Å³). In OATP2B1, no similar pose could be reproduced. The reason could be a more hydrophobic pocket environment compared to OATP1B3, as well as the presence of SER66 and thus increase of bulk water which might lead to steric clashes.

Overall, the conclusions drawn from these observations suggest that selectivity of hepatic OATPs is controlled by a limited number of residues at the TMH1/TMH2 interface, which affects pocket size and specific interactions.

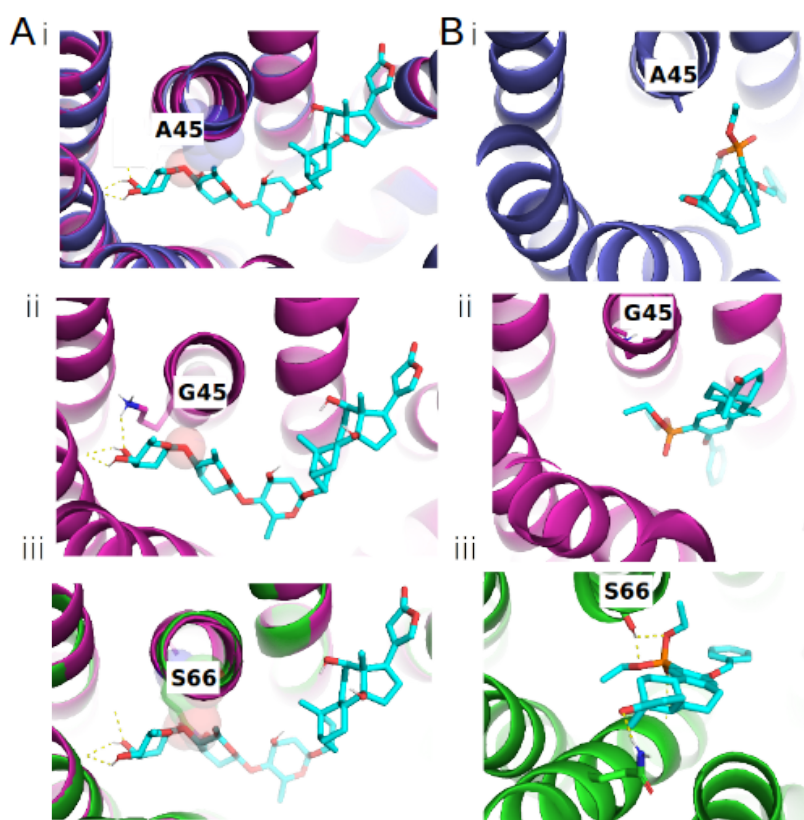


Figure 10

Selectivity switches across the hepatic OATPs. (A) Digoxin, (B) Phosphonated 13-epiestrones. Color coding:

docked compounds = cyan (carbon), red (oxygen), orange (phosphorus), OATP1B1 = blue, OATP1B3=magenta, OATP2B1 = green. Van der Waals radius of ALA45 (Ai), GLY45 (Aii), and SER66 (Aiii) is visualized as a transparent sphere. Poses are shown in the top view (from the extracellular part).

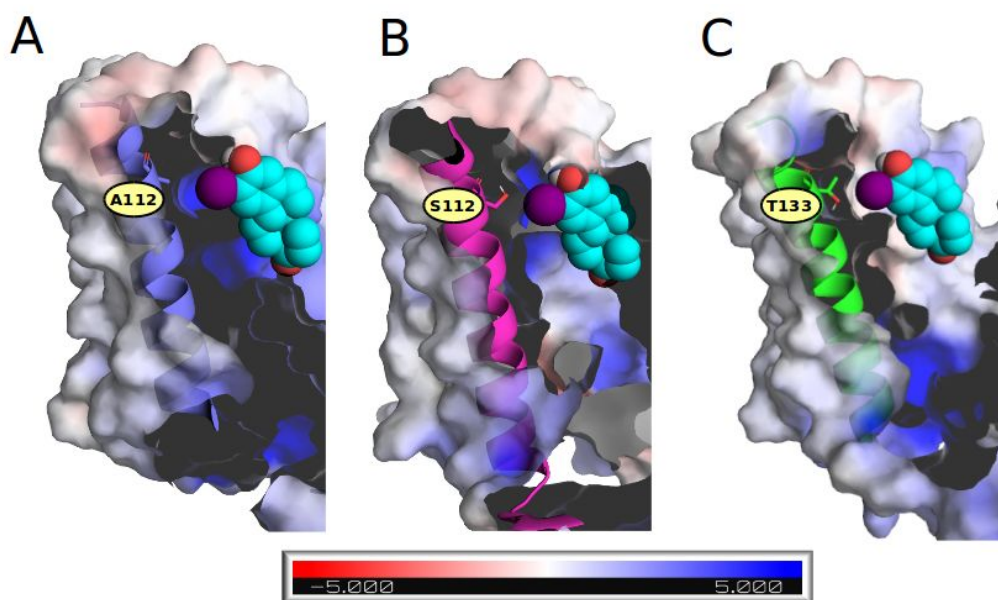


Figure 11

Putative halogen binding site in (A) OATP1B1 (B) OATP1B3 compared to the docked pose in (C) OATP2B1. Color coding: docked compound = cyan (carbon), red (oxygen), purple (iodine), OATP1B1 = blue, OATP1B3=magenta, OATP2B1 = green.

4. Summary and Conclusions

Molecular modeling of OATP-ligand interactions remains challenging due to the lack of

detailed knowledge about the protein structure. Here, we present an integrative computational approach, involving a systematic exploration of available structures with MFS fold by normal mode analysis, construction of multiple OATP models based on alternate conformations of selected templates, prioritization of the models on the basis of enrichment docking, and an in-depth analysis of molecular interactions for steroid analogs.

Signature dynamics of MFS proteins shows conserved fluctuations of specific protein subdomains. Among others, an increased flexibility at the TMH1/TMH2 interface was found to be an important determinant which contributes to the intrinsic dynamics of MFS proteins. These findings suggest functional importance of the TMH1/TMH2 interface. Therefore, the selected structural template (here: FucP transporter) was sampled along these modes in such a way that the final conformational ensemble covers movement of TMH1 and TMH2 helices. The calculated network models of the template were subsequently used to build OATP structural models trapped in different conformations. Enrichment docking of a large data set retrieved from the open domain (spiked with decoys from DUD-E database) was employed to validate the structural models on the basis of their ability to enrich known actives. Top prioritized models for OATP1B1, OATP1B3, and OATP2B1, are showing an out-of-plane shift of their TMH1/TMH2 helices. Structural models exhibit a different shape of the N-terminal binding site compared to the ones generated on the basis of the initial template structure. As shown in the Results & discussion section, a shift in TMH1/TMH2 was found to be crucial for the recognition of steroid analogs.

Mutational experiments for OATP1B1 have shown an involvement of residues at TMH1 (LYS41, GLY/ALA45, LYS49)(72) and TMH2 (ASP70, PHE73, GLU74, and GLY76) (27,70) in the transport of natural substrates (mostly estrone-3-sulfate or taurocholate). Similarly, alanine scanning of TMH1 on OATP2B1 helped identify key residues (VAL52, HIS55, GLN59, ALA61, GLN62, SER66, and LEU69) implicated in the transport of estrone-3-sulfate and taurocholate. (79) Thus, interactions of ligands with the TMH1/TMH2 interface seem to be of high relevance, as also shown by the docking results for steroid analogs in this paper. Interestingly, by retrospective re-docking of steroid analogs into the models based on the initial template structure, the established binding modes for steroids could not be fully reproduced.

Docked steroids generally show orientational versatility in the binding site(s). Specifically, the R-3 and R-17 substituents are capable of forming most of the key interactions.

Cluster analysis reveals two distinct sites for OATP1B1/OATP1B3 - one site in the central cavity and one N-terminal binding site. The site in the central cavity reveals conserved interactions between OATP1B1 and OATP1B3. Experimental data showed that several steroids (such as estrone-3-sulfate) exhibit biphasic kinetics, which led to the identification of low- and high-affinity binding sites. (27,68,80) Our computational analyses show two distinct binding sites in OATP1B1 and OATP1B3, which is in line with the published experimental data. Specifically, Li et al have shown that the alanine mutants of ASP70 in OATP1B1 affect both low- and high-affinity components, while PHE73, GLU74, and GLY76 only affects single site. As shown in Figure 8Ai, ASP70 is located in the upper part of TMH2, thus being partially involved in both the N-terminal binding site and inner cavity. On the contrary, the other mutated residues are oriented into the inner cavity to a larger extent. In another study on OATP1B1, LYS41 and LYS49 (both located at THM1) showed impaired K_m values at the high- and low-affinity binding site of estrone-3-sulfate, respectively. Moreover, LYS41ALA mutation possesses similar K_m values to the wild-type transporter for the low-affinity site of estrone-3-sulfate, thus suggesting that LYS41 is only implicated in a single (high-affinity) binding site. On the contrary, LYS49ALA mutation affected the low affinity component of estrone-3-sulfate. By comparing these findings to our OATP structural models, we reveal that LYS41 is buried in the N-terminal binding region, while LYS49 is exposed more towards the central cavity of a transporter. Mutagenesis study done to elucidate OATP2B1 transport has shown an effect of HIS579 at TMH7 on the modulation of the low-affinity binding site of estrone-3-sulfate. Interestingly, HIS579 is replaced by GLY552 in OATP1B1 and OATP1B3, which was identified by our computational analysis to be involved in the ligand binding in the central region of OATP1B1/OATP1B3 (see Table 4). Overall, the experimental data reported in the literature suggest that the N-terminal region might correspond to the high-affinity binding site for steroids, while the central region represents the low-affinity binding site. However, to fully assess which one of the two regions identified in this study might represent a high- or low-affinity site, free energy calculations have to be undertaken in future studies.

In contrast, OATP2B1 shows a single, N-terminal binding site, for interactions with steroid

analogs. However, we have to note that the conclusions drawn for OATP2B1 might be affected by the limited structural diversity of the OATP2B1 dataset, which is almost exclusively composed of 13-epiestrone structures. By comparing interactions adapted by OATP1B1 and OATP1B3 in the N-terminal binding site, differences in ligand accessibility appear to be the main cause for variations in ligand binding... This is partly affected by differently formed salt bridges in the cavity, as well as by the replacement of alanine (OATP1B1) at positions 45 and 216 to glycine (OATP1B3), which leads to disparities in pocket geometry. In OATP2B1, a single N-terminal binding site at the TMH1/TMH2 interface accommodates most of the OATP2B1 active steroid analogs. As a special use case, a distinct halogen binding site with the likelihood of halogen bond formation in the upper part of the N-terminal half was identified. The corresponding binding site was not found in OATP1B1 and OATP1B3, likely due to the replacement of THR133 (in OATP2B1) to ALA112 (in OATP1B1) and SER112 (in OATP1B3). Moreover, OATP2B1-specific binding of halogenated epiestrones likely happens due to the different composition of the electrostatic surface in OATP1B1/1B3 (positively charged) versus OATP2B1 (hydrophobic). As a follow up study, we plan to perform quantum mechanics optimization of the binding modes possessing halogen bonding.

Interestingly, a single residue at TMH1 (residue number 45 in OATP1B1/1B3, 66 in OATP2B1) seems to play a role as OATP selective switch given its chemical specificity and regiospecificity. In this paper, chemical specificity was exemplified, e.g., for phosphonated 13-epiestrones by their ability to form a hydrogen bond with SER66 in OATP2B1 (but not with ALA45 or GLY45 in OATP1B1/1B3). Regiospecificity was demonstrated for digoxin binding (a selective OATP3B1 ligand), where the size of the binding site is crucial for accommodating its bulky substituent at position R-3 of the steroidal core. As the N-terminal binding site in OATP1B3 possesses the highest volume - given the loss of the side chain at position 45 and 216, respectively (both GLY) - the OATP1B3-selective binding of digoxin can be explained. The role of residue number 45 for OATP1B1/1B3 selectivity was previously confirmed by mutational studies replacing GLY45 (OATP1B3) to ALA45 (OATP1B1) and vice versa. In this study, we found an indication that binding of steroids to SER66 in OATP2B1 could drive activity towards this transporter (as also shown in OATP2B1 mutational studies), and that ALA45/GLY45/SER66 might be responsible for

driving selectivity of steroid analogs across the three hepatic OATPs.

The TMH1/TMH2 interface has previously been identified as an essential substrate binding cavity for different types of MFS proteins, including the POT family of oligopeptide transporters.⁽⁸¹⁾ In future studies, different transporters with MFS fold should be explored in order to investigate whether ligand recognition at the N-terminal domain of MFS transporter represents a consistent pattern.

With this study, we ultimately contributed to the knowledge about structural determinants of hepatic OATPs, using a rigorous computational protocol for generating structural models, followed by the comparative analysis of important ligand interactions at the identified binding sites.

Insights about the interactions of steroid analogs with OATP1B1, OATP1B3, and OATP2B1 are contributing to the knowledge of compound requirements for the design of new chemical probes, which can further elucidate the physiological role of these emerging transporters.

Acknowledgments

This work received funding from the Austrian Science Fund (FWF) (Grant P 29712, “Elucidating hepatic OATP-ligand interactions and selectivity”). Financial support was also received by the National Research, Development and Innovation Office [OTKA FK 128751, Cs. Ö-L]. We thank Lars Richter for critical discussion especially concerning the ensemble docking procedure.

References

1. Lin L, Yee SW, Kim RB, Giacomini KM. SLC transporters as therapeutic targets: emerging opportunities. *Nat Rev Drug Discov.* 2015 Aug;14(8):543–60.
2. Giacomini KM, Huang S-M, Tweedie DJ, Benet LZ, Brouwer KLR, Chu X, et al. Membrane transporters in drug development. *Nat Rev Drug Discov.* 2010 Mar;9(3):215–36.
3. Stieger B, Hagenbuch B. Chapter Five - Organic Anion-Transporting Polypeptides. In: Bevensee MO, editor. *Current Topics in Membranes* [Internet]. Academic Press; 2014 [cited 2020 Dec 15]. p. 205–32. (Exchangers; vol. 73). Available from:

<http://www.sciencedirect.com/science/article/pii/B9780128002230000050>

4. Hagenbuch B, Gui C. Xenobiotic transporters of the human organic anion transporting polypeptides (OATP) family. *Xenobiotica*. 2008 Aug 1;38(7–8):778–801.
5. Shitara Y, Maeda K, Ikejiri K, Yoshida K, Horie T, Sugiyama Y. Clinical significance of organic anion transporting polypeptides (OATPs) in drug disposition: their roles in hepatic clearance and intestinal absorption. *Biopharm Drug Dispos*. 2013;34(1):45–78.
6. Kalliokoski A, Niemi M. Impact of OATP transporters on pharmacokinetics. *Br J Pharmacol*. 2009;158(3):693–705.
7. Cvetkovic M, Leake B, Fromm MF, Wilkinson GR, Kim RB. OATP and P-Glycoprotein Transporters Mediate the Cellular Uptake and Excretion of Fexofenadine. *Drug Metab Dispos*. 1999 Aug 1;27(8):866–71.
8. Obaidat A, Roth M, Hagenbuch B. The Expression and Function of Organic Anion Transporting Polypeptides in Normal Tissues and in Cancer. *Annu Rev Pharmacol Toxicol*. 2012;52(1):135–51.
9. Seithel A, Eberl S, Singer K, Auge D, Heinkele G, Wolf NB, et al. The Influence of Macrolide Antibiotics on the Uptake of Organic Anions and Drugs Mediated by OATP1B1 and OATP1B3. *Drug Metab Dispos*. 2007 May 1;35(5):779–86.
10. Kindla J, Müller F, Mieth M, Fromm MF, König J. Influence of Non-Steroidal Anti-Inflammatory Drugs on Organic Anion Transporting Polypeptide (OATP) 1B1- and OATP1B3-Mediated Drug Transport. *Drug Metab Dispos*. 2011 Jun 1;39(6):1047–53.
11. Khor BY, Tye GJ, Lim TS, Choong YS. General overview on structure prediction of twilight-zone proteins. *Theor Biol Med Model*. 2015 Sep 4;12(1):15.
12. Shaikh N, Sharma M, Garg P. Selective Fusion of Heterogeneous Classifiers for Predicting Substrates of Membrane Transporters. *J Chem Inf Model*. 2017 Mar 27;57(3):594–607.
13. Badolo L, Rasmussen LM, Hansen HR, Sveigaard C. Screening of OATP1B1/3 and OCT1 inhibitors in cryopreserved hepatocytes in suspension. *Eur J Pharm Sci*. 2010 Jul 11;40(4):282–8.
14. Soars MG, Barton P, Ismail M, Jupp R, Riley RJ. The Development, Characterization, and Application of an OATP1B1 Inhibition Assay in Drug Discovery. *Drug Metab Dispos*.

2012 Aug 1;40(8):1641–8.

15. Karlgren M, Ahlin G, Bergström CAS, Svensson R, Palm J, Artursson P. In Vitro and In Silico Strategies to Identify OATP1B1 Inhibitors and Predict Clinical Drug–Drug Interactions. *Pharm Res.* 2012 Feb 1;29(2):411–26.
16. Karlgren M, Vildhede A, Norinder U, Wisniewski JR, Kimoto E, Lai Y, et al. Classification of Inhibitors of Hepatic Organic Anion Transporting Polypeptides (OATPs): Influence of Protein Expression on Drug–Drug Interactions. *J Med Chem.* 2012 May 24;55(10):4740–63.
17. van de Steeg E, Venhorst J, Jansen HT, Nooijen IHG, DeGroot J, Wortelboer HM, et al. Generation of Bayesian prediction models for OATP-mediated drug–drug interactions based on inhibition screen of OATP1B1, OATP1B1*15 and OATP1B3. *Eur J Pharm Sci.* 2015 Apr 5;70:29–36.
18. Bruyn TD, Westen GJP van, IJzerman AP, Stieger B, Witte P de, Augustijns PF, et al. Structure-Based Identification of OATP1B1/3 Inhibitors. *Mol Pharmacol.* 2013 Jun 1;83(6):1257–67.
19. Chang C, Pang KS, Swaan PW, Ekins S. Comparative Pharmacophore Modeling of Organic Anion Transporting Polypeptides: A Meta-Analysis of Rat Oatp1a1 and Human OATP1B1. *J Pharmacol Exp Ther.* 2005 Aug 1;314(2):533–41.
20. Türková A, Jain S, Zdrazil B. Integrative Data Mining, Scaffold Analysis, and Sequential Binary Classification Models for Exploring Ligand Profiles of Hepatic Organic Anion Transporting Polypeptides. *J Chem Inf Model.* 2019 May 28;59(5):1811–25.
21. Wang P, Hata S, Xiao Y, Murray JW, Wolkoff AW. Topological assessment of oatp1a1: a 12-transmembrane domain integral membrane protein with three N-linked carbohydrate chains. *Am J Physiol-Gastrointest Liver Physiol.* 2008 Apr 1;294(4):G1052–9.
22. Meier-Abt F, Mokrab Y, Mizuguchi K. Organic Anion Transporting Polypeptides of the OATP/SLCO Superfamily: Identification of New Members in Nonmammalian Species, Comparative Modeling and a Potential Transport Mode. *J Membr Biol.* 2006 Jan 1;208(3):213–27.
23. Taank V, Zhou W, Zhuang X, Anderson JF, Pal U, Sultana H, et al. Characterization of tick organic anion transporting polypeptides (OATPs) upon bacterial and viral infections. *Parasit Vectors.* 2018 Nov 14;11(1):593.

24. Alam K, Crowe A, Wang X, Zhang P, Ding K, Li L, et al. Regulation of Organic Anion Transporting Polypeptides (OATP) 1B1- and OATP1B3-Mediated Transport: An Updated Review in the Context of OATP-Mediated Drug-Drug Interactions. *Int J Mol Sci*. 2018 Mar;19(3):855.
25. Hagenbuch B, Meier PJ. The superfamily of organic anion transporting polypeptides. *Biochim Biophys Acta BBA - Biomembr*. 2003 Jan 10;1609(1):1–18.
26. Mandery K, Sticht H, Bujok K, Schmidt I, Fahrmayr C, Balk B, et al. Functional and Structural Relevance of Conserved Positively Charged Lysine Residues in Organic Anion Transporting Polypeptide 1B3. *Mol Pharmacol*. 2011 Sep 1;80(3):400–6.
27. Li N, Hong W, Huang H, Lu H, Lin G, Hong M. Identification of amino acids essential for estrone-3-sulfate transport within transmembrane domain 2 of organic anion transporting polypeptide 1B1. *PLoS One*. 2012;7(5):e36647.
28. Hong W, Wu Z, Fang Z, Huang J, Huang H, Hong M. Amino Acid Residues in the Putative Transmembrane Domain 11 of Human Organic Anion Transporting Polypeptide 1B1 Dictate Transporter Substrate Binding, Stability, and Trafficking. *Mol Pharm*. 2015 Dec 7;12(12):4270–6.
29. Glaeser H, Mandery K, Sticht H, Fromm MF, König J. Relevance of conserved lysine and arginine residues in transmembrane helices for the transport activity of organic anion transporting polypeptide 1B3. *Br J Pharmacol*. 2010;159(3):698–708.
30. Khuri N, Zur AA, Wittwer MB, Lin L, Yee SW, Sali A, et al. Computational Discovery and Experimental Validation of Inhibitors of the Human Intestinal Transporter OATP2B1. *J Chem Inf Model*. 2017 Jun 26;57(6):1402–13.
31. Türková A, Zdrážil B. Current Advances in Studying Clinically Relevant Transporters of the Solute Carrier (SLC) Family by Connecting Computational Modeling and Data Science. *Comput Struct Biotechnol J*. 2019 Jan 1;17:390–405.
32. Ponzoni L, Zhang S, Cheng MH, Bahar I. Shared dynamics of LeuT superfamily members and allosteric differentiation by structural irregularities and multimerization. *Philos Trans R Soc B Biol Sci*. 2018 Jun 19;373(1749):20170177.
33. Bakan A, Bahar I. The intrinsic dynamics of enzymes plays a dominant role in determining the structural changes induced upon inhibitor binding. *Proc Natl Acad Sci*. 2009 Aug 25;106(34):14349–54.

34. Bowie JU, Luthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. *Science*. 1991 Jul 12;253(5016):164–70.
35. Lobley A, Sadowski MI, Jones DT. pGenTHREADER and pDomTHREADER: new methods for improved protein fold recognition and superfamily discrimination. *Bioinformatics*. 2009 Jul 15;25(14):1761–7.
36. Dang S, Sun L, Huang Y, Lu F, Liu Y, Gong H, et al. Structure of a fucose transporter in an outward-open conformation. *Nature*. 2010 Oct;467(7316):734–8.
37. Pei J, Grishin NV. PROMALS3D: Multiple Protein Sequence Alignment Enhanced with Evolutionary and Three-Dimensional Structural Information. In: Russell DJ, editor. *Multiple Sequence Alignment Methods* [Internet]. Totowa, NJ: Humana Press; 2014 [cited 2020 Sep 23]. p. 263–71. (Methods in Molecular Biology). Available from: https://doi.org/10.1007/978-1-62703-646-7_17
38. Carlsson J, Coleman RG, Setola V, Irwin JJ, Fan H, Schlessinger A, et al. Ligand discovery from a dopamine D 3 receptor homology model and crystal structure. *Nat Chem Biol*. 2011 Nov;7(11):769–78.
39. Bakan A, Meireles LM, Bahar I. ProDy: Protein Dynamics Inferred from Theory and Experiments. *Bioinformatics*. 2011 Jun 1;27(11):1575–7.
40. Holm L, Laakso LM. Dali server update. *Nucleic Acids Res*. 2016 Jul 8;44(W1):W351–5.
41. Lezon TR, Bahar I. Constraints Imposed by the Membrane Selectively Guide the Alternating Access Dynamics of the Glutamate Transporter GltPh. *Biophys J*. 2012 Mar 21;102(6):1331–40.
42. Lomize MA, Pogozheva ID, Joo H, Mosberg HI, Lomize AL. OPM database and PPM web server: resources for positioning of proteins in membranes. *Nucleic Acids Res*. 2012 Jan 1;40(D1):D370–6.
43. Tama F, Gadea FX, Marques O, Sanejouand Y-H. Building-block approach for determining low-frequency normal modes of macromolecules. *Proteins Struct Funct Bioinforma*. 2000;41(1):1–7.
44. Kabsch W, Sander C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*. 1983;22(12):2577–637.
45. Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, Hess B, et al. GROMACS: High

- performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*. 2015;1:19–25.
46. Lindorff-Larsen K, Piana S, Palmo K, Maragakis P, Klepeis JL, Dror RO, et al. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins Struct Funct Bioinforma*. 2010;78(8):1950–1958.
47. Williams CJ, Headd JJ, Moriarty NW, Prisant MG, Videau LL, Deis LN, et al. MolProbity: More and better reference data for improved all-atom structure validation. *Protein Sci*. 2018;27(1):293–315.
48. Adams PD, Afonine PV, Bunkóczi G, Chen VB, Davis IW, Echols N, et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr*. 2010 Feb 1;66(2):213–21.
49. Laczkó-Rigó R, Jójárt R, Mernyák E, Bakos É, Tuerkova A, Zdrazil B, et al. Structural dissection of 13-epiestrones based on the interaction with human Organic anion-transporting polypeptide, OATP2B1. *J Steroid Biochem Mol Biol*. 2020;105652.
50. Mysinger MM, Carchia M, Irwin JohnJ, Shoichet BK. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J Med Chem*. 2012 Jul 26;55(14):6582–94.
51. Trott O, Olson AJ. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem*. 2010;31(2):455–61.
52. Tuerkova A, Zdrazil B. A ligand-based computational drug repurposing pipeline using KNIME and Programmatic Data Access: case studies for rare diseases and COVID-19. *J Cheminformatics*. 2020;12(1):1–20.
53. Bemis GW, Murcko MA. The Properties of Known Drugs. 1. Molecular Frameworks. *J Med Chem*. 1996 Jan 1;39(15):2887–93.
54. O’Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: An open chemical toolbox. *J Cheminformatics*. 2011;3(1):33.
55. Varela-Salinas G, García-Pérez CA, Peláez R, Rodríguez AJ. Visual Clustering Approach for Docking Results from Vina and AutoDock. In: Martínez de Pisón FJ, Urraca R, Quintián H, Corchado E, editors. *Hybrid Artificial Intelligent Systems*. Cham: Springer International Publishing; 2017. p. 342–53. (Lecture Notes in Computer Science).

56. Berthold MR, Cebon N, Dill F, Gabriel TR, Kötter T, Meinl T, et al. KNIME - the Konstanz information miner: version 2.0 and beyond. *ACM SIGKDD Explor Newsl.* 2009 Nov 16;11(1):26–31.
57. Molecular Operating Environment (MOE) 2013 08. Chemical Computing Group ULC. 1010 Sherbooke St West Suite 910 Montr QC Can H3A 2R7 2018. 2018;
58. Wagner JR, Sørensen J, Hensley N, Wong C, Zhu C, Perison T, et al. POVME 3.0: software for mapping binding pocket flexibility. *J Chem Theory Comput.* 2017;13(9):4584–4592.
59. Patik I, Székely V, Német O, Szepesi Á, Kucsma N, Várady G, et al. Identification of novel cell-impermeant fluorescent substrates for testing the function and drug interaction of Organic Anion-Transporting Polypeptides, OATP1B1/1B3 and 2B1. *Sci Rep.* 2018;8(1):1–12.
60. Székely V, Patik I, Ungvári O, Telbisz Á, Szakács G, Bakos É, et al. Fluorescent probes for the dual investigation of MRP2 and OATP1B1 function and drug interactions. *Eur J Pharm Sci.* 2020;105395.
61. Deng D, Sun P, Yan C, Ke M, Jiang X, Xiong L, et al. Molecular basis of ligand recognition and transport by glucose transporters. *Nature.* 2015;526(7573):391–396.
62. Chang S, Li K, Hu J, Jiao X, Tian X. Allosteric and transport behavior analyses of a fucose transporter with network models. *Soft Matter.* 2011;7(10):4661–4671.
63. Jiang D, Zhao Y, Fan J, Liu X, Wu Y, Feng W, et al. Atomic resolution structure of the E. coli YajR transporter YAM domain. *Biochem Biophys Res Commun.* 2014;450(2):929–935.
64. Jiang D, Zhao Y, Wang X, Fan J, Heng J, Liu X, et al. Structure of the YajR transporter suggests a transport mechanism based on the conserved motif A. *Proc Natl Acad Sci.* 2013;110(36):14664–14669.
65. Deng D, Xu C, Sun P, Wu J, Yan C, Hu M, et al. Crystal structure of the human glucose transporter GLUT1. *Nature.* 2014;510(7503):121–125.
66. Oka Y, Asano T, Shibasaki Y, Lin J-L, Tsukuda K, Katagiri H, et al. C-terminal truncated glucose transporter is locked into an inward-facing form without transport activity. *Nature.* 1990;345(6275):550–553.

67. Sun L, Zeng X, Yan C, Sun X, Gong X, Rao Y, et al. Crystal structure of a bacterial homologue of glucose transporters GLUT1–4. *Nature*. 2012 Oct;490(7420):361–6.
68. Hoshino Y, Fujita D, Nakanishi T, Tamai I. Molecular localization and characterization of multiple binding sites of organic anion transporting polypeptide 2B1 (OATP2B1) as the mechanism for substrate and modulator dependent drug–drug interaction. *MedChemComm*. 2016;7(9):1775–1782.
69. Panek A, Świzdor A, Milecka-Tronina N, Panek JJ. Insight into the orientational versatility of steroid substrates—a docking and molecular dynamics study of a steroid receptor and steroid monooxygenase. *J Mol Model*. 2017;23(3):96.
70. Lee HH, Leake BF, Teft W, Tirona RG, Kim RB, Ho RH. Contribution of hepatic organic anion-transporting polypeptides to docetaxel uptake and clearance. *Mol Cancer Ther*. 2015;14(4):994–1003.
71. Miao Y, Hagenbuch B. Conserved positively charged amino acid residues in the putative binding pocket are important for OATP1B1 function. *FASEB J*. 2007;21(5):A196–A197.
72. DeGorter MK, Ho RH, Leake BF, Tirona RG, Kim RB. Interaction of three regiospecific amino acid residues is required for OATP1B1 gain of OATP1B3 substrate specificity. *Mol Pharm*. 2012;9(4):986–995.
73. Jójárt R, Pécsy S, Keglevich G, Szécsi M, Rigó R, Özveggy-Laczka C, et al. Pd-Catalyzed microwave-assisted synthesis of phosphonated 13 α -estrone as potential OATP2B1, 17 β -HSD1 and/or STS inhibitors. *Beilstein J Org Chem*. 2018;14(1):2838–2845.
74. Auffinger P, Hays FA, Westhof E, Ho PS. Halogen bonds in biological molecules. *Proc Natl Acad Sci*. 2004;101(48):16789–16794.
75. Wilcken R, Zimmermann MO, Lange A, Joerger AC, Boeckler FM. Principles and applications of halogen bonding in medicinal chemistry and chemical biology. *J Med Chem*. 2013;56(4):1363–1388.
76. Kuhn B, Gilberg E, Taylor R, Cole J, Korb O. How Significant Are Unusual Protein–Ligand Interactions? Insights from Database Mining. *J Med Chem*. 2019;62(22):10441–10455.
77. Riley KE, Tran K-A. Strength, character, and directionality of halogen bonds involving cationic halogen bond donors. *Faraday Discuss*. 2017;203:47–60.
78. Priimagi A, Cavallo G, Metrangolo P, Resnati G. The halogen bond in the design of

functional supramolecular materials: recent advances. *Acc Chem Res.* 2013;46(11):2686–2695.

79. Fang Z, Huang J, Chen J, Xu S, Xiang Z, Hong M. Transmembrane Domain 1 of Human Organic Anion Transporting Polypeptide 2B1 Is Essential for Transporter Function and Stability. *Mol Pharmacol.* 2018 Aug 1;94(2):842–9.
80. Wang X, Chen J, Xu S, Ni C, Fang Z, Hong M. Amino-terminal region of human organic anion transporting polypeptide 1B1 dictates transporter stability and substrate interaction. *Toxicol Appl Pharmacol.* 2019 Sep 1;378:114642.
81. Solcan N, Kwok J, Fowler PW, Cameron AD, Drew D, Iwata S, et al. Alternating access mechanism in the POT family of oligopeptide transporters. *EMBO J.* 2012 Aug 15;31(16):3411–21.

3.6 Combining AI-driven and Structure-based Approaches to Identify Novel Inhibitors of Hepatic Organic Anion Transporting Polypeptides (OATPs)

TUERKOVA, Alzbeta; BONGERS, Brandon J.; NORINDER, Ulf; UNGVÄRI, Orsolya; SZÉKELY, Virág; ÖZVEGY-LACZKA, Csilla; SZAKÁCS, Gergely; VAN WESTEN, Gerard J.P.; ZDRAZIL, Barbara. **2020** *In preparation*

* *Corresponding author: barbara.zdrazil@univie.ac.at*

A. Tuerkova performed structure-based virtual screening, data analysis, and established binding mode hypothesis for the experimentally-confirmed hits. B.J. Bongers, under the supervision of GJP. van Westen, developed proteochemometric and QSAR models, and performed virtual screening with respective models. U. Norinder developed conformal prediction models and performed virtual screening with respective models. C. Laczka and co-workers performed transport inhibition experiments with suggested hits and estimated IC₅₀ values. B. Zdrazil conceived the study and provided supervision for the structure-based virtual screening. A. Tuerkova wrote the following chapter.

The Supplementary Information can be found in Part V.

For references, see Bibliography.

3.6.1 Introduction

Integrating artificial intelligence with structure-based approaches into a single virtual screening pipeline is a promising strategy to detect novel compounds. [131] Since hepatic OATPs are largely understudied, virtual screening followed by subsequent experimental testing of the detected hits can be leveraged. Novel ligands would in turn enhance our understanding about molecular determinants and structural aspects of hepatic OATPs driving ligand binding. The most comprehensive screening study for hepatic OATPs done so far was performed by Karlgren et al. [27] Specifically, 225 drug-like compounds were tested for their activity on OATP1B1, OATP1B3, and OATP2B1. 91 OATP inhibitors with different overlapping profiles across the three hepatic OATPs were identified. Among those, some specific OATP1B1 (pravastatin, $IC_{50} = 3.6 \mu M$), and OATP2B1 (erlotinib, $IC_{50} = 0.55 \mu M$) inhibitors, were found. In another screening study, several OATP1B1 inhibitors with KC_i values ranging from 0.06 to $6.5 \mu M$ were identified and proteochemometric models were subsequently developed by using *in vitro* data. [132] Next, Khuri et al identified novel OATP2B1 inhibitors by a combination of random forest models and multiple structural models. [133]

In the presented study, we performed a consensus screening approach by using different types of machine learning models (proteochemometric models, conformal prediction models, and deep learning models for hepatic OATPs), followed by structure-based virtual screening of preselected hits using the structural models for hepatic OATPs generated in Study 5. Screening the diverse REAL drug-like set (a subset of ENAMINE REAL with 21M compounds) [134] has shown a comparable hit rate for OATP1B1/OATP1B3 (32%), while the hit rate for OATP2B1 was significantly higher (70.5%). By full dose-response measurement, IC_{50} values for the prioritized compounds were determined. Several strong OATP inhibitors were selected and the binding mode hypothesis was established. Structural comparison of the detected binding sites across the three transporters shows remarkable differences in the localization of aromatic residues in OATP1B1/OATP1B3 compared to OATP2B1, which helps explain different hit-rates delivered for the three hepatic transporters.

3.6.2 Material and Methods

Machine learning models

A dataset of (non-)inhibitors from the open domain published in Study 1 [135] was used to develop three different types of machine learning models. This part was done in collaboration with Dr. Ulf Norinder (OATP conformal prediction models) and Brandon J. Bongers, MSc and Prof. Gerard JP van Westen (OATP deep learning and proteochemometric models). A detailed methodology for conformal prediction [136], as well as for proteochemometric modeling [137] can be found in the literature. In this chapter, the focus lies on the structure-based modeling part done by the author of this thesis.

Molecular docking of pre-selected hits

Here, comparative models for OATP1B1, OATP1B3, and OATP2B1 generated in Study 5 were used. Potential interaction sites in OATP1B1, OATP1B3, and OATP2B1 transporter structures, were mapped via a small molecule mapping server, FTMap (available at <https://ftmap.bu.edu/serverhelp.php>). [88] For each transporter, three to four possible sites were found. Interestingly, an equivalent binding cavity in all the three transporters was identified, being lined by TMH1, TMH2, TMH4, TMH5, TMH7, TMH8, and TMH11. Concrete residues belonging to this region are listed in Table 1 3.1.

Prioritization of the identified hits from docking calculations

A selection of 3,291 out of ~8,5M compounds obtained from the screens by using the developed machine learning models were docked in AutoDock Vina 1.1.2 (exhaustiveness of the global search was set to 10). In this study, three different categories of compounds were defined as follows: G1 class (potentially OATP1B1 selective compounds), G2 class (potentially OATP1B3 selective compounds), and G3 class (potentially OATP2B1 selective compounds). These categories were based on different machine learning models which were developed for each of the three transporters individually. In a first instance, compounds were sorted according to their docking score. Top 30 ranked compounds per category were kept. Next, physico-chemical properties (SlogP, TPSA, SMR, number of rotatable bonds, and AMW) were calculated. Physico-chemical properties calculated

TABLE 3.1: Residues belonging to the predicted binding site. Residue positions indicated in brackets are specific to OATP2B1 due to the structural differences between OATP1B1/OATP1B3 and OATP2B1 transporters.

Residue position (1B1/1B3/2B1)	OATP1B1	OATP1B3	OATP2B1
78/78/99	LEU	LEU	THR
189/189/207	VAL	VAL	GLN
213/213/231	ASN	ASN	PHE
216/216/234	ALA	GLY	THR
217/217/235	MET	MET	MET
348/348/383	GLN	GLN	LEU
349/349/384	VAL	VAL	SER
352/352/387	TYR	PHE	ALA
356/356/391	PHE	PHE	ALA
385/385/420	ILE	THR	SER
549/549/576	ALA	ALA	CYS
552/552/579	GLY	GLY	HIS
577/577/604	MET	MET	MET
580/580/607	ARG	ARG	ARG

here were previously shown to be important molecular determinants of hepatic OATP activity. [135] Therefore, our intention was to check whether the newly identified hits are falling within the range of the known OATP ligands. Outlier values of the relevant phys/chem properties were not considered for this analysis. After hits pre-filtering (SlogP [-0.318, 8.054], TPSA [0.000, 275.640], SMR [37.627, 245.137], number of rotatable bonds [0, 20], AMW [136.154, 936.921] for OATP1B1, SlogP [-0.318, 8.713], TPSA [0.000, 297.120], SMR [33.117, 292.120], number of rotatable bonds [0, 20], AMW [128.558, 1,093.331] for OATP1B3, and SlogP [-1.144, 8.948], TPSA [9.230, 241.880], SMR [40.361, 216.039], number of rotatable bonds [1, 17], AMW [144.214, 794.040] for OATP2B1, respectively), structural diversity of the retained hits was examined as follows: Similarities between the pairs of compounds were calculated. The maximum common substructure (MCS) of the compound pairs was defined as a similarity metric. A distance matrix was used to perform a hierarchical compound clustering. Complete linkage method was applied to perform hierarchical clustering. Compounds were assigned to a common cluster with the distance threshold of 0.5. A single compound per

each cluster was retained. Selection of a single representative per cluster was guided by the docking score of the corresponding compounds. Filtered compounds were further sorted according to the differences in the docking score to check which compounds are likely to preferentially interact with one of the three transporters. At the end, 15 compounds per class were kept.

The selection procedure and subsequent data analyses were performed using Konstanz Information Miner (version 4.0.). [50]

Generation and maintenance of cell lines

The experimental part of this study was done by Dr. Csilla Özvegy-Laczka and co-workers. A431 cells overexpressing OATP1B1, OATP1B3 or OATP2B1, or their mock transfected controls were generated previously [138], and were maintained in Dulbeccó modified eagle medium (DMEM, Gibco, Thermofisher Scientific, Waltham, MA, US) supplemented with 10% fetal bovine serum, 2 mM L-glutamine, 100 units/mL penicillin and 100 µg/mL streptomycin. Expression and function of OATPs in the cell lines was checked regularly.

Transport inhibition measurements

Interaction with OATP1B1, OATP1B3 and OATP2B1 was tested in an indirect transport assay using pyranine (8-Hydroxypyrene-1,3,6-trisulfonic acid trisodium salt, H1529, Sigma, Merck, Budapest, Hungary) as test substrate. [139, 140] A431 cells overexpressing OATP1B1, OATP1B3 or OATP2B1, or their mock transfected controls were seeded on 96-well plates in a density of 8×10^4 cells per well in 200 µl cell culture medium 1 day prior to the transport measurements. After 16-24 hours the medium was removed, the cells were washed three times with 200 µl PBS (phosphate buffered saline, pH 7.2) and preincubated for 5 minutes at 37°C with 50 µl uptake buffer (125 mM NaCl, 4.8 mM KCl, 1.2 mM CaCl₂, 1.2 mM KH₂PO₄, 12 mM MgSO₄, 25 mM MES [2-(N-morpholino)ethanesulfonic acid and 5.6 mM glucose, pH 5.5) with or without the tested compound. During the initial screen, the compounds were tested in three different concentrations, 1, 10 or 100 µM, though in some cases due to poor solubility the maximum concentrations were 20 or 50 µ. Each test compound was dissolved in

DMSO (that did not exceed 0.5% in samples); solvent controls were also applied. Hit compounds were then tested at 8 different concentrations (see Figure 3.1). Transport reaction was started by the addition of 50 μ l uptake buffer containing pyranine in a final concentration of 10 μ M (OATP1B1) or 20 μ M (OATP1B3 and OATP2B1), and the cells were further incubated at 37°C for 15 minutes (OATP1B1 and OATP2B1), or 30 minutes (OATP1B3). The reaction was stopped by removing the supernatant. After repeated washing with ice-cold PBS, fluorescence was determined in an Enspire plate reader (Perkin Elmer, Waltham, MA) with excitation/emission wavelengths of 460/510 nm. OATP-dependent transport was calculated by extracting fluorescence measured in mock transfected cells. Transport activity was calculated based on the fluorescence signal in the absence (100%) of the tested compounds. Experiments were repeated in at least 3 biological replicates.

Determining IC₅₀ values

IC₅₀ values were calculated by Hill1 fit, using the Origin Pro 2018 software (OriginLab Corporation, Northampton, MA, USA).

3.6.3 Results & Discussion

On the basis of our multistep computational protocol, 44 compounds which would potentially act as OATP inhibitors, were selected (Table 3.2 and Supplementary Figure S1). Initial experimental screens detected 32% of OATP1B1, 32% of OATP1B3, and 70.5% of OATP2B1 compounds with an activity threshold ≤ 10 μ M. The novel OATP dataset covers different areas of chemical space compared to the OATP bioactivity data gathered from the open domain. [135] Specifically, newly measured OATP inhibitors do possess a low degree structural similarity, as evidenced by comparison of their Murcko scaffolds. On the contrary, distribution of relevant physico-chemical properties remains similar across the novel dataset and the data set retrieved from the open domain. The only exception was found for the fraction of sp³-hybridized carbons (“FractionCSP3”), which is generally lower for novel compounds compared to the public domain (median value 0.25 versus 0.50 for OATP1B1, 0.23 versus 0.53 for OATP1B3 and 0.25 vs 0.35 for OATP2B1 inhibitors, respectively, see Supplementary Figure S2 to S4).

TABLE 3.2: Table showing 44 compounds measured in the initial screens. Four categories were determined: 50% inhibition observed below 1 μM (colored red), 50% inhibition between 1 and 10 μM (colored yellow), 50% inhibition above 10 μM (colored light gray), and no effect on transport (white). In addition, several compounds were identified as activators of the transport (colored blue).

Code	OATP1B1	OATP1B3	OATP2B1
C7	$\sim 10 \mu\text{M}$ / 10-50 μM	1-10 μM	$<1 \mu\text{M}$
H5	10-50 μM	10-25 μM	$<1 \mu\text{M}$
A5	no effect	no effect	$>25 \mu\text{M}$
E3	1-10 μM	1-10 μM	$\sim 1 \mu\text{M}$
E5	1-10 μM	$\sim 10 \mu\text{M}$	$\sim 1 \mu\text{M}$
C3	1-10 μM	10-100 μM	$\sim 10 \mu\text{M}$
D5	10-50 μM	10-50 μM	$\sim 10 \mu\text{M}$
A6	$\sim \mu\text{M}$ / no effect	$\sim 25 \mu\text{M}$	$\sim 25 \mu\text{M}$
A4	no effect	no effect	$\sim 50 \mu\text{M}$
A2	1-10 μM	1-10 μM	1-10 μM
A3	1-10 μM	10-25 μM	1-10 μM
A7	$>25 \mu\text{M}$ / no effect	activated	1-10 μM
B2	no effect	$>10 \mu\text{M}$	1-10 μM
B3	10-50 μM	10-50 μM	1-10 μM
B4	no effect	1-10 μM	1-10 μM
B5	10-100 μM	activated	1-10 μM
B7	1-10 μM	10-100 μM	1-10 μM
C2	1-10 μM	$\sim 10 \mu\text{M}$	1-10 μM
C6	1-10 μM	1-10 μM	1-10 μM
D6	10-100 μM	$\sim 10 \mu\text{M}$	1-10 μM
E2	10-50 μM	1-10 μM	1-10 μM
E4	$\sim 10 \mu\text{M}$	10-100 μM	1-10 μM
E6	10-50 μM	10-50 μM	1-10 μM
F2	$\sim 10 \mu\text{M}$	10-50 μM	1-10 μM
F5	1-10 μM	1-10 μM	1-10 μM
F6	10-100 μM	10-100 μM	1-10 μM
G2	1-10 μM	1-10 μM	1-10 μM
G3	$\sim 10 \mu\text{M}$	10-100 μM	1-10 μM
G4	no effect	no effect	1-10 μM

Table 3.2 continued from previous page

Code	OATP1B1	OATP1B3	OATP2B1
G5	10-25 μM	1-10 μM	1-10 μM
H2	10-100 μM	10-100 μM	1-10 μM
H3	1-10 μM	~ 10 μM	1-10 μM
H6	no effect	no effect	1-10 μM
D4	1-10 μM / 10-50 μM	~ 10 μM	1-10 μM (~ 1)
C4	> 50 μM	no effect	10-100 μM
D3	> 50 μM	no effect	10-100 μM
D7	10-100 μM	~ 100 μM	10-100 μM
G6	no effect	~ 100 μM	10-100 μM
F3	10-100 μM	10-100 μM	10-100 μM (~ 10)
B6	> 25 μM	10-25 μM	10-25 μM
C5	> 50 μM	no effect	10-25 μM
D2	10-50 μM	10-25 μM	10-50 μM
F4	~ 50 μM	activated	10-50 μM
H4	no effect	no effect	no effect

Volumetric maps for different pocket features (hydrophobicity, hydrophilicity, hydrogen bond donors/acceptors, and aromaticity, Supplementary Figure S5 - S9) show a remarkable difference in the localization of aromatic residues between OATP1B1/OATP1B3 and OATP2B1 transporters (Supplementary Figure S5). By a close inspection of the respective binding regions (lined by TMH5, TMH7, TMH8, and TMH11), several “aromatic residue-to-glycine/alanine” replacements were observed in OATP1B1/OATP1B3 compared to OATP2B1 (see Figure 3.1). Specifically, TYR352/PHE352 in OATP1B1/OATP1B3 at TMH7 are replaced by ALA387 in OATP2B1, PHE356 in OATP1B1/OATP1B3 at TMH7 is replaced by ALA391 in OATP2B1, and GLY552 in OATP1B1/OATP1B3 at TMH8 is replaced by HIS579 in OATP2B1. Other amino acid substitutions include ASN213 in OATP1B1/OATP1B3 at TMH5 being replaced by PHE231 in OATP2B1, VAL556/ILE556 in OATP1B1/OATP1B3 at TMH8 being replaced by PHE583 in OATP2B1, and SER576 in OATP1B1/OATP1B3 at TMH11 being replaced to PHE603.

Calculation of the electrostatic potential and mapping the surface onto the transporter's binding site shows the substitutions at positions (at TMH7) that are crucial for ligands to get partially accommodated in the sub cavity located in the C-terminal domain (Figure 3.2). Accession of the C-terminal sub cavity in OATP1B1 and OATP1B3 is blocked due to the presence of aromatic residues in TMH7 at positions 352 (387) and 356 (391) (PHE and TYR respectively). On the contrary, the electrostatic surface of OATP2B1 shows a small region at the TMH7/TMH8 interface which is accessible from the central cavity of the transporter (Supplementary Figure S15). Indeed, the strong OATP2B1 inhibitors identified in this study possess a high shape complementarity with the accessible surface in OATP2B1 (Supplementary Figure S15B). Further, the replacement of GLY552 in OATP1B1/OATP1B3 to HIS549 in OATP2B1 at TMH8 has an additional effect on the ligand recognition in the C-terminal domain, as it further restricts the space where the ligands can bind to. Since HIS579 in OATP2B1 is pointing towards the centre of the transporter cavity and thus restricts the translocation pore of the transporter, the question arises whether HIS579 could adopt different rotameric state, which in turn could have an impact on ligand binding. Therefore, rotamer analysis was performed to model alternative side chain orientations of HIS579. The probability of adopting different rotamers gets lower due to the steric clashes with the neighboring residues (see Supplementary Figure S14). We can therefore conclude that our structural model for OATP2B1 depicts the correct orientation of HIS579.

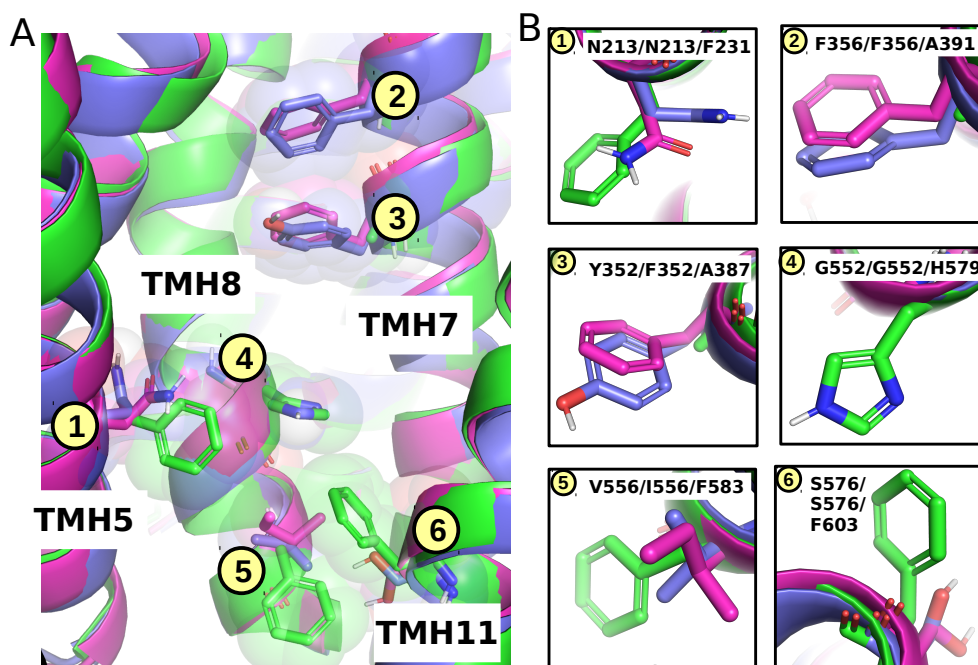


FIGURE 3.1: Different localization of the aromatic residues in OATP1B1 (the blue structure), OATP1B3 (the magenta structure), and OATP2B1 (the green structure) impacts pocket geometry and accessibility.

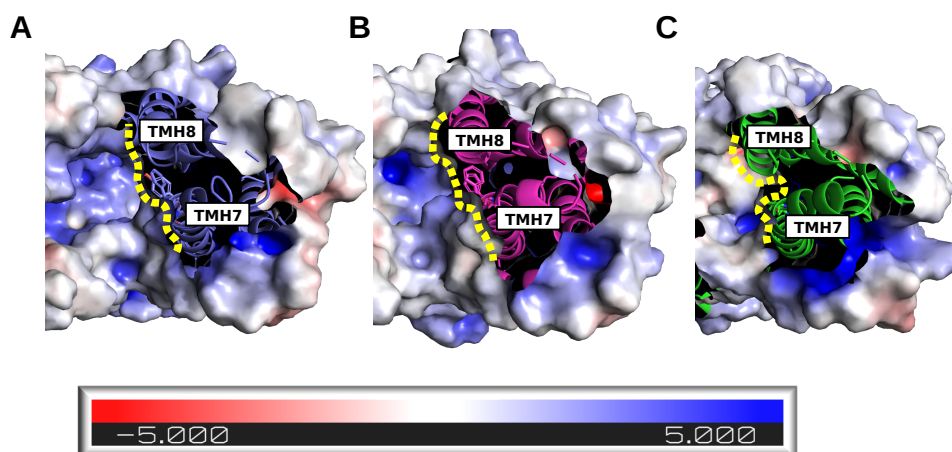


FIGURE 3.2: Electrostatic potentials mapped onto the protein surfaces of the (A) OATP1B1, (B) OATP1B3, and (C) OATP2B1 inner cavity (top view). Electrostatic potential ranges from the negative potential (red) through zero potential (white) to the positive potential (blue). Substitution of the two aromatic residues in OATP1B1/OATP1B3 at position 352 (387 in OATP2B1) and 356 (391 in OATP2B1) to alanine in OATP2B1 leads to the increase of bulk at the TMH7/TMH8 interface and therefore increases the accessibility of the TMH7/8 interface.

Calculation of protein-ligand interaction fingerprints (PLIFs) led to the identification of key residues which interact with the newly measured compounds (activity threshold $\leq 10 \mu\text{M}$, Supplementary Figure S10-S12). The most frequent residues implicated in ligand binding in OATP1B1 (ASN213, MET217, GLN348, GLY552, ARG580), and OATP1B3 (ASN213, MET217, GLY552, MET577, ARG580) are largely overlapping. The similarity of protein-ligand interactions between the two transporters might be attributed to their high sequence similarity ($\sim 80\%$). Interestingly, all of these residues also appeared to be implicated in binding of steroidal analogs from Study 5. These findings provide us with a consistent picture on the structural determinants for ligand binding in the central cavity, as already shown in Study 5 for OATP1B1 and OATP1B3 transporters. On the contrary, OATP2B1 ligand interactions point to the contribution of different residues compared to Study 5, such as PHE231, SER420, ALA575, CYS576, and HIS579. The majority of the frequently interacting residues are non-conserved across the OATP members (such as SER420, CYS576 and HIS579). Interestingly, HIS579 has already been confirmed by the mutational experiments to be crucial for OATP2B1 ligand recognition. [34]

From the set of 44 compounds measured in transport inhibition experiments, eight compounds have been selected for further IC_{50} value determination (see Figure 3.3). The manual selection of the eight compounds was based on the following criteria: a strong inhibitory potential in the initial screens, a tendency towards selectivity for one of the transporters, and diversity in chemical structures.

By determining IC_{50} values for the eight prioritized compounds (Figure 3.3, Supplementary Table S2), five strong OATP2B1 (IC_{50} values are ranging $2.37 \mu\text{M}$ - 20 nM) inhibitors, as well a single strong OATP1B1 ($\text{IC}_{50} = 2.69 \mu\text{M}$) inhibitor, and a single strong OATP1B3 ($1.53 \mu\text{M}$) inhibitors. Selected compounds (codes: B4, C7, E3, E5) were studied in further detail to elucidate their binding modes (Figure 3.4).

Distribution of physico-chemical properties across the actives and inactives for a given hepatic OATP transporter reveals that OATP2B1 actives exhibit higher values of FractionCSP3 (median value 0.25) compared to inactives (median value 0.2). For OATP1B1 and OATP1B3, no distinction between actives and inactives in terms of FractionCSP3 has been observed. By investigating the chemical structures of the measured compounds, flexible linkers (which can be responsible for an increase of the FractionCSP3 value) seem to be important for the ligands to get properly accommodated in the binding pocket

at the TMH7/TMH8 interface. Indeed, docking poses of the strong OATP2B inhibitors reveal that the ligands get accommodated in such a way that one end of the ligands gets stabilized at the TMH7/TMH8 interface, while the second end of the ligand gets tilted via a flexible linker to reach the interface between TMH7 and TMH11 in a “L-shaped” fashion (Figure 3.4B). Furthermore, compounds B4 and E3 show an additional pi-pi interaction with HIS579 (see Figure 3.4A). For other compounds C7 and E5, HIS579 acts as a mechanical barrier disabling the ligand to get bound more deeply to the inner cavity. As a consequence, compounds tend to preferentially bind to the interface between TMH7 and TMH8 at the level of the two alanine residues (ALA391 and ALA387 at TMH7), which appears favorable for adopting the optimal shape complementarity with OATP2B1 binding site.

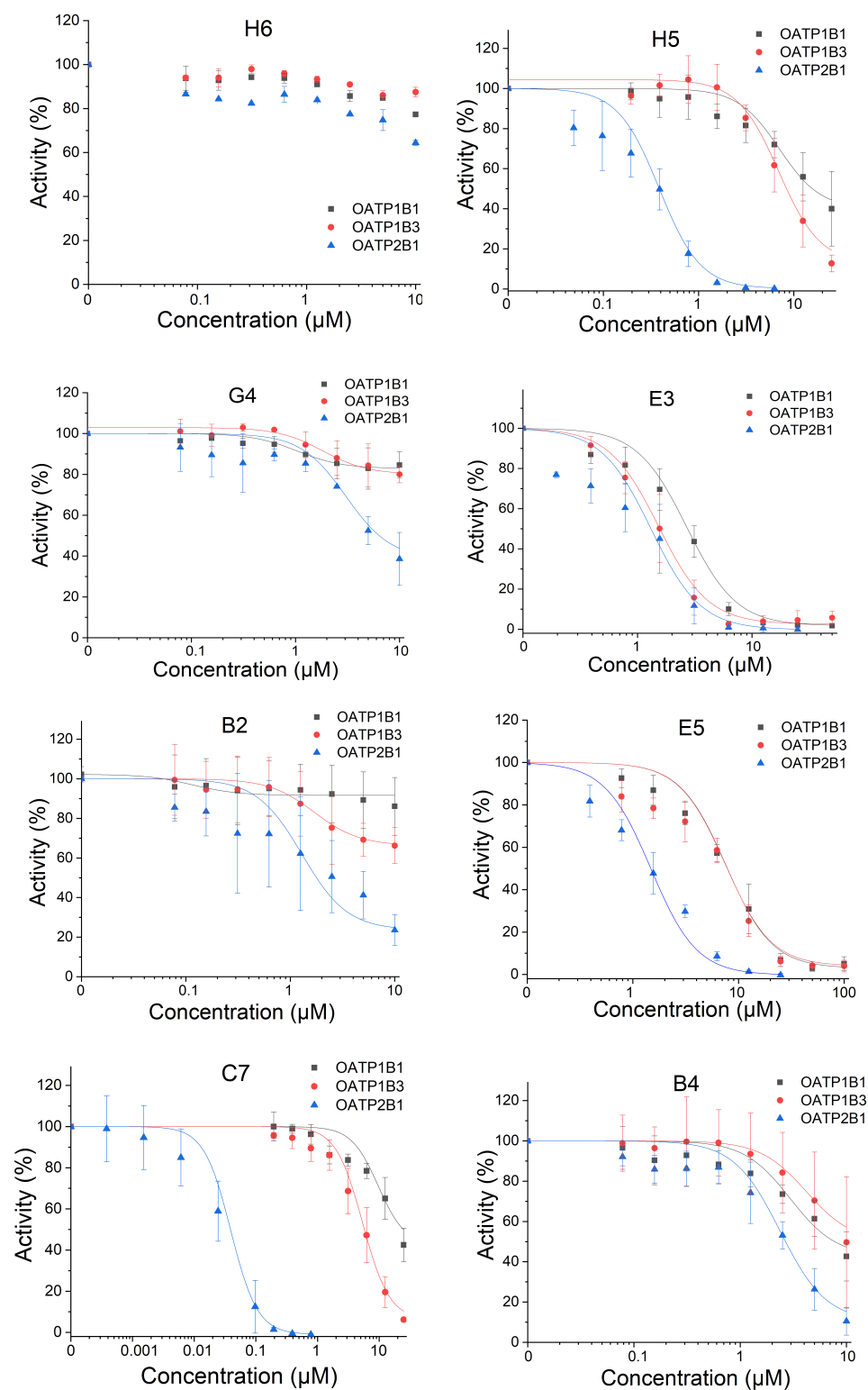


FIGURE 3.3: Graphs showing full dose-response curves for the eight selected inhibitors.

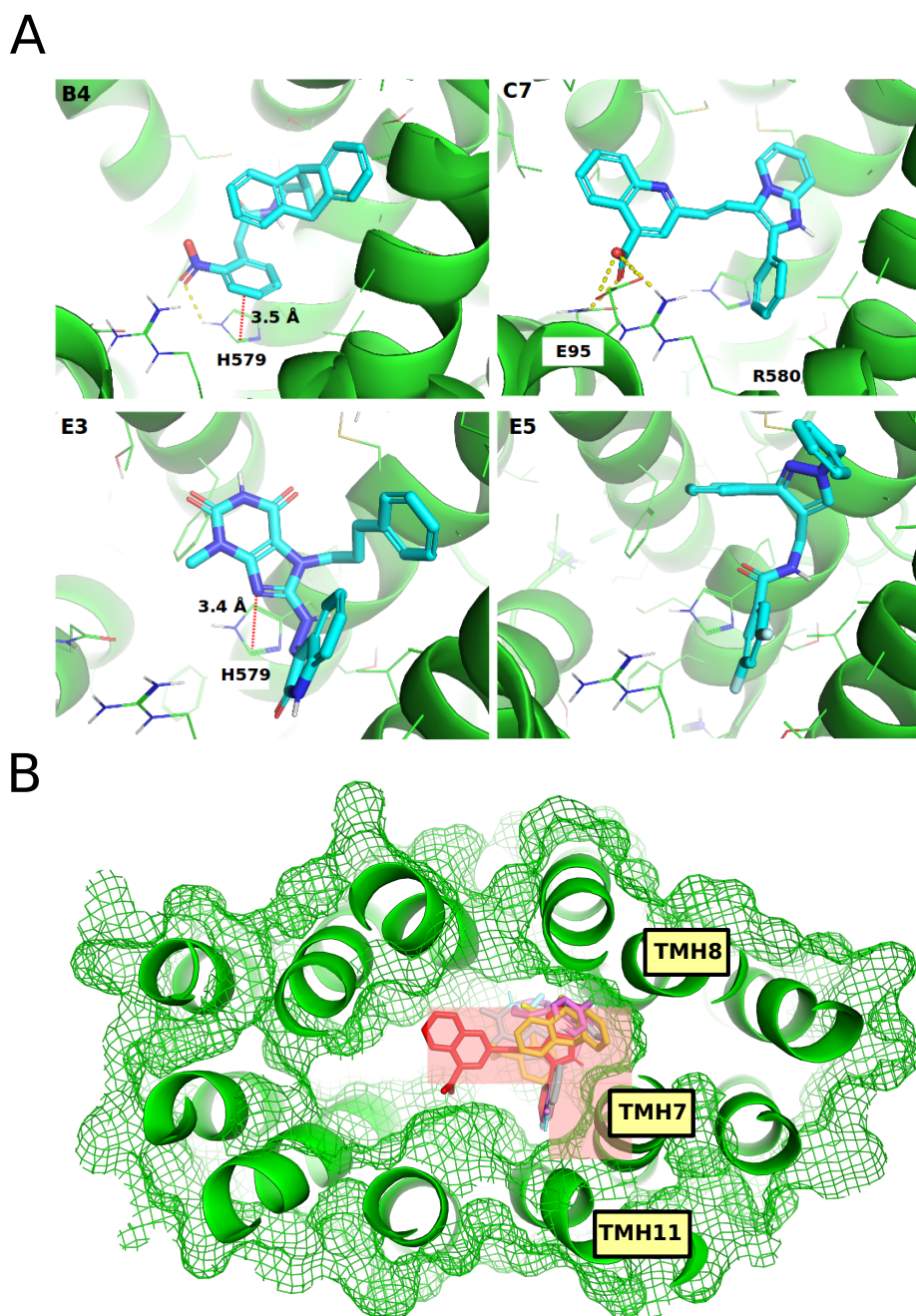


FIGURE 3.4: Interactions of the most potent OATP2B1 inhibitors (codes: B4, C7, E3, E5). (A) In case of B4 and E3 compound pi-pi interaction between the ligand and H579 was observed (as indicated by the red dashed line). In addition, several hydrogen bonds have occurred (as indicated by the yellow dashed line). (B) Representative poses for B4 (the yellow structure), C7 (the red structure), E3 (the cyan structure), E5 (the magenta structure) are showing the ‘L-shaped’ binding mode. The poses shown here represent the most populated poses per compound identified by hierarchical pose clustering.

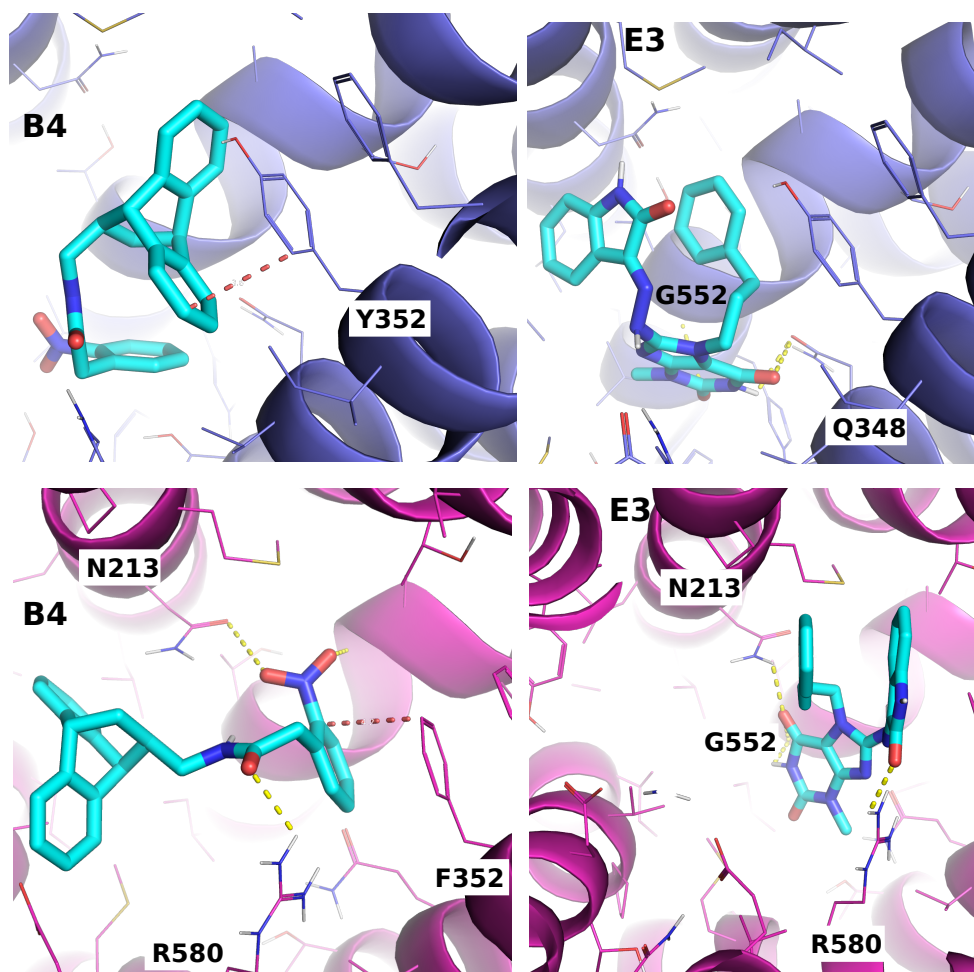


FIGURE 3.5: Prominent interactions in OATP1B1 (the blue structure) and OATP1B3 (the magenta structure). Face-to-face pi-pi interaction between compound B4 and Y352/F352 was observed (as indicated by the red dashed line). In addition, several hydrogen bonds have occurred (as indicated by the yellow dashed line). E3 compound forms a hydrogen bond with GLN348 in OATP1B1 and with ASN213 and ARG580 in OATP1B3, respectively. The poses shown here represent the most populated poses per compound identified upon hierarchical pose clustering.

3.6.4 Conclusions

In silico identification of novel OATP inhibitors confirmed by experimental validation is a promising approach which can be exploited to guide the design of novel chemical probes. Such compounds can be used as tools to study the physiological role of these emerging transporters. In this study, the diverse REAL drug-like set was initially screened by different sorts of machine learning models. By consensus ranking of the identified hits from the ligand-based screening, 3,291 compounds could be identified that were further

docked into the OATP1B1, OATP1B3, and OATP2B1 structural models to prioritize 44 compounds for subsequent transporter inhibition assay experiments. By this procedure, 31 new active compounds (activity threshold $\leq 10 \mu\text{M}$) with either selective, dual, or pan inhibitory activity have been identified. Interestingly, the strongest OATP2B1 inhibitor (compound C7, $\text{IC}_{50}=40 \text{ nM}$), is almost 14-times more potent compared to the strongest OATP2B1 inhibitor reported in literature (erlotinib, $\text{IC}_{50}=550 \text{ nM}$). These findings indicate that the developed integrative modeling pipeline, combining AI-based and structure-based methods, is capable of detecting highly active compounds. Further, it also approves our ensemble-based docking methodology from Study 5, as the structural models delivered by enrichment docking led to the identification of a dataset with a high hit rate (70,5% hit rate).

The newly measured compounds were inspected and the binding mode hypothesis was established. A remarkable difference between the OATP1B1/OATP1B3 and OATP2B1 hit rate was explained by different constitution of the binding site of the hepatic OATPs, being characterized by different localization of aromatic residues in the inner cavity which extends into the C-terminal domain. We have shown that the “L-shaped” inhibitors better fit into the OATP2B1 binding site with respect to their shape complementarity.

Overall, by combining the results from molecular docking of steroidal compounds (Study 5) and the identification of novel inhibitors upon a rigorous virtual screening approach followed by experimental testing, we elucidated the structural basis of the two major binding sites in hepatic OATPs (the N-terminal and inner/C-terminal binding site).

Part IV

CONCLUDING DISCUSSION

In the presented thesis a combination of structure-based modeling and data science (cheminformatics) approaches are being used to study ligand recognition for the selected hepatocellular transporters of pharmacological importance: OCT1, OATP1B1, OATP1B3, and OATP2B1.

In Study 1 we performed (semi)automated fusion of OATP bioactivity data from public data sources (ChEMBL, Metrabase, DrugBank, IUPHAR/Guide-To-Pharmacology, and UCSF-FDA TransPortal). Querying heterogeneous data sources helped to enrich the OATP datasets in terms of both the number of enumerated compounds and the coverage of chemical space. The former aspect was beneficial to make a more rigorous decision about bioactivity cut-off setting for binary label assignment. The latter aspect (increase of the chemical space) was demonstrated, e.g., for Metrabase data, as it became a rich source of microcystin and porphyrin derivatives which were not represented elsewhere. With the aim to separate substrates (here: KC_m and EC_{50} end-points) and inhibitors (here: IC_{50} , K_i , percentage inhibition end-points) Study 1 provides further insights into the constitution of different databases; while Metrabase became a major source of substrate data (69% of all substrates/non-substrates), inhibition data were predominantly collected from ChEMBL (94% of all inhibitors/non-inhibitors). These findings could be of value for a broader scientific community in the field of in silico modeling. To date, modelers predominantly use ChEMBL as a major source of data, as reflected by the number of citations in Google Scholar (118 citations of the ChEMBL update paper from 2019). [54] On the contrary, compound-target pairs from Metrabase have been exploited to a lesser extent (29 citations). [141] The reason for the favored usage of ChEMBL databases could be the possibility to query ChEMBL via an Application Programming Interface (API). In contrast, Metabase provides data access via MySQL Database Service which is less exploited within the cheminformatics community. Alternatively, Metrabase can be accessed by parsing the HTML file which can be done via KNIME, as shown in Study 1.

Substructure analysis yielded enriched scaffold series with pronounced selectivity profiles (selective, dual-, or pan- OATP activity). Here, we did not perform substructure searches based on Murcko scaffolds due to the high heterogeneity of the collected dataset (as reflected by the scaffold-to-compound ratio). Instead, a comparative analysis of the dataset was conducted by using the maximum common substructure as a similarity

metric. Scaffold significance was calculated by the Fisher exact test ($p < 0.05$). Enriched substructures possessing a specific activity profile were additionally grouped on the basis of hierarchical scaffold clustering. Structural analysis delivered substructures which were OATP1B1-specific (pravastatin, estrone, porphyrin, gedunin, khivorin, N-phenylpyrimidin-4-amine, and the valsartan-like derivatives). A substantial fraction of enriched substructures has shown OATP1B1/OATP1B3 dual activity (e.g., cyclosporine derivatives). The tendency towards a high degree of shared substructures could be explained by the high sequence identity between OATP1B1 and OATP1B3 (80% sequence identity). When using a weaker significance level ($p < 0.1$), the steroid scaffold was identified as a common substructure for OATP1B1, OATP1B3, and OATP2B1. R-group scaffold decomposition enabled us to calculate the distribution of distinct substituents across all three hepatic OATPs. R-17 and R-3 positions have shown the highest variety of functional groups for all three transporters. In addition to the scaffold analysis, important molecular features (as revealed by descriptor importance from binary classification) were inspected across the three OATPs. Using two levels of binary classification models - (1) A general OATP model and (2) three individual models for OATP1B1, OATP1B3, and OATP2B1 - we have identified features which contribute to the general OATP inhibition and features conferring OATP1B/OATP2B specificity. Feature analysis helped identify a higher lipophilicity (SlogP, 3.9 for OATP1B1, 4.2 for OATP1B3, 3.64 for OATP2B1 inhibitors, respectively), molecular weight (471.6 for OATP1B1, 504.6 for OATP1B3, 503.3 for OATP2B1 inhibitors, respectively), and increased polarity (topological polar surface area - TPSA, 126.3 for OATP1B1, 133.7 for OATP1B3, and 120.6 for OATP2B1 inhibitors, respectively), as general features driving OATP inhibition. The conclusions drawn here are in agreement with previously reported studies. [27] In addition, molecular refractivity (141.4 for OATP1B1, 132 for OATP1B3, and 127.4 for OATP2B1 inhibitors, respectively), number of rotatable bonds (5 for OATP1B1, 6 for OATP1B3, and 7 for OATP2B1 inhibitors, respectively) and number of rings (4 for OATP1B1, OATP1B3, and OATP2B1 inhibitors, respectively) were identified as novel features which contributed to general OATP inhibition. The fraction of sp^3 -hybridized carbons and number of aromatic carbocycles - have shown divergent importance for OATP1B and OATP2B subfamilies. We conclude that OATP2B1 inhibitors tend to be more planar than OATP1B1/OATP1B3 inhibitors. The information gained on the systematic data analysis provided us with the basic knowledge on OATP ligands for the follow-up structure-based modeling studies (Study 5 and Study 6). The knowledge

about the variability of distinct functional groups on the steroidal scaffold was further elaborated in Study 2.

Study 2 was centered around 13-epiestrones with different modifications (here: at the R-2/R-4 positions halogenation, phenylalkylation). In addition, two different variations of R-3 position (3-hydroxy or 3-methoxy group) were considered. Study 2 was tailored to OATP2B1 due to the fact that this transporter has been found to be involved in hormone dependent cancers. Inhibition of OATP2B1 function, alongside with other proteins implicated in the cancer (such as 17 β -hydroxysteroid-dehydrogenase type 1 or steroid sulfatase), thus represents a promising strategy to prevent tumor progression. [142] In addition, the OATP2B1 transporter is largely understudied, as demonstrated in Study 1 by the number of available inhibition data - 1340 (non)inhibitors for OATP1B1, 1250 (non)inhibitors for OATP1B3, versus 230 (non)inhibitors for OATP2B1. The feature analysis of 13-epiestrones provided us with further insights into the OATP-steroids interactions. In Study 2, we have performed SAR analysis for a series of steroidal compounds showing activity towards OATP2B1. In this context, SAR analysis applied in Study 2 is somewhat similar to matched molecular pair analysis. [143] However, we did not compare compound pairs, but extrapolated from the whole group of compounds with identical scaffolds to derive general trends. Our aim was to identify important physico-chemical features of specific substitution sites which might trigger gain (positive correlation) or loss (negative correlation) of OATP2B1 bioactivity. In general, halogenated substituents at position R-2 (reflected by HallKier alpha descriptor) possessed positive correlations with EC₅₀ measurements. The effect of halogenated substituents in the context of hepatic OATPs has not been thoroughly explored yet. A substructural fragment analysis performed by Shaikh et al is the only evidence found in literature which points to the importance of fluor and 4-fluorophenyl groups in inducing OATP2B1 activity. [144] In contrast, the presence of bulky substituents at the R-2 position has been negatively correlated with OATP2B1 bioactivity. The SAR analysis of the substituents at position R-4 did not prioritize any pronounced trends. We conclude Study 2 with the likelihood of halogen-bond formation in OATP2B1-ligand binding complexes.

From a cheminformatics perspective, Study 1 and Study 2 have demonstrated the advantage of using the KNIME Analytics Platform for data integration and analysis. These findings prompted us to apply KNIME workflows for another cheminformatic approach - a ligand-based drug repurposing pipeline (Study 3). The workflow developed herein

involves a targeted download of protein and bioassay data through web services, data curation and standardization, detection of enriched structural patterns, and retrospective analysis of known drugs by performing substructure mining. Here, workflow flexibility was demonstrated on the basis of two case studies: (1) A novel disease (COVID-19) and (2) a rare disease involving GLUT-1 transporter (GLUT-1 deficiency syndrome). Identified compounds for COVID-19 can be categorized into five separate clusters - (1) compounds with open-chain structures, (2) nucleoside/nucleotide analogs, (3) cyclopropane derivatives, (4) adamantane derivatives, and (5) miscellaneous cluster with ubiquitous structures. For GLUT-1 majority of the detected hits contained a quinoline or quinazoline scaffold. As evidenced in literature, quinoline/quinazoline analogs were reported to be important for their anticonvulsant activity. [145–147] Since the patients with GLUT-1 deficiency syndrome possess seizures, quinoline/quinazoline derivatives could become promising therapeutic candidates for further investigations. In conclusion, Study 3 provides a framework for performing *in silico* drug repurposing, which can be fully reused for other disease of interest.

In Study 4 we focused on OCT1 as a relevant hepatic uptake transporter of immense clinical importance. A comparative analysis of human and mouse OCT1 was carried out to test the possibility to exploit mouse OCT1 as rodent animal model for prediction of OCT1-associated pharmacokinetics and drug efficacy in humans. Human/mouse chimeric OCTs have shown an impact of TMH2 and TMH3 on the uptake characteristics of metformin and thiamine. Here, we supplemented the experimental findings by structural modeling of human and mouse OCT1 structure. We have identified a role of ILE35 at TMH1 (conserved residue) and LEU155 in human/VAL156 in mouse (TMH2) in the formation of hydrophobic packing interactions. Indeed, the effect of VAL156/LEU155 replacement has led to a decrease of mOCT1 transport activity. Here, we relate the decrease of compound affinity to the tendency to form hydrophobic contacts and thus suppress the conformational change of a transporter. Hydrophobic interactions observed here are analogous to the coiled-coil or leucine zipper structural motifs. Interestingly, similar packing interactions were experimentally confirmed for other SLC transporters with MFS fold, such as lactose permease. [148]. The conclusions drawn here are revealing an important role of tertiary interactions in SLC transporters which might impact substrate transport.

In Study 5 we unify the knowledge gained on steroidal compounds from Study 1 and

Study 2. We further expand on the investigations by including structure-based modeling. In a first instance, an ensemble of OATP structural models was generated. Inhibition data collected in Study 1 were used as a docking library to prioritize models on the basis of ligand enrichment. Steroid analogs identified by the substructure analysis in Study 1 and newly measured 13-epiestrones were docked into the final OATP models to establish a binding mode hypothesis. Interestingly, several general consistent trends were identified by the docking calculations. The R-17, R-2, and R-3 substituents of steroidal compounds have predominantly contributed to the protein-ligand interactions. The R-17 and R-3 positions have already been investigated in Study 1 presenting the broadest spectrum of different substitutions. The additional significance of position R-2 in Study 2 might be attributed to the presence of new bioactivity data for 13-epiestrones. Calculated binding modes are describing distinct interaction sites within transmembrane regions of OATPs. Orientational versatility of the steroidal compounds has been observed for the prominent poses. These findings were already reported in different docking studies involving steroidal compounds. [149] Here, we conclude that the steroidal compounds can get flipped within a single binding site as long as the key interactions are at least partially preserved. These observations might be attributed to the fact that the key protein-ligand interactions are primarily mediated via substituents at either end (R-2/R-3 or R-17) of the steroidal scaffold. In a broader context, orientational versatility of steroids reflects the bolaamphiphilic nature of this class of compounds. [150, 151]. As bolaamphiphilic compounds possess polar groups at both ends of a bulky hydrophobic core, flexibility of steroids to adopt two possible orientations becomes more intuitive. When performing molecular docking of steroid analogs, the entire transmembrane region has been defined as a putative binding site. Here, we followed the hypothesis that compounds possessing common scaffold (here: the steroidal core) tend to adopt a similar binding mode. [152] Cluster analysis uncovered two prevalent binding regions for OATP1B1 and OATP1B3 (located in the inner cavity and the N-terminal domain, respectively) and a single binding site in OATP2B1 which is located in the N-terminal transporter domain. We have to note, however, that the cluster analysis for OATP2B1 might be influenced by the limited diversity of the OATP2B1 steroids dataset. Newly measured bioactivities for 13-epiestrones helped identify differences in OATP1B1/OATP1B3 versus OATP2B1 ligand recognition. As a special use case, halogen bond formation in the upper part of the N-terminal domain of OATP2B1 was detected. As opposed to the orientational versatility of the steroidal analogs, the R-2 halogenated 13-epiestrones appeared to adopt

directional-dependent halogen bond with THR133 compared to their R-4 counterparts. Regiospecificity of the halogenated 13-epiestrones is also indicated by the drop in the bioactivity of R-4 halogenated substituents compared to the ones with the halogen atom at the R-2 position. These findings are in line with the SAR analysis done in Study 2. The corresponding poses showing halogen bond formation could not have been detected in OATP1B1/OATP1B3 due to the increased positive electrostatic potential of the N-terminal domain, as well as replacement of THR133 in OATP2B1 to ALA112 in OATP1B1 and SER112 in OATP1B3, respectively. These findings are in agreement with measured bioactivities, indicating a higher activity of halogenated epiestrones against OATP2B1 compared to OATP1B1/OATP1B3.

Most remarkably, a single residue at TMH1 - ALA45 in OATP1B1, GLY45 in OATP1B3, and SER66 in OATP2B1 - was suggested as a key structural determinant acting as a selectivity switch across the three hepatic OATPs. This residues affects ligand accessibility and recognition by modulating bulk of the N-terminal domain and by the ability or disability to form hydrogen bonds with 13-epiestrones. Here, our computational analyses supported by the newly measured 13-epiestrones and available mutagenesis data suggest that the N-terminal domain of hepatic OATPs acts as a selectivity filter for steroidal compounds.

Enrichment docking into a representative subset of protein conformations led to the prioritization of the structural models capable of accomodating highly active compounds (activity threshold < 1 μ M). The active compounds used for ligand enrichment were spread over different areas of chemical space. The top prioritized structural models were therefore not biased towards recognizing compounds with specific molecular/structural properties. As a possible extension of our ensemble docking strategy, ligand enrichment calculations could be repeated using the compounds with a limited structural diversity (e.g., compounds possessing a common scaffold). Structural models generated upon such a restricted dataset could be further exploited to, e.g., explore conformation-specific ligands.

In Study 6, we further examine the predictive power of the generated structural OATP models by performing virtual screening of ENAMINE Real database to identify novel OATP inhibitors. The compounds pre-selected by different sorts of machine learning models (proteochemometric models, QSAR models, and conformational prediction models)

were docked into the inner cavity of OATP structural models. After compound prioritization, 44 compounds were tested in the transport assay. Initial screens identified 32% OATP1B1, 32% OATP1B3, and 70.5% OATP2B1 potent inhibitors (activity threshold $\leq 10 \mu\text{M}$). By subsequent full-dose response measurements IC_{50} values for eight prioritized compounds were determined. Quite remarkably, several strong OATP2B1 inhibitors were identified (IC_{50} values ranging from $2.5 \mu\text{M}$ to 40 nM). Binding of newly found inhibitors is affected by different localization of aromatic residues in OATP1B1/OATP1B3 (TYR352/PHE352 in OATP1B1/OATP1B3 versus ALA387 at corresponding position in OATP2B1, and PHE356 in OATP1B1/OATP1B3 versus ALA391 at the corresponding position in OATP2B1) and OATP2B1 (HIS579 in OATP2B1 versus GLY552 at the corresponding position in OATP1B1/OATP1B3, PHE231 in OATP2B1 versus ASN213 at the corresponding position in OATP1B1/OATP1B3, PHE583 in OATP2B1 versus VAL556 at the corresponding position in OATP1B1 and ILE556 at the corresponding position in OATP1B3, and PHE603 in OATP2B1 versus SER576 at the corresponding position in OATP1B1/OATP1B3). HIS579 has already been tested in mutagenesis studies and was found to be crucial for OATP2B1 activity. [34] Here, we show that replacement of HIS579 in OATP2B1 to GLY552 in OATP1B1/OATP1B3 impacts pocket shape and accessibility, which, together with other replacements of aromatic residues, causes differences in OATP1B1/OATP1B3 versus OATP2B1 binding. Different geometry of the inner cavity in the three hepatic transporters has, in turn, led to the identification of the novel inhibitors with a predominant activity against OATP2B1. The novel OATP inhibitors identified here can be used for the design of chemical tools used to study physiological function of hepatic OATPs. Study 6 provided further insights into the selectivity switches which are located in the inner cavity of the three hepatic OATPs. Overall, Study 5 and 6 provided a comprehensive picture of the two major binding sites in hepatic OATPs.

A consistent trend has been observed for the TMH1/TMH2 interface in OATP1B1 (Figure 3.6A), OATP1B3 (Figure 3.6B), OATP2B1 (Figure 3.6C), human OCT1 (Figure 3.6D), mouse OCT1 (Figure 3.6E), and normal modes of motion for available MFS transporters (Figure 3.6F, directions of the fluctuations are indicated by the white arrows). Specifically, a functional importance of TMH1 and TMH2 has been identified in all the structure-based modeling studies done within the scope of this thesis. TMH1 and TMH2 in hepatic OATPs carry residues, adopting key interactions with steroidal

compounds, such as ALA45/GLY45/SER66 (TMH1), or ASP70/ or GLU74 (TMH1). As indicated in Study 5, some of these residues might act as selectivity switches across the three transporters. The role of the residues at TMH1 and TMH2 has also been extensively explored by alanine scanning studies. [153, 154] Furthermore, ALA45GLY mutations in OATP1B1 confer OATP1B3 transport specificity, as confirmed by mutagenesis studies. [155] The TMH1/TMH2 interface in OATPs is therefore suggested to be a selectivity filter important for ligand recognition. The conclusions drawn from Study 4 points to the effect of hydrophobic packing interactions in between ILE35 at TMH1 and LEU155 at TMH2 (human OCT1, Figure 3.6D), and ILE35 at TMH1 and VAL156 at TMH2 (mouse OCT1, Figure 3.6E). Here, the role of hydrophobic contacts between TMH1 and TMH2 was suggested to play an additional role in substrate transport. NMA performed for all available MFS structures helped identify shared fluctuations across the whole protein family. Among others, the TMH1/TMH2 interface exhibits a considerable flexibility. Figure 3.6F shows the lowest frequency mode calculated for FucP transporter used as an example of MFS proteins herein. This evolutionary conserved motion across the whole MFS proteins might represent functional importance, e.g., in adopting the correct geometry of binding cavity to accommodate ligands. Residues at the TMH1/TMH2 interface were shown to drive ligand recognition also for other representatives of MFS proteins, such as peptide transporters (e.g., for rabbit PepT1, PepTS [156], or PepTSo [157]). Overall, these findings are suggesting the crucial role of TMH1/TMH2 in proteins with MFS fold which forms a basis for the comparative analyses of this class of transporters in the future.

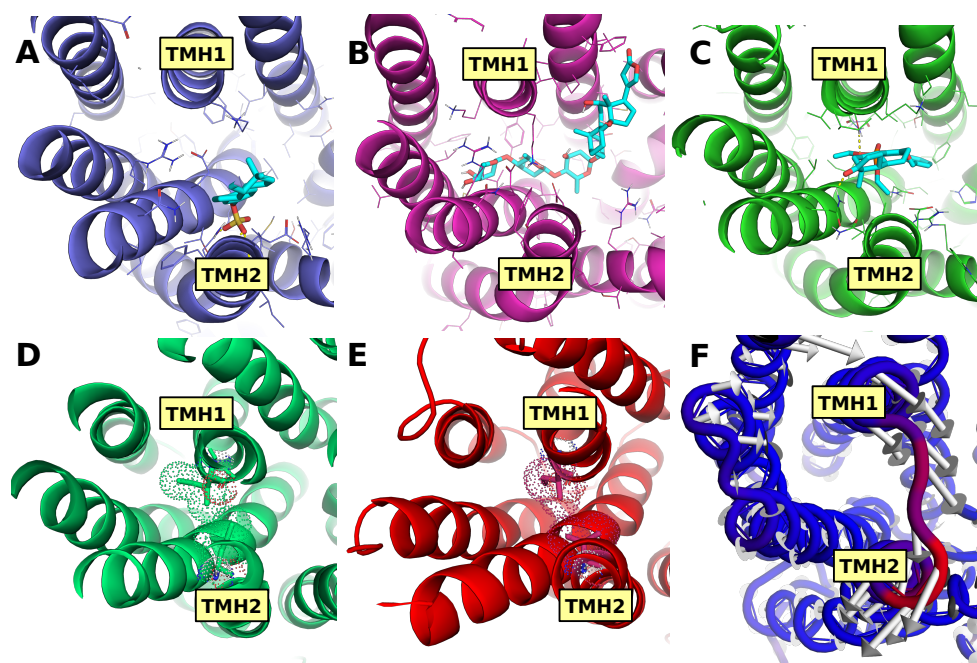


FIGURE 3.6: Different structural aspects of TMH1 and TMH2 helices. (A) OATP1B1, (B) OATP1B3, (C) OATP2B1, (D) human OCT1, (E) mouse OCT1, (F) the lowest frequency mode of motion calculated for FucP transporter.

Overall, these findings show that structure-based modeling, in conjunction with a thorough data integration and analysis, represents a promising strategy which can outperform classic, non-integrative, modeling approaches. In a biological context, the presented thesis ultimately contributed to the elucidation of molecular determinants of hepatic uptake transporters with a special focus on hepatic OATP-ligand interactions and selectivity.

Bibliography

- [1] Hans Popper, Fenton Schaffner, et al. Liver: structure and function. *Liver: structure and function.*, 1957.
- [2] H Remmer. The role of the liver in drug metabolism. *The American journal of medicine*, 49(5):617–629, 1970.
- [3] Alexander Jetter and Gerd A Kullak-Ublick. Drugs and hepatic transporters: A review. *Pharmacological research*, 154:104234, 2020.
- [4] Houfu Liu and Jasminder Sahi. Role of hepatic drug transporters in drug development. *The Journal of Clinical Pharmacology*, 56:S11–S22, 2016.
- [5] Dietrich Keppler. The roles of mrp2, mrp3, oatp1b1, and oatp1b3 in conjugated hyperbilirubinemia. *Drug Metabolism and Disposition*, 42(4):561–565, 2014.
- [6] Christiane Pauli-Magnus and Peter J Meier. Hepatobiliary transporters and drug-induced cholestasis. *Hepatology*, 44(4):778–787, 2006.
- [7] Kathleen M Giacomini, Shiew-Mei Huang, Donald J Tweedie, Leslie Z Benet, Kim LR Brouwer, Xiaoyan Chu, Amber Dahlin, Raymond Evers, Volker Fischer, Kathleen M Hillgren, et al. Membrane transporters in drug development. *Nature reviews Drug discovery*, 9(3):215, 2010.
- [8] Matthias A Hediger, Michael F Romero, Ji-Bin Peng, Andreas Rolfs, Hitomi Takanaga, and Elspeth A Bruford. The abcs of solute carriers: physiological, pathological and therapeutic implications of human membrane transport proteins. *Pflügers Archiv*, 447(5):465–468, 2004.
- [9] Eva Meixner, Ulrich Goldmann, Vitaly Sedlyarov, Stefania Scorzoni, Manuele Reb-samen, Enrico Girardi, and Giulio Superti-Furga. A substrate-based ontology for human solute carriers. *Molecular systems biology*, 16(7):e9652, 2020.
- [10] Lena Schaller and Volker M Lauschke. The genetic landscape of the human solute carrier (slc) transporter superfamily. *Human genetics*, 138(11-12):1359–1377, 2019.

- [11] Avner Schlessinger, Sook Wah Yee, Andrej Sali, and Kathleen M Giacomini. Slc classification: an update. *Clinical Pharmacology & Therapeutics*, 94(1):19–23, 2013.
- [12] Mike Mueckler and Bernard Thorens. The slc2 (glut) family of membrane transporters. *Molecular aspects of medicine*, 34(2-3):121–138, 2013.
- [13] Andreas Stahl. A current review of fatty acid transport proteins (slc27). *Pflügers Archiv*, 447(5):722–727, 2004.
- [14] Hermann Koepsell. The slc22 family with transporters of organic cations, anions and zwitterions. *Molecular aspects of medicine*, 34(2-3):413–435, 2013.
- [15] Bruno Hagenbuch and Bruno Stieger. The slco (former slc21) superfamily of transporters. *Molecular aspects of medicine*, 34(2-3):396–412, 2013.
- [16] Tatiana Claro Da Silva, James E Polli, and Peter W Swaan. The solute carrier family 10 (slc10): beyond bile acid transport. *Molecular aspects of medicine*, 34(2-3):252–269, 2013.
- [17] Helgi B Schiöth, Sahar Roshanbin, Maria GA Hägglund, and Robert Fredriksson. Evolutionary origin of amino acid transporter families slc32, slc36 and slc38 and physiological, pathological and therapeutic aspects. *Molecular aspects of medicine*, 34(2-3):571–585, 2013.
- [18] Dimitrios Fotiadis, Yoshikatsu Kanai, and Manuel Palacín. The slc3 and slc7 families of amino acid transporters. *Molecular aspects of medicine*, 34(2-3):139–158, 2013.
- [19] Lawrence Lin, Sook Wah Yee, Richard B Kim, and Kathleen M Giacomini. Slc transporters as therapeutic targets: emerging opportunities. *Nature reviews Drug discovery*, 14(8):543–560, 2015.
- [20] Anders S Kristensen, Jacob Andersen, Trine N Jørgensen, Lena Sørensen, Jacob Eriksen, Claus J Loland, Kristian Strømgaard, and Ulrik Gether. Slc6 neurotransmitter transporters: structure, function, and regulation. *Pharmacological reviews*, 63(3):585–640, 2011.
- [21] Hisao Imai, Kyoichi Kaira, Noboru Oriuchi, Kimihiro Shimizu, Hideyuki Tomimaga, Noriko Yanagitani, Noriaki Sunaga, Tamotsu Ishizuka, Shushi Nagamori, Kanyarat Promchan, et al. Inhibition of l-type amino acid transporter 1 has antitumor activity in non-small cell lung cancer. *Anticancer research*, 30(12):4819–4828, 2010.

- [22] Mohamed Hassanein, Megan D Hoeksema, Masakazu Shiota, Jun Qian, Bradford K Harris, Heidi Chen, Jonathan E Clark, William E Alborn, Rosana Eisenberg, and Pierre P Massion. Slc1a5 mediates glutamine transport required for lung cancer cell growth and survival. *Clinical cancer research*, 19(3):560–570, 2013.
- [23] Mathew G Soars, Peter JH Webborn, and Robert J Riley. Impact of hepatic uptake transporters on pharmacokinetics and drug- drug interactions: use of assays and models for decision making in the pharmaceutical industry. *Molecular pharmaceutics*, 6(6):1662–1677, 2009.
- [24] Bruno Hagenbuch. Drug uptake systems in liver and kidney: a historic perspective. *Clinical Pharmacology & Therapeutics*, 87(1):39–47, 2010.
- [25] Gerd A Kullak-Ublick, Bruno Stieger, and Peter J Meier. Enterohepatic bile salt transporters in normal physiology and liver disease. *Gastroenterology*, 126(1):322–342, 2004.
- [26] Evita van de Steeg, Viktor Stránecký, Hana Hartmannová, Lenka Nosková, Martin Hřebíček, Els Wagenaar, Anita van Esch, Dirk R de Waart, Ronald PJ Oude Elferink, Kathryn E Kenworthy, et al. Complete oatp1b1 and oatp1b3 deficiency causes human rotor syndrome by interrupting conjugated bilirubin reuptake into the liver. *The Journal of clinical investigation*, 122(2):519–528, 2012.
- [27] Maria Karlgren, Anna Vildhede, Ulf Norinder, Jacek R Wisniewski, Emi Kimoto, Yurong Lai, Ulf Haglund, and Per Artursson. Classification of inhibitors of hepatic organic anion transporting polypeptides (oatps): influence of protein expression on drug–drug interactions. *Journal of medicinal chemistry*, 55(10):4740–4763, 2012.
- [28] Manfred G Ismail, Bruno Stieger, Valentino Cattori, Bruno Hagenbuch, Michael Fried, Peter J Meier, and Gerd A Kullak-Ublick. Hepatic uptake of cholecystokinin octapeptide by organic anion-transporting polypeptides oatp4 and oatp8 of rat and human liver. *Gastroenterology*, 121(5):1185–1190, 2001.
- [29] Pijun Wang, Soichiro Hata, Yansen Xiao, John W Murray, and Allan W Wolkoff. Topological assessment of oatp1a1: a 12-transmembrane domain integral membrane protein with three n-linked carbohydrate chains. *American Journal of Physiology-Gastrointestinal and Liver Physiology*, 294(4):G1052–G1059, 2008.
- [30] Vikas Taank, Wenshuo Zhou, Xuran Zhuang, John F Anderson, Utpal Pal, Hameeda Sultana, and Girish Neelakanta. Characterization of tick organic anion transporting polypeptides (oatps) upon bacterial and viral infections. *Parasites & vectors*, 11(1):1–12, 2018.

- [31] B Hagenbuch and C Gui. Xenobiotic transporters of the human organic anion transporting polypeptides (oatp) family. *Xenobiotica*, 38(7-8):778–801, 2008.
- [32] Khondoker Alam, Alexandra Crowe, Xueying Wang, Pengyue Zhang, Kai Ding, Lang Li, and Wei Yue. Regulation of organic anion transporting polypeptides (oatp) 1b1-and oatp1b3-mediated transport: an updated review in the context of oatp-mediated drug-drug interactions. *International journal of molecular sciences*, 19(3):855, 2018.
- [33] Biochim Hagenbuch and Peter J Meier. The superfamily of organic anion transporting polypeptides. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1609(1):1–18, 2003.
- [34] Yusuke Hoshino, Daichi Fujita, Takeo Nakanishi, and Ikumi Tamai. Molecular localization and characterization of multiple binding sites of organic anion transporting polypeptide 2b1 (oatp2b1) as the mechanism for substrate and modulator dependent drug–drug interaction. *MedChemComm*, 7(9):1775–1782, 2016.
- [35] Fabienne Meier-Abt, Younes Mokrab, and Kenji Mizuguchi. Organic anion transporting polypeptides of the oatp/slco superfamily: identification of new members in nonmammalian species, comparative modeling and a potential transport mode. *The Journal of membrane biology*, 208(3):213–227, 2006.
- [36] Jeff Abramson, Irina Smirnova, Vladimir Kasho, Gillian Verner, H Ronald Kaback, and So Iwata. Structure and mechanism of the lactose permease of escherichia coli. *Science*, 301(5633):610–615, 2003.
- [37] Yafei Huang, M Joanne Lemieux, Jinmei Song, Manfred Auer, and Da-Neng Wang. Structure and mechanism of the glycerol-3-phosphate transporter from escherichia coli. *Science*, 301(5633):616–620, 2003.
- [38] Shan Chang, Kang-shun Li, Jian-ping Hu, Xiong Jiao, and Xu-hong Tian. Allosteric and transport behavior analyses of a fucose transporter with network models. *Soft Matter*, 7(10):4661–4671, 2011.
- [39] Bruno Hagenbuch. Cellular entry of thyroid hormones by organic anion transporting polypeptides. *Best practice & research Clinical endocrinology & metabolism*, 21(2):209–221, 2007.
- [40] Simone Leuthold, Bruno Hagenbuch, Nilufar Mohebbi, Carsten A Wagner, Peter J Meier, and Bruno Stieger. Mechanisms of ph-gradient driven transport mediated by organic anion polypeptide transporters. *American Journal of Physiology-Cell Physiology*, 296(3):C570–C582, 2009.

- [41] Daisuke Kobayashi, Takashi Nozawa, Kozue Imai, Jun-ichi Nezu, Akira Tsuji, and Ikumi Tamai. Involvement of human organic anion transporting polypeptide oatp-b (slc21a9) in ph-dependent transport across intestinal apical membrane. *Journal of pharmacology and experimental therapeutics*, 306(2):703–708, 2003.
- [42] Johan W Jonker and Alfred H Schinkel. Pharmacological and physiological functions of the polyspecific organic cation transporters: Oct1, 2, and 3 (slc22a1-3). *Journal of Pharmacology and Experimental Therapeutics*, 308(1):2–9, 2004.
- [43] Hermann Koepsell. Polyspecific organic cation transporters: their functions and interactions with drugs. *Trends in pharmacological sciences*, 25(7):375–381, 2004.
- [44] Eriko Shikata, Rei Yamamoto, Hiroshi Takane, Chiaki Shigemasa, Tadasu Ikeda, Kenji Otsubo, and Ichiro Ieiri. Human organic cation transporter (oct1 and oct2) gene polymorphisms and therapeutic effects of metformin. *Journal of human genetics*, 52(2):117–122, 2007.
- [45] Hermann Koepsell and Thorsten Keller. Functional properties of organic cation transporter oct1, binding of substrates and inhibitors, and presumed transport mechanism. In *Organic Cation Transporters*, pages 49–72. Springer, 2016.
- [46] Hermann Koepsell. Multiple binding sites in organic cation transporters require sophisticated procedures to identify interactions of novel drugs. *Biological Chemistry*, 400(2):195–207, 2019.
- [47] Frank K Brown et al. Chemoinformatics: what is it and how does it impact drug discovery. *Annual reports in medicinal chemistry*, 33:375–384, 1998.
- [48] Stephan Beisen, Thorsten Meinl, Bernd Wiswedel, Luis F de Figueiredo, Michael Berthold, and Christoph Steinbeck. Knime-cdk: Workflow-driven cheminformatics. *BMC bioinformatics*, 14(1):257, 2013.
- [49] Moises Hassan, Robert D Brown, Shikha Varma-O’Brien, and David Rogers. Cheminformatics analysis and learning in a data pipelining environment. *Molecular diversity*, 10(3):283–299, 2006.
- [50] Michael R Berthold, Nicolas Cebron, Fabian Dill, Thomas R Gabriel, Tobias Kötter, Thorsten Meinl, Peter Ohl, Kilian Thiel, and Bernd Wiswedel. Knime-the konstanz information miner: version 2.0 and beyond. *AcM SIGKDD explorations Newsletter*, 11(1):26–31, 2009.
- [51] Wendy A Warr. Scientific workflow systems: Pipeline pilot and knime. *Journal of computer-aided molecular design*, 26(7):801–804, 2012.

- [52] Timothy NC Wells, Paul Willis, Jeremy N Burrows, and Rob Hooft van Huijsduinen. Open data in drug discovery and development: lessons from malaria. *Nature Reviews Drug Discovery*, 15(10):661, 2016.
- [53] Konstantin V Balakin. *Pharmaceutical data mining: approaches and applications for drug discovery*, volume 6. John Wiley & Sons, 2009.
- [54] David Mendez, Anna Gaulton, A Patrícia Bento, Jon Chambers, Marleen De Veij, Eloy Félix, María Paula Magariños, Juan F Mosquera, Prudence Mutowo, Michal Nowotka, et al. ChEMBL: towards direct deposition of bioassay data. *Nucleic acids research*, 47(D1):D930–D940, 2019.
- [55] David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, et al. Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic acids research*, 46(D1):D1074–D1082, 2018.
- [56] Simon D Harding, Joanna L Sharman, Elena Faccenda, Chris Southan, Adam J Pawson, Sam Ireland, Alasdair JG Gray, Liam Bruce, Stephen PH Alexander, Stephen Anderton, et al. The iuphar/bps guide to pharmacology in 2018: updates and expansion to encompass the new guide to immunopharmacology. *Nucleic acids research*, 46(D1):D1091–D1106, 2018.
- [57] UniProt Consortium. Uniprot: a hub for protein information. *Nucleic acids research*, 43(D1):D204–D212, 2015.
- [58] David S Goodsell, Christine Zardecki, Luigi Di Costanzo, Jose M Duarte, Brian P Hudson, Irina Persikova, Joan Segura, Chenghua Shao, Maria Voigt, John D Westbrook, et al. Rcsb protein data bank: Enabling biomedical research and drug discovery. *Protein Science*, 29(1):52–65, 2020.
- [59] Denise Carvalho-Silva, Andrea Pierleoni, Miguel Pignatelli, ChuangKee Ong, Luca Fumis, Nikiforos Karamanis, Miguel Carmona, Adam Faulconbridge, Andrew Hercules, Elaine McAuley, Alfredo Miranda, Gareth Peat, Michaela Spitzer, Jeffrey Barrett, David G Hulcoop, Eliseo Papa, Gautier Koscielny, and Ian Dunham. Open Targets Platform: new developments and updates two years on. *Nucleic Acids Research*, 47(D1):D1056–D1065, 11 2018.
- [60] Domenico Gadaleta, Anna Lombardo, Cosimo Toma, and Emilio Benfenati. A new semi-automated workflow for chemical data retrieval and quality checking for modeling applications. *Journal of Cheminformatics*, 10(1), dec 2018.

- [61] Mario Lovrić, José Manuel Molero, and Roman Kern. PySpark and RDKit: Moving towards big data in cheminformatics. *Molecular Informatics*, 38(6):1800082, mar 2019.
- [62] The IUPAC international chemical identifier (InChI). *Chemistry International – Newsmagazine for IUPAC*, 31(1), jan 2009.
- [63] David Weininger. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Modeling*, 28(1):31–36, feb 1988.
- [64] Gilles Klopmand. Concepts and applications of molecular similarity, by mark a. johnson and gerald m. maggiora, eds., john wiley & sons, new york, 1990, 393 pp. price: \$65.00. *Journal of Computational Chemistry*, 13(4):539–540, may 1992.
- [65] Yoan Martínez-López, Yovani Marrero-Ponce, Stephen J. Barigye, Enrique Teran, Oscar Martínez-Santiago, Cesar H. Zambrano, and F. Javier Torres. When global and local molecular descriptors are more than the sum of its parts: Simple, but not simpler? *Molecular Diversity*, oct 2019.
- [66] Dávid Bajusz, Anita Rácz, and Károly Héberger. Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics*, 7(1), may 2015.
- [67] Yiqun Cao, Tao Jiang, and Thomas Girke. A maximum common substructure-based algorithm for searching and predicting drug-like compounds. *Bioinformatics*, 24(13):i366–i374, jul 2008.
- [68] Jonas Bostrom and Andrew Grant. Smarts smiles arbitrary target specification smiles simplified molecular input line entry system. *Molecular Drug Properties: Measurement and Prediction*, 37:183, 2008.
- [69] So much more to know . *Science*, 309(5731):78b–102b, jul 2005.
- [70] J. Moult and M. N. G. James. An algorithm for determining the conformation of polypeptide segments in proteins by systematic search. *Proteins: Structure, Function, and Genetics*, 1(2):146–163, feb 1986.
- [71] Robert E. Bruccoleri and Martin Karplus. Prediction of the folding of short polypeptide segments by uniform conformational sampling. *Biopolymers*, 26(1):137–168, jan 1987.

- [72] R. M. Fine, H. Wang, P. S. Shenkin, D. L. Yarmush, and C. Levinthal. Predicting antibody hypervariable loop conformations II: Minimization and molecular dynamics studies of MCP603 from many randomly generated loop conformations. *Proteins: Structure, Function, and Genetics*, 1(4):342–362, apr 1986.
- [73] C. Chothia and A.M. Lesk. The relation between the divergence of sequence and structure in proteins. *The EMBO Journal*, 5(4):823–826, apr 1986.
- [74] Andrej Šali and Tom L. Blundell. Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, 234(3):779–815, dec 1993.
- [75] Burkhard Rost. Twilight zone of protein sequence alignments. *Protein Engineering, Design and Selection*, 12(2):85–94, feb 1999.
- [76] Steven E. Brenner and Ichael Levitt. Expectations from structural genomics. *Protein Science*, 9(1):197–200, dec 2008.
- [77] Anna Lobley, Michael I Sadowski, and David T Jones. pgenthreader and pdomthreader: new methods for improved protein fold recognition and superfamily discrimination. *Bioinformatics*, 25(14):1761–1767, 2009.
- [78] Dong Deng, Pengcheng Sun, Chuangye Yan, Meng Ke, Xin Jiang, Lei Xiong, Wenlin Ren, Kunio Hirata, Masaki Yamamoto, Shilong Fan, et al. Molecular basis of ligand recognition and transport by glucose transporters. *Nature*, 526(7573):391–396, 2015.
- [79] Manfred J Sippl. Recognition of errors in three-dimensional structures of proteins. *Proteins: Structure, Function, and Bioinformatics*, 17(4):355–362, 1993.
- [80] Themis Lazaridis and Martin Karplus. Discrimination of the native from misfolded protein models with an energy function including implicit solvation. *Journal of molecular biology*, 288(3):477–487, 1999.
- [81] Christopher J Williams, Jeffrey J Headd, Nigel W Moriarty, Michael G Prisant, Lizbeth L Videau, Lindsay N Deis, Vishal Verma, Daniel A Keedy, Bradley J Hintze, Vincent B Chen, et al. Molprobity: More and better reference data for improved all-atom structure validation. *Protein Science*, 27(1):293–315, 2018.
- [82] Hao Fan, John J Irwin, Benjamin M Webb, Gerhard Klebe, Brian K Shoichet, and Andrej Sali. Molecular docking screens using comparative models of proteins. *Journal of chemical information and modeling*, 49(11):2512–2527, 2009.
- [83] Michael M Mysinger, Michael Carchia, John J Irwin, and Brian K Shoichet. Directory of useful decoys, enhanced (dud-e): better ligands and decoys for better benchmarking. *Journal of medicinal chemistry*, 55(14):6582–6594, 2012.

- [84] Ajay N Jain and Anthony Nicholls. Recommendations for evaluation of computational methods. *Journal of computer-aided molecular design*, 22(3-4):133–139, 2008.
- [85] John J Irwin, Brian K Shoichet, Michael M Mysinger, Niu Huang, Francesco Colizzi, Pascal Wassam, and Yiqun Cao. Automated docking screens: a feasibility study. *Journal of medicinal chemistry*, 52(18):5712–5720, 2009.
- [86] Thomas Lengauer and Matthias Rarey. Computational methods for biomolecular docking. *Current opinion in structural biology*, 6(3):402–406, 1996.
- [87] A Chang Chia-en, Wei Chen, and Michael K Gilson. Ligand configurational entropy and protein binding. *Proceedings of the National Academy of Sciences*, 104(5):1534–1539, 2007.
- [88] Dima Kozakov, Laurie E Grove, David R Hall, Tanggis Bohnuud, Scott E Mottarella, Lingqi Luo, Bing Xia, Dmitri Beglov, and Sandor Vajda. The ftmap family of web servers for determining and characterizing ligand-binding hot spots of proteins. *Nature protocols*, 10(5):733–755, 2015.
- [89] Oleg Trott and Arthur J Olson. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 31(2):455–461, 2010.
- [90] Renxiao Wang, Luhua Lai, and Shaomeng Wang. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *Journal of computer-aided molecular design*, 16(1):11–26, 2002.
- [91] Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.
- [92] DE Koshland Jr. Application of a theory of enzyme specificity to protein synthesis. *Proceedings of the National Academy of Sciences of the United States of America*, 44(2):98, 1958.
- [93] Daniel E Koshland Jr. The key–lock theory and the induced fit theory. *Angewandte Chemie International Edition in English*, 33(23-24):2375–2378, 1995.
- [94] Jacque Monod, Jeffries Wyman, and Jean-Pierre Changeux. On the nature of allosteric transitions: a plausible model. *J Mol Biol*, 12(1):88–118, 1965.
- [95] Dror Tobi and Ivet Bahar. Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state. *Proceedings of the National Academy of Sciences*, 102(52):18908–18913, 2005.

- [96] Kei-ichi Okazaki and Shoji Takada. Dynamic energy landscape view of coupled binding and protein conformational change: induced-fit versus population-shift mechanisms. *Proceedings of the National Academy of Sciences*, 105(32):11182–11187, 2008.
- [97] Leo C James and Dan S Tawfik. Structure and kinetics of a transient antibody binding intermediate reveal a kinetic discrimination mechanism in antigen recognition. *Proceedings of the National Academy of Sciences*, 102(36):12730–12735, 2005.
- [98] Ahmet Bakan and Ivet Bahar. The intrinsic dynamics of enzymes plays a dominant role in determining the structural changes induced upon inhibitor binding. *Proceedings of the National Academy of Sciences*, 106(34):14349–14354, 2009.
- [99] Maxim Totrov and Ruben Abagyan. Flexible ligand docking to multiple receptor conformations: a practical alternative. *Current opinion in structural biology*, 18(2):178–184, 2008.
- [100] Wilfredo Evangelista Falcon, Sally R Ellingson, Jeremy C Smith, and Jerome Baudry. Ensemble docking in drug discovery: how many protein configurations from molecular dynamics simulations are needed to reproduce known ligand binding? *The Journal of Physical Chemistry B*, 123(25):5189–5195, 2019.
- [101] Sally R Ellingson, Yinglong Miao, Jerome Baudry, and Jeremy C Smith. Multi-conformer ensemble docking to difficult protein targets. *The Journal of Physical Chemistry B*, 119(3):1026–1034, 2015.
- [102] Anhui Wang, Yuebin Zhang, Huiying Chu, Chenyi Liao, Zhichao Zhang, and Guohui Li. Higher accuracy achieved for protein-ligand binding pose prediction by elastic network model based ensemble docking. *Journal of Chemical Information and Modeling*, 2020.
- [103] Jens Carlsson, Ryan G Coleman, Vincent Setola, John J Irwin, Hao Fan, Avner Schlessinger, Andrej Sali, Bryan L Roth, and Brian K Shoichet. Ligand discovery from a dopamine d 3 receptor homology model and crystal structure. *Nature chemical biology*, 7(11):769–778, 2011.
- [104] Luca Ponzoni, She Zhang, Mary Hongying Cheng, and Ivet Bahar. Shared dynamics of LeuT superfamily members and allosteric differentiation by structural irregularities and multimerization. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1749):20170177, may 2018.

- [105] Timothy R Lezon, I Srivastava, Y Zheng, and Ivet Bahar. Elastic network models for biomolecular dynamics: theory and application to membrane proteins and viruses. *Handbook on Biological Networks*, pages 129–58, 2009.
- [106] AJ Rader, Chakra Chennubhotla, Lee-Wei Yang, Ivet Bahar, and Q Cui. The gaussian network model: Theory and applications. *Normal mode analysis: Theory and applications to biological and chemical systems*, 9:41–64, 2006.
- [107] Akio Kitao and Nobuhiro Go. Investigating protein dynamics in collective coordinate space. *Current opinion in structural biology*, 9(2):164–169, 1999.
- [108] Konrad Hinsen. Analysis of domain motions by approximate normal mode calculations. *Proteins: Structure, Function, and Bioinformatics*, 33(3):417–429, 1998.
- [109] Ji Guo Su, Chun Hua Li, Rui Hao, Wei Zu Chen, and Cun Xin Wang. Protein unfolding behavior studied by elastic network model. *Biophysical journal*, 94(12):4586–4596, 2008.
- [110] Pemra Doruker, Ali Rana Atilgan, and Ivet Bahar. Dynamics of proteins predicted by molecular dynamics simulations and analytical approaches: Application to α -amylase inhibitor. *Proteins: Structure, Function, and Bioinformatics*, 40(3):512–524, 2000.
- [111] Monique M Tirion. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Physical review letters*, 77(9):1905, 1996.
- [112] Sanzo Miyazawa and Robert L. Jernigan. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*, 18(3):534–552, may 1985.
- [113] I. Bahar and R.L. Jernigan. Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation. *Journal of Molecular Biology*, 266(1):195–214, feb 1997.
- [114] B. Halle. Flexibility and packing in proteins. *Proceedings of the National Academy of Sciences*, 99(3):1274–1279, jan 2002.
- [115] Sibsankar Kundu, Julia S. Melton, Dan C. Sorensen, and George N. Phillips. Dynamics of proteins in crystals: Comparison of experiment with simple models. *Biophysical Journal*, 83(2):723–732, aug 2002.
- [116] A.R. Atilgan, S.R. Durell, R.L. Jernigan, M.C. Demirel, O. Keskin, and I. Bahar. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophysical Journal*, 80(1):505–515, jan 2001.

- [117] K.P. Wilson, B.A. Malcolm, and B.W. Matthews. STRUCTURAL AND THERMODYNAMIC ANALYSIS OF COMPENSATING MUTATIONS WITHIN THE CORE OF CHICKEN EGG WHITE LYSOZYME, oct 1993.
- [118] Statistical thermodynamics of random networks. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, 351(1666):351–380, nov 1976.
- [119] David W Miller and David A Agard. Enzyme specificity under dynamic control: A normal mode analysis of α -lytic protease. *Journal of Molecular Biology*, 286(1):267–278, feb 1999.
- [120] Mengmeng Wang, Ronald T. Borchardt, Richard L. Schowen, and Krzysztof Kucze-
ra. Domain motions and the open-to-closed conformational transition of an en-
zyme: a normal mode analysis of S-adenosyl-L-homocysteine hydrolase†. *Biochem-
istry*, 44(19):7228–7239, may 2005.
- [121] Ivet Bahar. On the functional significance of soft modes predicted by coarse-
grained models for membrane proteins. *Journal of General Physiology*, 135(6):563–
573, may 2010.
- [122] Timothy R. Lezon, Andrej Sali, and Ivet Bahar. Global motions of the nuclear
pore complex: Insights from elastic network models. *PLoS Computational Biology*,
5(9):e1000496, sep 2009.
- [123] F. Tama, M. Valle, J. Frank, and C. L. Brooks. Dynamic reorganization of the
functionally active ribosome explored by normal mode analysis and cryo-electron
microscopy. *Proceedings of the National Academy of Sciences*, 100(16):9319–9323,
jul 2003.
- [124] Florence Tama and Charles L. Brooks. Diversity and identity of mechanical proper-
ties of icosahedral viral capsids studied with elastic network normal mode analysis.
Journal of Molecular Biology, 345(2):299–314, jan 2005.
- [125] Herman W.T. van Vlijmen and Martin Karplus. Normal mode calculations of
icosahedral viruses with full dihedral flexibility by use of molecular symmetry.
Journal of Molecular Biology, 350(3):528–542, jul 2005.
- [126] Gareth Williams. Elastic network model of allosteric regulation in protein kinase
PDK1. *BMC Structural Biology*, 10(1):11, 2010.
- [127] Lu Jin. Allosteric transitions of glutamine-binding protein studied by the elastic
network model. *American Journal of Bioscience and Bioengineering*, 3(6):162,
2015.

- [128] Tomasz Oliwa and Yang Shen. cNMA: a framework of encounter complex-based normal mode analysis to model conformational changes in protein interactions. *Bioinformatics*, 31(12):i151–i160, jun 2015.
- [129] L. Yang, G. Song, and R. L. Jernigan. Protein elastic network models and the ranges of cooperativity. *Proceedings of the National Academy of Sciences*, 106(30):12347–12352, jul 2009.
- [130] Mert Gur, Jeffrey D. Madura, and Ivet Bahar. Global transitions of proteins explored by a multiscale hybrid methodology: Application to adenylate kinase. *Biophysical Journal*, 105(7):1643–1652, oct 2013.
- [131] Eduardo Habib Bechelane Maia, Letícia Cristina Assis, Tiago Alves de Oliveira, Alisson Marques da Silva, and Alex Gutterres Taranto. Structure-based virtual screening: from classical to artificial intelligence. *Frontiers in Chemistry*, 8, 2020.
- [132] Tom De Bruyn, Gerard JP Van Westen, Adriaan P IJzerman, Bruno Stieger, Peter de Witte, Patrick F Augustijns, and Pieter P Annaert. Structure-based identification of oatp1b1/3 inhibitors. *Molecular pharmacology*, 83(6):1257–1267, 2013.
- [133] Natalia Khuri, Arik A Zur, Matthias B Wittwer, Lawrence Lin, Sook Wah Yee, Andrej Sali, and Kathleen M Giacomini. Computational discovery and experimental validation of inhibitors of the human intestinal transporter oatp2b1. *Journal of Chemical Information and Modeling*, 57(6):1402–1413, 2017.
- [134] AN Shivanyuk, SV Ryabukhin, A Tolmachev, AV Bogolyubsky, DM Mykytenko, AA Chupryna, W Heilman, and AN Kostyuk. Enamine real database: Making chemical diversity real. *Chemistry today*, 25(6):58–59, 2007.
- [135] Alzbeta Turkova, Sankalp Jain, and Barbara Zdrazil. Integrative data mining, scaffold analysis, and sequential binary classification models for exploring ligand profiles of hepatic organic anion transporting polypeptides. *Journal of chemical information and modeling*, 59(5):1811–1825, 2018.
- [136] Ulf Norinder, Lars Carlsson, Scott Boyer, and Martin Eklund. Introducing conformational prediction in predictive modeling. a transparent and flexible alternative to applicability domain determination. *Journal of chemical information and modeling*, 54(6):1596–1603, 2014.
- [137] Brandon J Bongers, Adriaan P IJzerman, and Gerard JP Van Westen. Proteochemometrics—recent developments in bioactivity and selectivity modeling. *Drug Discovery Today: Technologies*, 2020.

- [138] Izabel Patik, Virág Székely, Orsolya Német, Áron Szepesi, Nóra Kucsma, György Várady, Gergely Szakács, Éva Bakos, and Csilla Özvegy-Laczka. Identification of novel cell-impermeant fluorescent substrates for testing the function and drug interaction of organic anion-transporting polypeptides, oatp1b1/1b3 and 2b1. *Scientific reports*, 8(1):1–12, 2018.
- [139] Virág Székely, Izabel Patik, Orsolya Ungvári, Ágnes Telbisz, Gergely Szakács, Éva Bakos, and Csilla Özvegy-Laczka. Fluorescent probes for the dual investigation of mrp2 and oatp1b1 function and drug interactions. *European Journal of Pharmaceutical Sciences*, page 105395, 2020.
- [140] Violetta Mohos, Eszter Fliszár-Nyúl, Orsolya Ungvári, Éva Bakos, Katalin Kuffa, Tímea Bencsik, Balázs Zoltán Zsidó, Csaba Hetényi, Ágnes Telbisz, Csilla Özvegy-Laczka, et al. Effects of chrysin and its major conjugated metabolites chrysin-7-sulfate and chrysin-7-glucuronide on cytochrome p450 enzymes and on oatp, p-gp, bcrp, and mrp2 transporters. *Drug Metabolism and Disposition*, 48(10):1064–1073, 2020.
- [141] Lora Mak, David Marcus, Andrew Howlett, Galina Yarova, Guus Duchateau, Werner Klaffke, Andreas Bender, and Robert C Glen. Metrabase: a cheminformatics and bioinformatics database for small molecule transporter data analysis and (q)SAR modeling. *Journal of Cheminformatics*, 7(1), jun 2015.
- [142] Ildikó Bacsa, Bianka Edina Herman, Rebeka Jójárt, Kevin Stefán Herman, János Wölfling, Gyula Schneider, Mónika Varga, Csaba Tömböly, Tea Lanišnik Rižner, Mihály Szécsi, and Erzsébet Mernyák. Synthesis and structure–activity relationships of 2- and/or 4-halogenated 13 β - and 13 α -estrone derivatives as enzyme inhibitors of estrogen biosynthesis. *Journal of Enzyme Inhibition and Medicinal Chemistry*, 33(1):1271–1282, jan 2018.
- [143] Peter W. Kenny and Jens Sadowski. Structure modification in chemical databases, jan 2005.
- [144] Naeem Shaikh, Mahesh Sharma, and Prabha Garg. Selective fusion of heterogeneous classifiers for predicting substrates of membrane transporters. *Journal of Chemical Information and Modeling*, 57(3):594–607, mar 2017.
- [145] Li-Jing Cui, Zhi-Feng Xie, Hu-Ri Piao, Gao Li, Kyu-Yun Chai, and Zhe-Shan Quan. Synthesis and anticonvulsant activity of 1-substituted-7-benzyloxy-4, 5-dihydro-[1, 2, 4] triazolo [4, 3-a] quinoline. *Biological and Pharmaceutical Bulletin*, 28(7):1216–1220, 2005.

- [146] Zhi-Feng Xie, Kyu-Yun Chai, Hu-Ri Piao, Kyung-Chell Kwak, and Zhe-Shan Quan. Synthesis and anticonvulsant activity of 7-alkoxyl-4, 5-dihydro-[1, 2, 4] triazolo [4, 3-a] quinolines. *Bioorganic & medicinal chemistry letters*, 15(21):4803–4805, 2005.
- [147] Hong-Guang Jin, Xian-Yu Sun, Kyu-Yun Chai, Hu-Ri Piao, and Zhe-Shan Quan. Anticonvulsant and toxicity evaluation of some 7-alkoxy-4, 5-dihydro-[1, 2, 4] triazolo [4, 3-a] quinoline-1 (2h)-ones. *Bioorganic & medicinal chemistry*, 14(20):6868–6873, 2006.
- [148] Hemant Kumar, Vladimir Kasho, Irina Smirnova, Janet S. Finer-Moore, H. Ronald Kaback, and Robert M. Stroud. Structure of sugar-bound LacY. *Proceedings of the National Academy of Sciences*, 111(5):1784–1788, jan 2014.
- [149] Anna Panek, Alina Świzdor, Natalia Milecka-Tronina, and Jarosław J Panek. Insight into the orientational versatility of steroid substrates—a docking and molecular dynamics study of a steroid receptor and steroid monooxygenase. *Journal of Molecular Modeling*, 23(3):96, 2017.
- [150] Jürgen-Hinrich Fuhrhop and Tianyu Wang. Bolaamphiphiles. *Chemical Reviews*, 104(6):2901–2938, jun 2004.
- [151] Mayur Fariya, Ankitkumar Jain, Vivek Dhawan, Sanket Shah, and Mangal S. Nargarsenker. Bolaamphiphiles: A pharmaceutical review. *Adv Pharm Bull*, 4(6):483–491, 2014.
- [152] Jonas Boström, Anders Hogner, and Stefan Schmitt. Do structurally similar ligands bind in a similar fashion? *Journal of Medicinal Chemistry*, 49(23):6716–6725, nov 2006.
- [153] Nan Li, Weifang Hong, Hong Huang, Hanping Lu, Guangyun Lin, and Mei Hong. Identification of amino acids essential for estrone-3-sulfate transport within transmembrane domain 2 of organic anion transporting polypeptide 1b1. *PLoS ONE*, 7(5):e36647, may 2012.
- [154] Zihui Fang, Jiujiu Huang, Jie Chen, Shaopeng Xu, Zhaojian Xiang, and Mei Hong. Transmembrane domain 1 of human organic anion transporting polypeptide 2b1 is essential for transporter function and stability. *Molecular Pharmacology*, 94(2):842–849, jun 2018.
- [155] Marianne K. DeGorter, Richard H. Ho, Brenda F. Leake, Rommel G. Tirona, and Richard B. Kim. Interaction of three regiospecific amino acid residues is required

- for OATP1b1 gain of OATP1b3 substrate specificity. *Molecular Pharmaceutics*, 9(4):986–995, mar 2012.
- [156] Nieng Yan. Structural advances for the major facilitator superfamily (MFS) transporters. *Trends in Biochemical Sciences*, 38(3):151–159, mar 2013.
- [157] Nicolae Solcan, Jane Kwok, Philip W Fowler, Alexander D Cameron, David Drew, So Iwata, and Simon Newstead. Alternating access mechanism in the POT family of oligopeptide transporters. *The EMBO Journal*, 31(16):3411–3421, jun 2012.

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

Part V

SUPPLEMENTARY

Supplementary Information 1

This section includes the supplementary information for Study 1: Alžběta Türková, Sankalp Jain, Barbara Zdrazil. Integrative data mining, scaffold analysis, and sequential binary classification models for exploring ligand profiles of hepatic organic anion transporting polypeptides. *Journal of chemical information and modeling*, 2018.

Supporting Information

Integrative Data Mining, Scaffold Analysis, and Sequential Binary Classification Models for Exploring Ligand Profiles of Hepatic Organic Anion Transporting Polypeptides (OATPs)

Alžběta Türková,[†] Sankalp Jain,[†] Barbara Zdrazil^{†}*

[†]University of Vienna, Department of Pharmaceutical Chemistry, Division of Drug Design and Medicinal Chemistry, Althanstraße 14, A-1090 Vienna, Austria.

*Corresponding author: barbara.zdrazil@univie.ac.at

Table of Contents

Table S1.....	p3
Table S2.....	p7
Table S3.....	p10
Table S4.....	p10
Table S5.....	p11
Table S6.....	p12
Table S7.....	p13
Table S8.....	p14
Table S9.....	p15
Table S10.....	p15
Table S11.....	p16
Table S12.....	p17
Table S13.....	p20
Table S14.....	p21
Table S15.....	p22
Table S16.....	p23
Table S17.....	p24

Figure S1.....	p25
Figure S2.....	p26
Figure S3.....	p27
Figure S4.....	p28
Figure S5.....	p30
Figure S6.....	p31
Figure S7.....	p32
Figure S8.....	p33
Figure S9.....	p34
Figure S10.....	p35
Description of Supplementary Data Files.....	p36

Table S1. Data sets annotated with bioactivity end point “inhibition” from ChEMBL: Reference to the original manuscript, PMID, compound concentration used in the experiment, number of unique compounds (after standardization procedure) and recommended activity thresholds are given.

Reference	PMID	c [uM]	Cmpd. number	Recommended threshold
Marada, V.V.; Flörl, S.; Kühne, A.; Burckhardt, G.; Hagos, Y. Interaction of human organic anion transporter polypeptides 1B1 and 1B3 with antineoplastic compounds. <i>Eur. J. Med. Chem.</i> 2015 , <i>92</i> , 723-731.	25618019	100	8	> 60% is active
De Bruyn, T.; van Westen, G.J.P., IJzerman, A.P., Stieger, B., de Witte, P., Augustijns, P.F., Annaert, P.P. Structure-Based Identification of OATP1B1/3 Inhibitors. <i>Mol. Pharmacol.</i> 2013 , <i>83</i> (6), 1257-1267.	23571415	10	1367	> 50% is active
Karlgrén, M.; Vildhede, A.; Norinder, U.; Wisniewski, J. R.; Kimoto, E.; Lai, Y.; Haglund, U.; Artursson, P. Classification of Inhibitors of Hepatic Organic Anion Transporting Polypeptides (OATPs): Influence of Protein Expression on Drug–Drug Interactions. <i>J. Med. Chem.</i> 2012 , <i>55</i> (10), 4740–4763.	22541068	20	221	> 50% is active
Kobayashi, D.; Nozawa, T.; Imai, K.; Nezu, J.; Tsuji, A.; Tamai, I. Involvement of human organic anion transporting polypeptide OATP-B (SLC21A9) in pH-dependent transport across intestinal apical membrane. <i>J. Pharmacol. Exp. Ther.</i> 2003 , <i>306</i> (2), 703-708.	12724351	1000 10000	2	REMOVED

St-Pierre, M. V.; Hagenbuch, B.; Ugele, B.; Meier, P. J.; Stallmach, T. Characterization of an organic anion-transporting polypeptide (OATP-B) in human placenta. <i>J. Clin. Endocrinol. Metab.</i> 2002 , 87(4), 1856-1863	11932330	100	1	> 60% is active (based on the threshold from Marada, V.V.; Flörl, S.; Kühne, A.; Burckhardt, G.; Hagos, Y. Interaction of human organic anion transporter polypeptides 1B1 and 1B3 with antineoplastic compounds. <i>Eur. J. Med. Chem.</i> 2015 , 92, 723-731.)
Ishiguro, N.; Shimizu, H.; Kishimoto, W.; Ebner, T.; Schaefer, O. Evaluation and prediction of potential drug-drug interactions of linagliptin using in vitro cell culture methods. <i>Drug Metab. Dispos.</i> 2013 , 41(1), 149-158	23073734	100	1	> 60% is active (based on the threshold from Marada, V.V.; Flörl, S.; Kühne, A.; Burckhardt, G.; Hagos, Y. Interaction of human organic anion transporter polypeptides 1B1 and 1B3 with antineoplastic compounds. <i>Eur. J. Med. Chem.</i> 2015 , 92, 723-731.)

<p>Sandhu, P.; Lee, W.; Xu, X.; Leake, B. F.; Yamazaki, M.; Stone, J. A.; Lin, J. H.; Pearson, P. G.; Kim, R. B. Hepatic uptake of the novel antifungal agent caspofungin. <i>Drug Metab. Dispos.</i> 2005, 33(1), 676</p>	<p>15716364</p>	<p>100 10</p>	<p>2</p>	<p>> 60% is active (for c=100 uM)</p> <p>(based on the threshold from Marada, V.V.; Flörl, S.; Kühne, A.; Burckhardt, G.; Hagos, Y. Interaction of human organic anion transporter polypeptides 1B1 and 1B3 with antineoplastic compounds. <i>Eur. J. Med. Chem.</i> 2015, 92, 723-731.)</p> <p>> 50% is active (for c=10 uM)</p> <p>(based on the threshold from De Bruyn, T.; van Westen, G.J.P., IJzerman, A.P., Stieger, B., de Witte, P., Augustijns, P.F., Annaert, P.P. Structure-Based Identification of OATP1B1/3 Inhibitors. <i>Mol. Pharmacol.</i> 2013, 83 (6), 1257-1267.)</p>
<p>Nozawa, T.; Minami, H.; Sugiura, S.; Tsuji, A.; Tamai, I. Role of organic anion transporter OATP1B1 (OATP-C) in hepatic uptake of irinotecan and its active metabolite, 7-ethyl-10-hydroxycamptothecin: in vitro evidence and effect of single nucleotide polymorphisms. <i>Drug Metab. Dispos.</i>, 2005, 33(3), 434-439</p>	<p>15608127</p>	<p>10</p>	<p>2</p>	<p>> 50% is active</p> <p>(based on the threshold from De Bruyn, T.; van Westen, G.J.P., IJzerman, A.P., Stieger, B., de Witte, P., Augustijns, P.F., Annaert, P.P. Structure-Based Identification of OATP1B1/3 Inhibitors. <i>Mol. Pharmacol.</i> 2013, 83 (6), 1257-1267.)</p>

Nozawa, T.; Sugiura, S.; Nakajima, M.; Goto, A.; Yokoi, T.; Nezu, J. I.; Tsuji, A.; Tamai, I. Involvement of organic anion transporting polypeptides in the transport of troglitazone sulfate: implications for understanding troglitazone hepatotoxicity. <i>Drug Metab. Dispos.</i> 2004 , 32(3), 291-294	14977862	1 10	6	> 50% is active (based on the threshold from De Bruyn, T.; van Westen, G.J.P., IJzerman, A.P., Stieger, B., de Witte, P., Augustijns, P.F., Annaert, P.P. Structure-Based Identification of OATP1B1/3 Inhibitors. <i>Mol. Pharmacol.</i> 2013 , 83 (6), 1257-1267.)
Nozawa, T.; Tamai, I.; Sai, Y.; Nezu, J. I.; Tsuji, A. Contribution of organic anion transporting polypeptide OATP-C to hepatic elimination of the opioid pentapeptide analogue [d-Ala 2, d-Leu5]-enkephalin. <i>J. Pharm. Pharmacol.</i> 2003 , 55(7), 1013-1020	12906759	1000 5000	3	REMOVED
König, J.; Cui, Y.; Nies, A. T.; Keppler, D. A novel human organic anion transporting polypeptide localized to the basolateral hepatocyte membrane. <i>Am. J. Physiol. Gastrointest. Liver Physiol.</i> 2000 , 278(1): G156-G164	10644574	50 1000	2	REMOVED
Satoh, H.; Yamashita, F.; Tsujimoto, M.; Murakami, H.; Koyabu, N.; Ohtani, H.; Sawada, Y. Citrus juices inhibit the function of human organic anion transporting polypeptide OATP-B. Drug metabolism and disposition. <i>Drug Metab. Dispos.</i> 2005 , 33(1): 518	15640378	1 10	4	> 50% is active (based on the threshold from De Bruyn, T.; van Westen, G.J.P., IJzerman, A.P., Stieger, B., de Witte, P., Augustijns, P.F., Annaert, P.P. Structure-Based Identification of OATP1B1/3 Inhibitors. <i>Mol. Pharmacol.</i> 2013 , 83 (6), 1257-1267.)

Table S2. Data sets annotated with bioactivity end point “inhibition” from Metrabase¹: Reference to the original manuscript, PMID, compound concentration used in the experiment, number of unique compounds (after standardization procedure) and recommended activity thresholds are given.

Reference	PMID	c [uM]	Cmpd. number	Recommended threshold
Karlgren, M.; Vildhede, A.; Norinder, U.; Wisniewski, J. R.; Kimoto, E.; Lai, Y.; Haglund, U.; Artursson, P. Classification of Inhibitors of Hepatic Organic Anion Transporting Polypeptides (OATPs): Influence of Protein Expression on Drug–Drug Interactions. <i>J. Med. Chem.</i> 2012 , 55 (10), 4740–4763.	22541068	20	5	> 50% is active
Sai, Y.; Kaneko, Y.; Ito, S.; Mitsuoka, K.; Kato, Y.; Tamai, I.; Artursson, P.; Tsuji, A. Predominant contribution of organic anion transporting polypeptide OATP-B (OATP2B1) to apical uptake of estrone-3-sulfate by human intestinal Caco-2 cells. <i>Drug Metab Dispos.</i> 2006 , 34(8):1423-31	16714376	1000 10000	2	REMOVED

¹ This table does not include ~150 data sets from Metrabase where there were additional sources available (e.g. from ChEMBL).

Fuchikami, H.; Satoh, H.; Tsujimoto, M.; Ohdo, S.; Ohtani, H.; Sawada, Y. Effects of Herbal Extracts on the Function of Human Organic Anion Transporting Polypeptide, OATP-B. <i>Drug Metab Dispos.</i> 2006 , 34(4):577-82	16415120	1 10 100	4	> 60% is active (based on the threshold from Marada, V.V.; Flörl, S.; Kühne, A.; Burckhardt, G.; Hagos, Y. Interaction of human organic anion transporter polypeptides 1B1 and 1B3 with antineoplastic compounds. <i>Eur. J. Med. Chem.</i> 2015 , 92, 723-731.)
Kis, O.; Zastre, J. A.; Ramaswamy, M.; Bendayan, R. pH dependence of organic anion-transporting polypeptide 2B1 in Caco-2 cells: potential role in antiretroviral drug oral bioavailability and drug–drug interactions. <i>J Pharmacol Exp Ther.</i> 2010 , 334(3):1009-22	20507927	100	4	> 60% is active (based on the threshold from Marada, V.V.; Flörl, S.; Kühne, A.; Burckhardt, G.; Hagos, Y. Interaction of human organic anion transporter polypeptides 1B1 and 1B3 with antineoplastic compounds. <i>Eur. J. Med. Chem.</i> 2015 , 92, 723-731.)
Grube, M.; Köck, K.; Oswald, S.; Draber, K.; Meissner, K.; Eckel, L.; Böhm, M.; Felix, S. B.; Vogelgesang, S.; Jedlitschky, G.; Siegmund, W.; Warzok, R.; Kroemer, H. K. Organic anion transporting polypeptide 2B1 is a high-affinity transporter for atorvastatin and is expressed in the human heart. <i>Clin Pharmacol Ther.</i> 2006 , 80(6):607-20	17178262	10	1	> 50% is active

Grube, M.; Kock, K.; Karner, S.; Reuther, S.; Ritter, C. A.; Jedlitschky, G.; Kroemer, H. K. Modification of OATP2B1 mediated transport by steroid hormones. Molecular pharmacology. <i>Mol Pharmacol.</i> 2006 , 70(5):1735-41.	16908597	10	2	> 50% is active (based on the threshold from De Bruyn, T.; van Westen, G.J.P., IJzerman, A.P., Stieger, B., de Witte, P., Augustijns, P.F., Annaert, P.P. Structure-Based Identification of OATP1B1/3 Inhibitors. <i>Mol. Pharmacol.</i> 2013 , 83 (6), 1257-1267.)
Reyes, M.; Benet, L. Z. Effects of uremic toxins on transport and metabolism of different biopharmaceutics drug disposition classification system xenobiotics. <i>J Pharm Sci.</i> 2011 , 100(9):3831-42	21618544	400	2	REMOVED

Table S3. Removed substrates from OATP1B1 substrate data set with conflicting annotations (median value equals to 0.5). Compound names and InChIKeys are provided.

Name	InChIKey
Methylaminopterin	FBOZXECLQNJBKD-UHFFFAOYSA-N
No name	HMBKEXWSKGGBBT-WCHIAOBISA-N
3,7-Dihydroxycholan-24-oic acid	RUDATBOHQWOJDD-UHFFFAOYSA-N
(8S,9S,13S,14S)-17-ethynyl-17-hydroxy-13-methyl-3-sulfooxy-7,8,9,11,12,14,15,16-octahydro-6H-cyclopenta[a]phenanthrene	WLGIWVFFGMPRLM-UHFFFAOYSA-N

Table S4. Removed substrates from OATP1B3 substrate data set with conflicting annotations (median value equals to 0.5). Compound names and InChIKeys are provided.

Name	InChIKey
Caloxetic acid	AQOXEJNYXXLRQQ-UHFFFAOYSA-N

Table S5. Removed substrates from OATP2B1 substrate data set with conflicting annotations (median value equals to 0.5). Compound names and InChIKeys are provided.

Name	InChIKey
Breviscapine	DJSISFGPUUYILV-UHFFFAOYSA-N
Ethyl-diisopropylamine	FUWQJNSUHNKFNP-UHFFFAOYSA-N
Bosentan	GJPICJJRGTNOD-UHFFFAOYSA-N
6-{{5,7-dihydroxy-2-(4-hydroxyphenyl)-4-oxo-4H-chromen-6-yl}oxy}-3,4,5-trihydroxyoxane-2-carboxylic acid	HBLWMMBFOKSEKW-UHFFFAOYSA-N
Talinolol	MXFWWQICDIZSOA-UHFFFAOYSA-N
No name	QQCSUWGQBREWRO-CYVLTUHYSA-N
2,2',4,4'-Tetrabromodiphenyl ether	XYBSIYMGXVUVGY-UHFFFAOYSA-N

Table S6. Removed inhibitors from OATP1B1 inhibitor data set with conflicting annotations (median value equals to 0.5). Compound names and InChIKeys are provided.

Name	InChIKey
Lanosteryl acetate	BQPPJGMMIYJVBR-UHFFFAOYSA-N
Capsazepine	DRCMAZOSEIMCHM-UHFFFAOYSA-N
Zopiclone	GBBSUAFBMRNDJC-UHFFFAOYSA-N
3-Episarsasapogenin	GMBQZIIUCVWOC-UHFFFAOYSA-N
Quinine	LOUPRKONTZGTKE-UHFFFAOYSA-N
Spironolactone	LXMSZDCAJNLERA-UHFFFAOYSA-N
Estradiol-3-sulfate	QZIGLSSUDXBTJ-UHFFFAOYSA-N
Taurocholic acid	WBWWGRHZICKQZ-UHFFFAOYSA-N
10-acetyloxy-1,2,6a,6b,9,9,12a-heptamethyl-13-oxo-1,2,3,4,5,6,6a,7,8,8a,10,11,12,14b-tetradecahydronicene-4a-carboxylic acid	XDHCWTUZCOFKRH-UHFFFAOYSA-N
Eltrombopag	XDWLKQMMKQXPV-QYQHSDTDSA-N
Cephalothin	XIURVHNZVLADCM-UHFFFAOYSA-N

Table S7. Removed inhibitors from OATP1B3 inhibitor data set with conflicting annotations (median value equals to 0.5). Compound names and InChIKeys are provided.

Name	InChIKey
Lanosteryl acetate	BQPPJGMMIYJVBR-UHFFFAOYSA-N
Capsazepine	DRCMAZOSEIMCHM-UHFFFAOYSA-N
Zopiclone	GBBSUAFBMRNDJC-UHFFFAOYSA-N
17beta-estradiol 17beta-D-glucuronide	MTKNDQYHASLID-UHFFFAOYSA-N
Vincristine	OGWKCGZFUXNPDA-UHFFFAOYSA-N
Hoechst 33342	PRDFBSVERLRMY-UHFFFAOYSA-N
(1-[[[(1-{3-[(E)-2-(7-chloroquinolin-2-yl)ethenyl]phenyl}-3-[2-(2-hydroxypropan-2-yl)phenyl]propyl)sulfanyl]methyl}cyclopropyl)acetic acid	UCHDWCPVSPXUMX-XNTDXEJSSA-N
No name	UWNQSONNONTGTF-UHFFFAOYSA-N
Nefazodone	VRBKIVRKKCLPHA-UHFFFAOYSA-N
No name	WTDQOIVETOYZMZ-UHFFFAOYSA-N
Antamanide	WTINJQXJTHUFRF-UHFFFAOYSA-N
10-acetyloxy-1,2,6a,6b,9,9,12a-heptamethyl-13-oxo-1,2,3,4,5,6,6a,7,8,8a,10,11,12,14b-tetradecahydronicene-4a-carboxylic acid	XDHCWTUZCOFKRH-UHFFFAOYSA-N
Levothyroxine	XUIIKFGFIJCVMT-UHFFFAOYSA-N

Table S8. Removed inhibitors from OATP1B3 inhibitor data set with conflicting annotations (median value equals to 0.5). Compound names and InChIKeys are provided.

Name	InChIKey
Dehydroepiandrosterone sulfate	CZWCKYRVOZZJNM-UHFFFAOYSA-N
Estrone-d4	DNXHEGUUPJUMQT-UHFFFAOYSA-N
Repaglinide	FAEKWTJYAYMJKF-UHFFFAOYSA-N
6',7'-dihydroxy Bergamottin	IXZUPBUEKFXTSD-MDWZMJQESA-N
Epicatechin-3-gallate	LSHVYAFMTMFKBA-UHFFFAOYSA-N
Nobiletin	MRIAQLRQZPPDS-UHFFFAOYSA-N
Nelfinavir	QAGYKUNXZHKKMR-UHFFFAOYSA-N
Quercetin	REFJWTPEDEVJJIY-UHFFFAOYSA-N
Tangeretin	ULSUXBXHSYSGDT-UHFFFAOYSA-N
Efavirenz	XPOQHMRABVBWPR-UHFFFAOYSA-N
Rosiglitazone	YASAKCUCGLMORW-UHFFFAOYSA-N

Table S9. Percentage of conflicting compound activities based on the comparison of data from ChEMBL and Metrabase.

Activity	OATP1B1	OATP1B3	OATP2B1
(non-)substrates	40%	36%	60%
(non-)inhibitors	64%	74%	10%

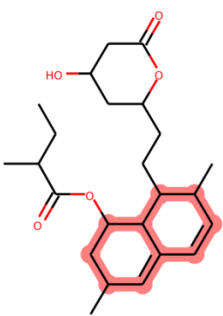
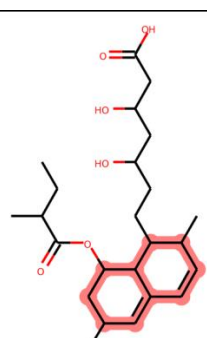
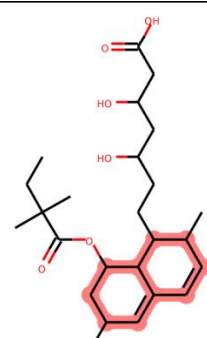
Table S10. “Dense dataset” for hepatic OATP substrates.

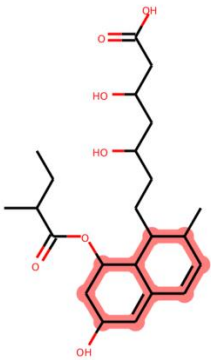
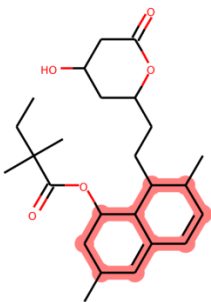
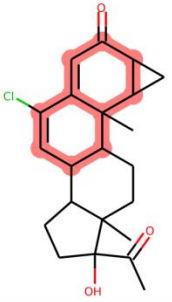
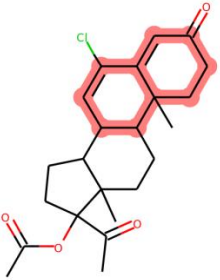
Class	Nr.	Description
0	6	General substrates
1	0	1B1+1B3 overlapping substrates
2	2	1B1+2B1 overlapping substrates
3	0	1B3+2B1 overlapping substrates
4	0	Selective 1B1 substrates
5	2	Selective 1B3 substrates
6	2	Selective 2B1 substrates
7	1	General non-substrates

Table S11. “Dense dataset” for hepatic OATP inhibitors.

Class	Nr.	Description
0	26	General inhibitors
1	14	1B1+1B3 overlapping inhibitors
2	10	1B1+2B1 overlapping inhibitors
3	1	1B3+2B1 overlapping inhibitors
4	12	Selective 1B1 inhibitors
5	2	Selective 1B3 inhibitors
6	10	Selective 2B1 inhibitors
7	88	General non- inhibitors

Table S12. Ten detected compounds with hexahydronaphthalene-associated scaffold (highlighted in red) with pharmacological profiles included: “1”....active; “0”....inactive; “?”....missing annotation.

Name	Structure	OATP1B1	OATP1B3	OATP2B1
Mevinolin		1	0	0
Lovastatin Acid		1	?	?
Simvastatin Acid		1	0	?

Pravastatin		1	0	0
Simvastatin		1	0	0
Cyprotene		1	?	?
Chlormadinone Acetate		1	1	?

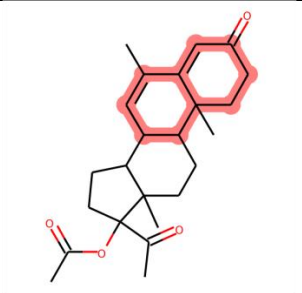
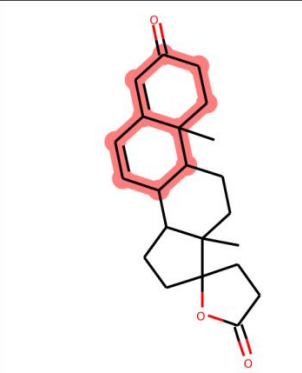
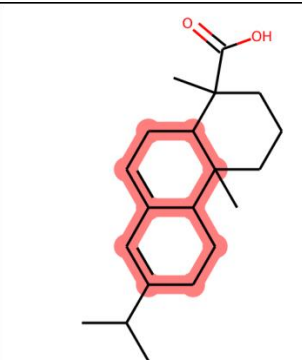
Megestrol Acetate		0	0	?
Canrenone		1	0	?
Abietic Acid		1	1	?

Table S13: Results on Level 1 (All inhibitors + general non-inhibitors) for all calculated statistics metrics: Sensitivity, Specificity, Balanced Accuracy and MCC. The performance is given for both 10-fold cross-validation and on the external test set. With bold font are depicted those models that gave the best results.

Model	Validation	Sensitivity	Specificity	Balanced Accuracy	MCC	Descriptor (% involvement in subset selection of attributes)
RandomTree	Training set 10fold-CV	0.893	0.339	0.616	0.255	SlogP, AMW
RandomTree	Test Set	0.805	0.346	0.576	0.142	
CostSensitive Classifier	Training set 10fold-CV	0.889	0.355	0.622	0.264	
CostSensitive Classifier	Test Set	0.788	0.462	0.625	0.222	
Stratified bagging	Training set	0.760	0.790	0.775	0.455	
Stratified bagging	Test Set	0.796	0.769	0.783	0.477	

Level 1: All inhibitors + general non-inhibitors (Training set)

CostSensitive Classifier: Cost matrix [0.0, 4.0; 1.0, 0.0]

Stratified Bagging with 64 bags

Table S14: Results on OPATP1B1 inhibition data set for all calculated statistics metrics: Sensitivity, Specificity, Balanced Accuracy and MCC. The performance is given for both 10-fold cross-validation and on the external test set. With bold font are depicted those models that gave the best results.

Model	Validation	Sensitivity	Specificity	Balanced Accuracy	MCC	Descriptor (% involvement in subset selection of attributes)
RandomTree	Training set 10fold-CV	0.547	0.830	0.689	0.370	SlogP, SMR, AMW, TPSA
RandomTree	Test Set	0.580	0.828	0.704	0.396	
CostSensitive Classifier	Training set 10fold-CV	0.539	0.817	0.678	0.345	
CostSensitive Classifier	Test Set	0.540	0.802	0.671	0.328	
Stratified bagging	Training set	0.703	0.799	0.751	0.462	
Stratified bagging	Test Set	0.730	0.809	0.769	0.497	

CostSensitive Classifier: Cost matrix [0.0, 1.0; 3.0, 0.0]

Stratified Bagging with 64 bags

Table S15: Results on OPATP1B3 inhibition data set for all calculated statistics metrics: Sensitivity, Specificity, Balanced Accuracy and MCC. The performance is given for both 10-fold cross-validation and on the external test set. With bold font are depicted those models that gave the best results.

Model	Validation	Sensitivity	Specificity	Balanced Accuracy	MCC	Descriptor (% involvement in subset selection of attributes)
RandomTree	Training set 10fold-CV	0.496	0.895	0.696	0.384	SlogP, SMR, AMW, TPSA
RandomTree	Test Set	0.373	0.899	0.636	0.282	
CostSensitive Classifier	Training set 10fold-CV	0.504	0.891	0.697	0.383	
CostSensitive Classifier	Test Set	0.424	0.892	0.658	0.316	
Stratified bagging	Training set	0.748	0.834	0.791	0.486	
Stratified bagging	Test Set	0.746	0.829	0.787	0.476	

CostSensitive Classifier: Cost matrix [0.0, 1.0; 5.0, 0.0]

Stratified Bagging with 64 bags

Table S16: Results on OPATP2B1 inhibition data set for all calculated statistics metrics: Sensitivity, Specificity, Balanced Accuracy and MCC. The performance is given for both 10-fold cross-validation and on the external test set. With bold font are depicted those models that gave the best results.

Model	Validation	Sensitivity	Specificity	Balanced Accuracy	MCC	Descriptor (% involvement in subset selection of attributes)
RandomTree	Training set 10fold-CV	0.535	0.856	0.695	0.400	SMR, AMW
RandomTree	Test Set	0.526	0.840	0.683	0.373	
CostSensitive Classifier	Training set 10fold-CV	0.535	0.839	0.687	0.377	
CostSensitive Classifier	Test Set	0.526	0.860	0.693	0.400	
Stratified bagging	Training set	0.698	0.771	0.734	0.434	
Stratified bagging	Test Set	0.632	0.840	0.736	0.464	

CostSensitive Classifier: Cost matrix [0.0, 1.0; 3.0, 0.0]

Stratified Bagging with 64 bags

Table S17: Summary statistics for molecular descriptors calculated for inhibitors of OATP1B1, OATP1B3, and OATP2B1. The median and mean values for lipophilicity (SlogP), molecular refractivity (SMR), the topological polar surface area (TPSA), average molecular weight (AMW), the number of rotatable bonds (RotB), the number of amide bonds (AmideBonds), the number of rings (NumRings), and the number of aromatic carbocycles (Aromatic Carbocycles) are given.

	SlogP(median)	SlogP(mean)	SMR(median)	SMR(mean)	TPSA(median)	TPSA(mean)	AMW(median)	AMW(mean)
OATP1B1 inhibitors	3.99	3.88	123.70	141.42	106.86	126.30	471.64	537.04
OATP1B1 non-inhibitors	2.24	2.20	81.58	86.91	66.76	81.31	300.27	327.92
OATP1B3 inhibitors	4.22	4.21	132.00	148.31	113.08	133.75	504.67	564.47
OATP1B3 non-inhibitors	2.33	2.26	82.39	87.05	68.22	81.78	303.60	330.61
OATP2B1 inhibitors	3.64	3.71	127.47	131.36	109.33	120.63	481.99	503.34
OATP2B1 non-inhibitors	2.48	2.22	95.40	104.36	84.83	99.87	358.16	394.43
	RotB(median)	RotB(mean)	AmideBonds(median)	AmideBonds(mean)	NumRings(median)	NumRings(mean)	AromaticCarbo-cycles(median)	AromaticCarbo-cycles(mean)
OATP1B1 inhibitors	5	6.56	0	1.03	4	4.03	1	1.10
OATP1B1 non-inhibitors	3	3.96	0	0.34	3	2.77	1	0.92
OATP1B3 inhibitors	6	6.90	0	0.80	4	4.37	1	1.21
OATP1B3 non-inhibitors	3	3.87	0	0.35	3	2.79	1	0.92
OATP2B1 inhibitors	7	7.52	0	0.74	4	3.74	2	1.76
OATP2B1 non-inhibitors	4	4.99	0	0.68	3	3.18	1	1.20

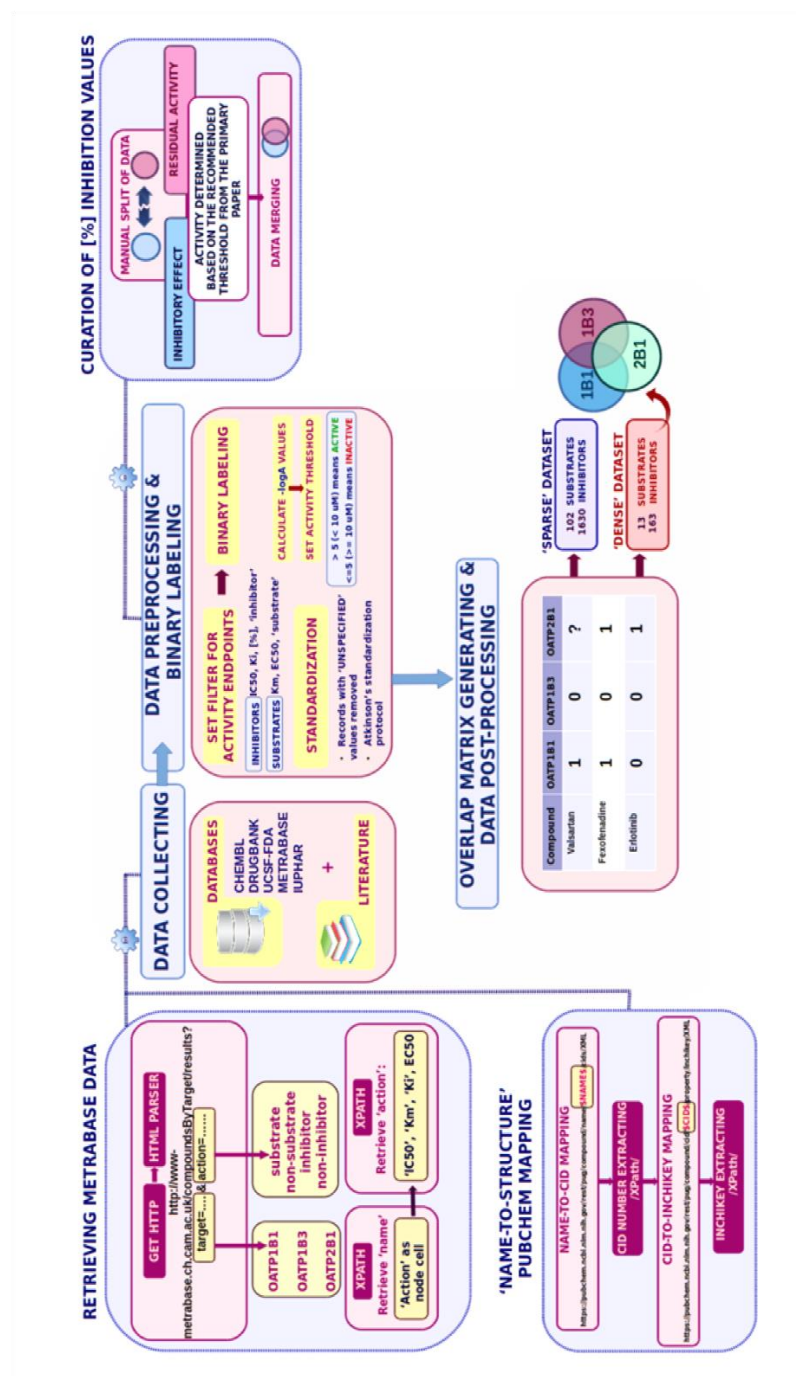


Figure S1: Schematic workflow for integrative data mining and curation.

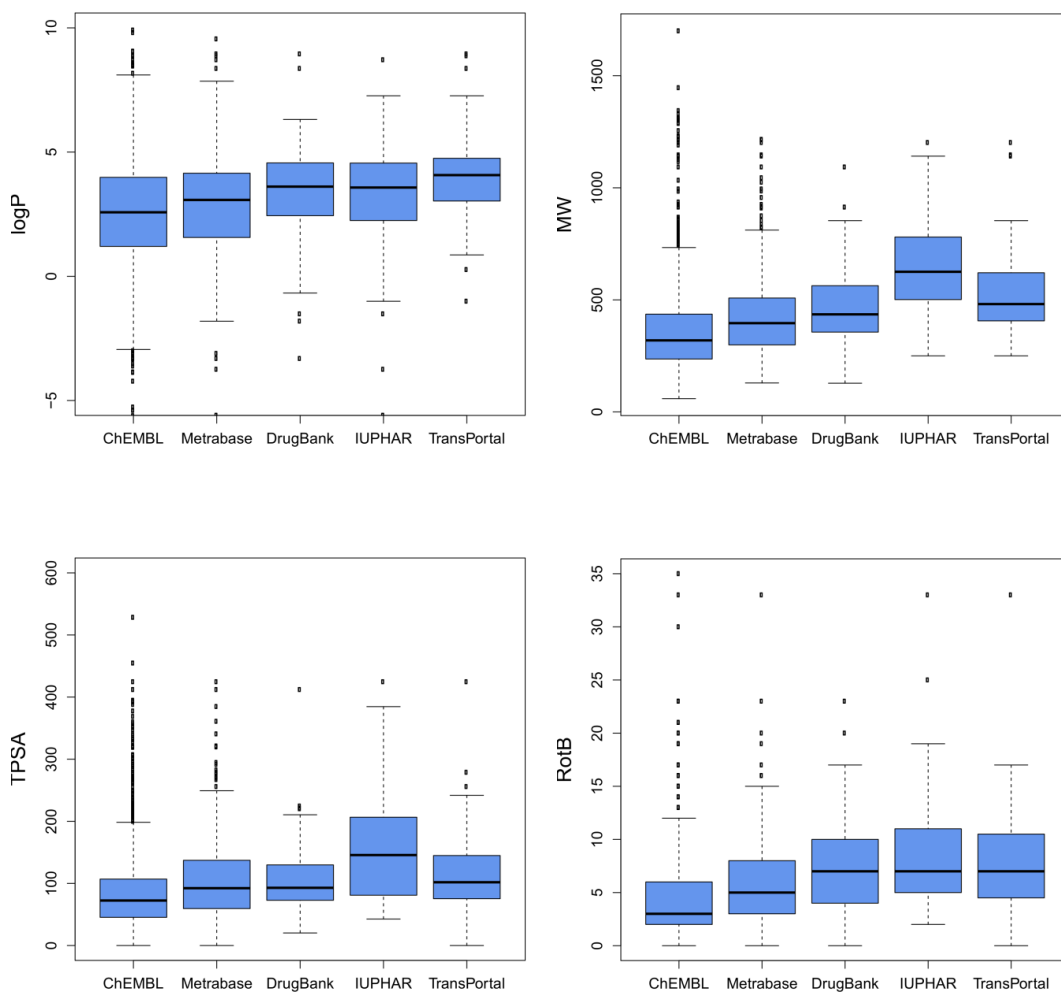


Figure S2. Box- and Whisker-Plots showing the distribution of molecular properties for compounds measured against human OATP1B1, OATP1B3, and OATP2B1 originating from five different data sources (ChEMBL, Metrabase, DrugBank, IUPHAR, TransPortal): partition coefficient (logP), molecular weight (MW), topological surface area (TPSA), number of rotatable bonds (RotB).

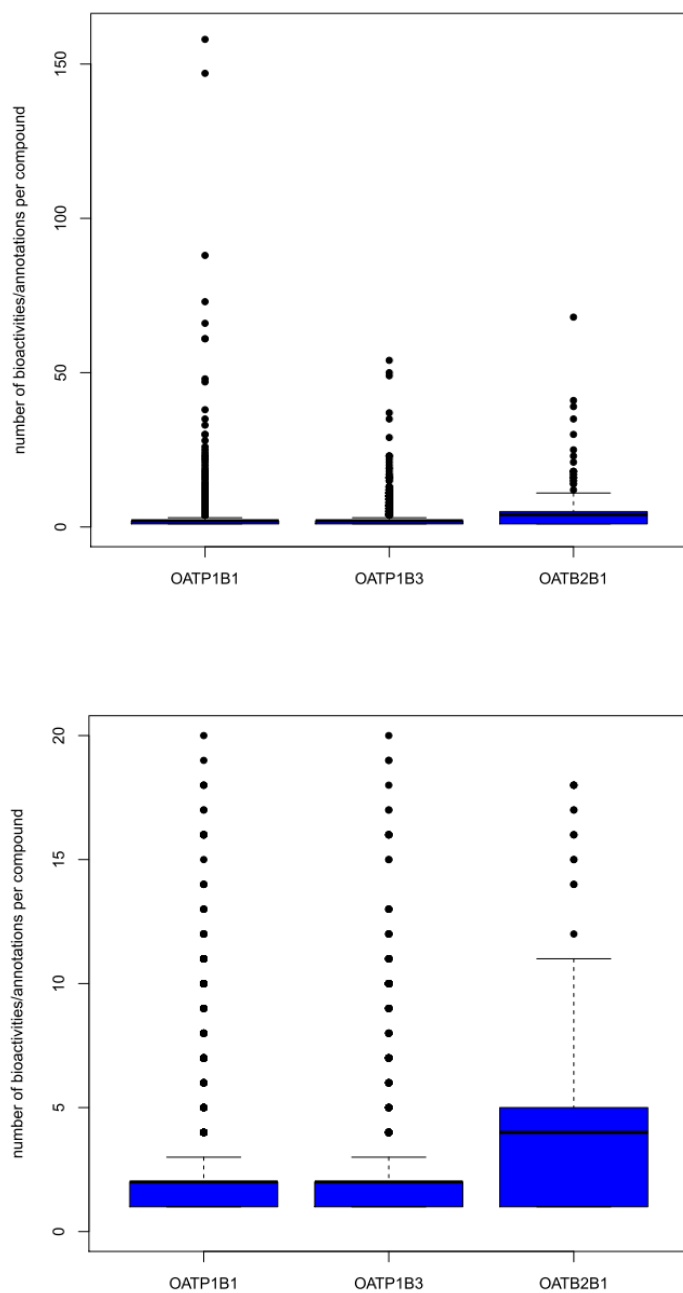
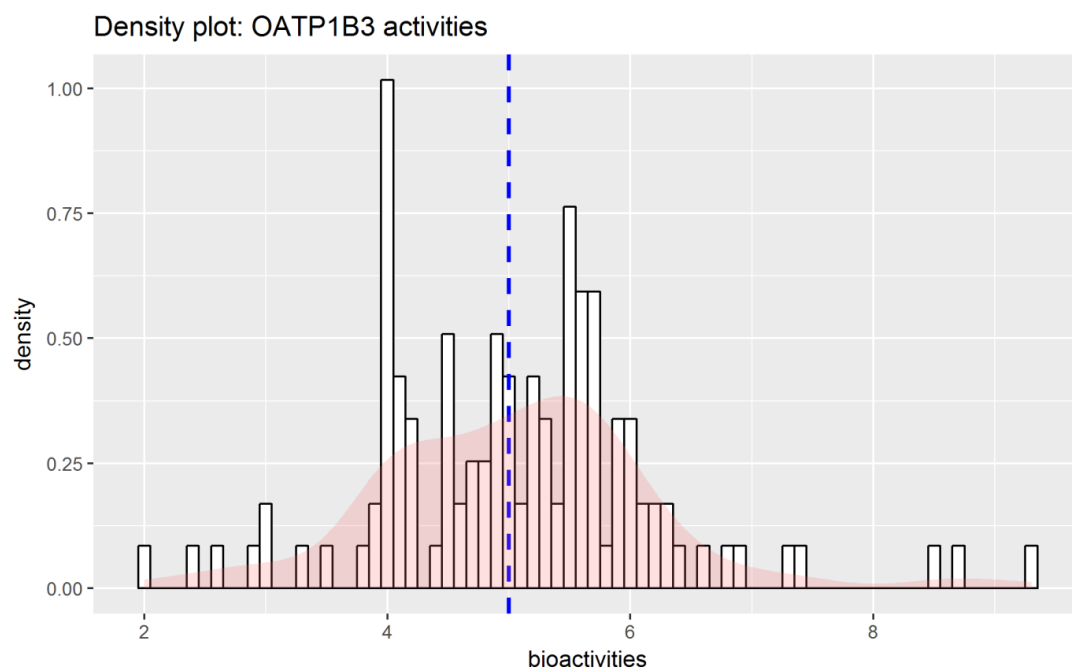
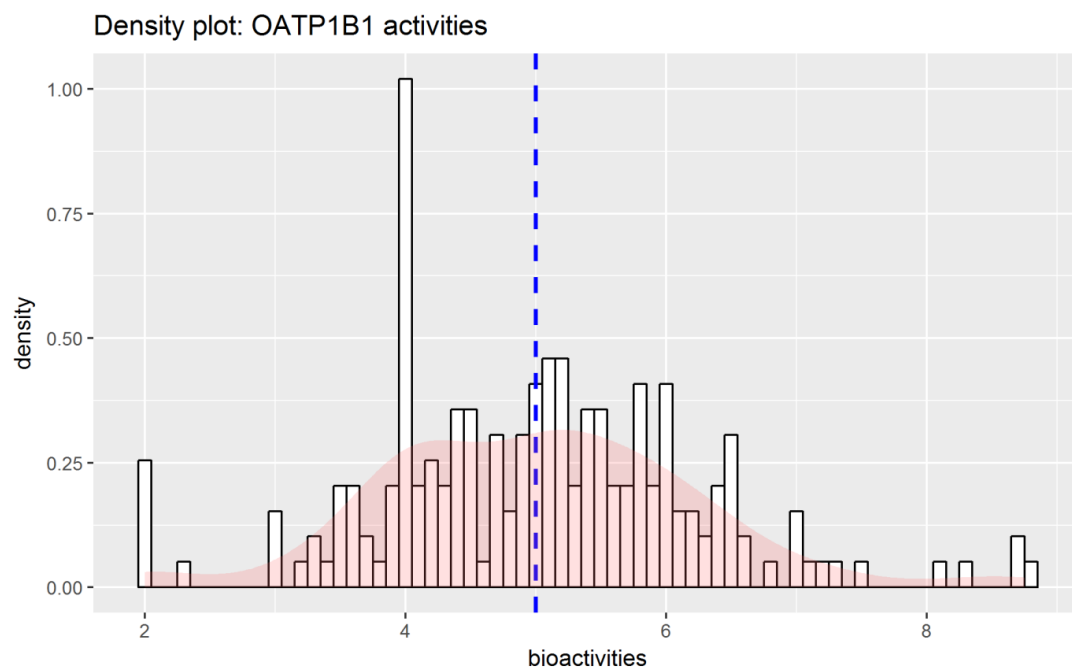


Figure S3. Number of bioactivities/annotations per unique compound for hepatic OATPs: upper plot....full range of bioactivity values displayed; lower plot....zoomed-in view with max. 20 bioactivities displayed.



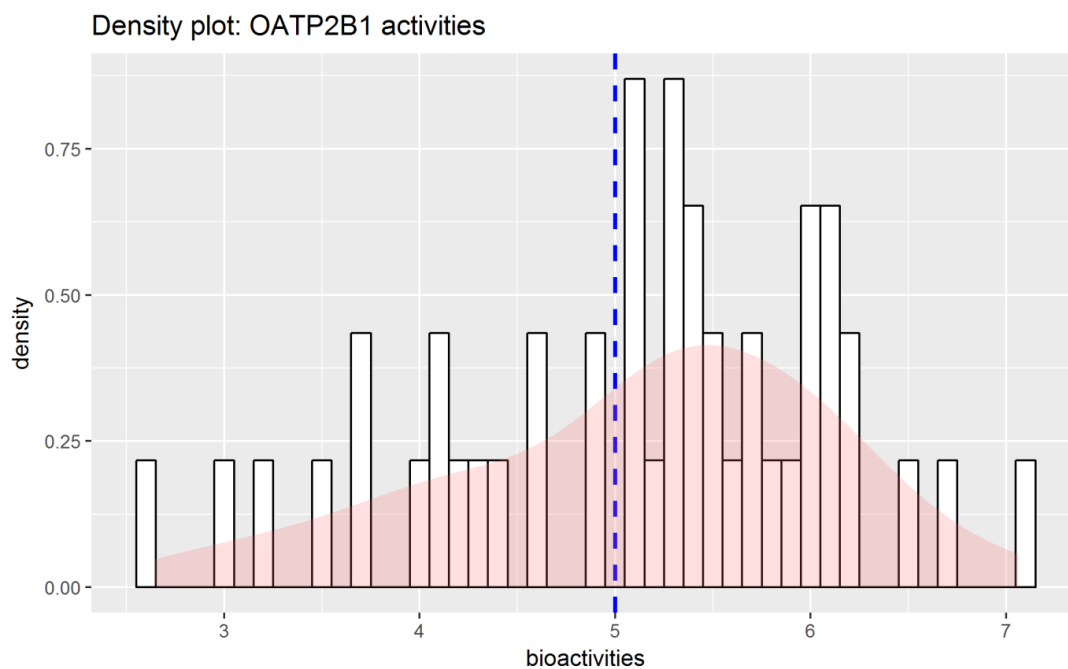


Figure S4. Histograms showing the distribution of median bioactivities [in negative logarithmic representation (molar)] for OATP1B1, OATP1B3, and OATP2B1. The chosen cut off for classifying compounds into actives and inactives ($10\mu\text{M}$; $-\log(\text{activity}[\text{molar}]) = 5$) is shown as a dashed blue line.

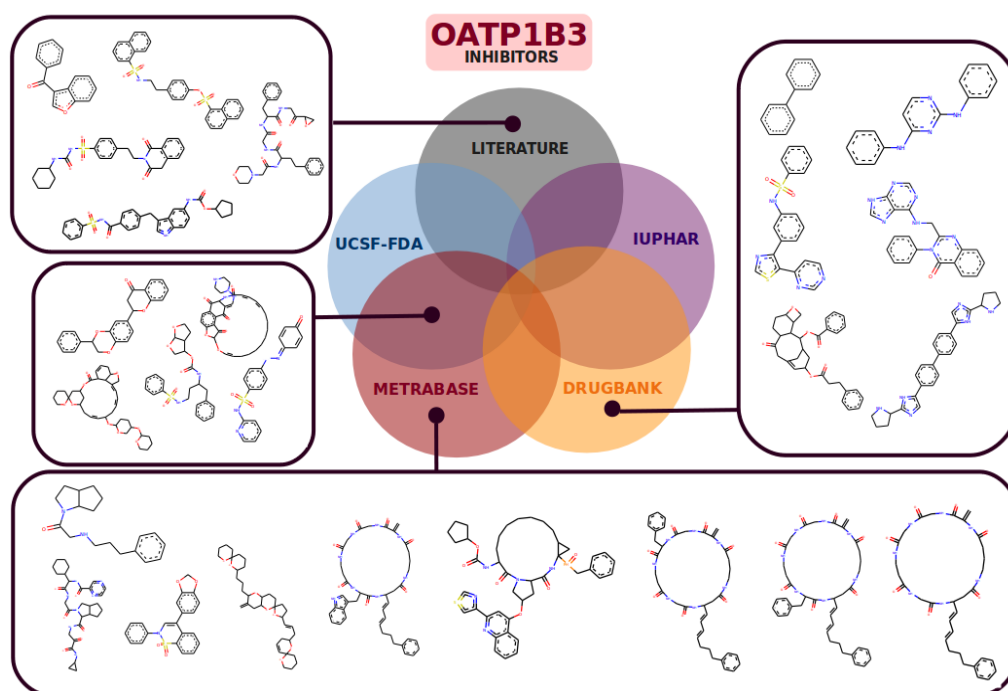


Figure S5. Murcko Scaffolds for OATP1B3 inhibitors retrieved from other databases than ChEMBL.

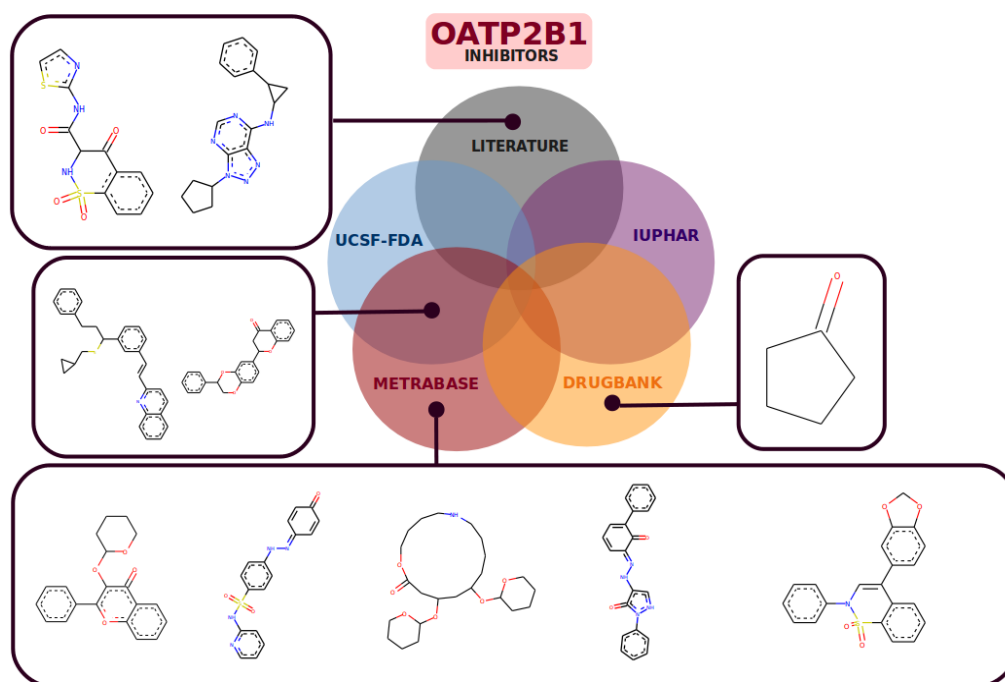


Figure S6. Murcko Scaffolds for OATP2B1 inhibitors retrieved from other databases than CHEMBL.

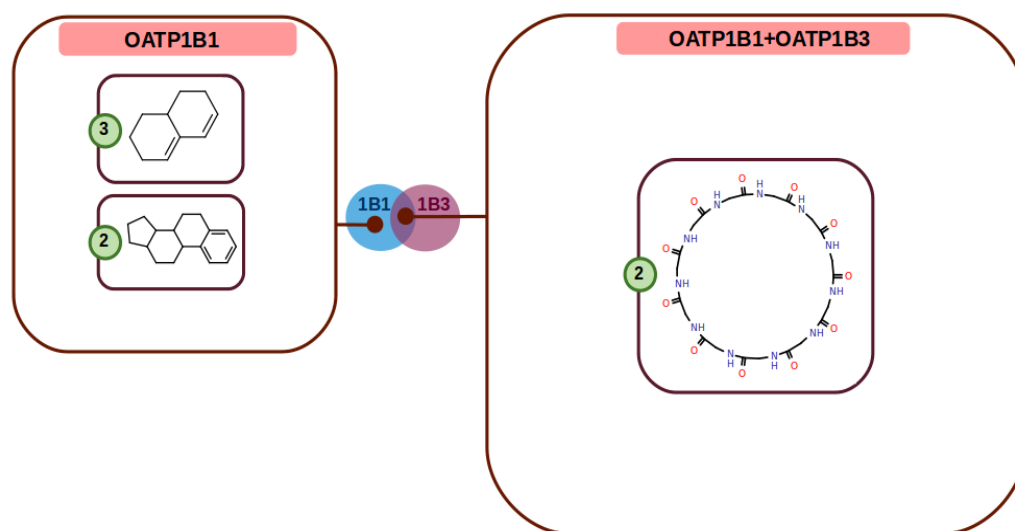


Figure S7. Enriched scaffolds ($p\text{-value} < 0.05$) for hepatic OATP inhibitors considering the dense data set (with complete pharmacological profile). Numbers in green circles correspond to the numbers of associated compounds for a respective scaffold cluster.

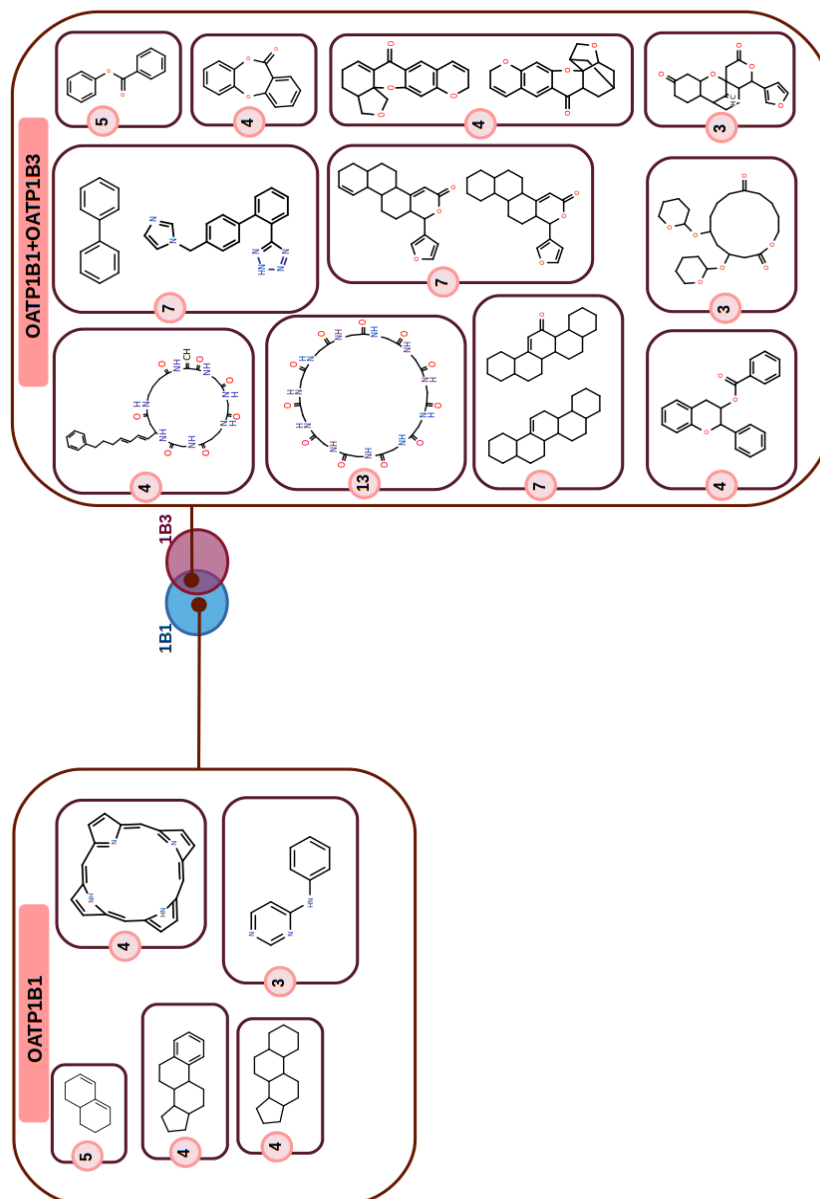


Figure S8. Enriched scaffolds (p-value < 0.05) for hepatic OATP inhibitors excluding data with the activity endpoint “percentage inhibition”. Numbers in pink circles correspond to the number of associated compounds for a respective scaffold cluster.

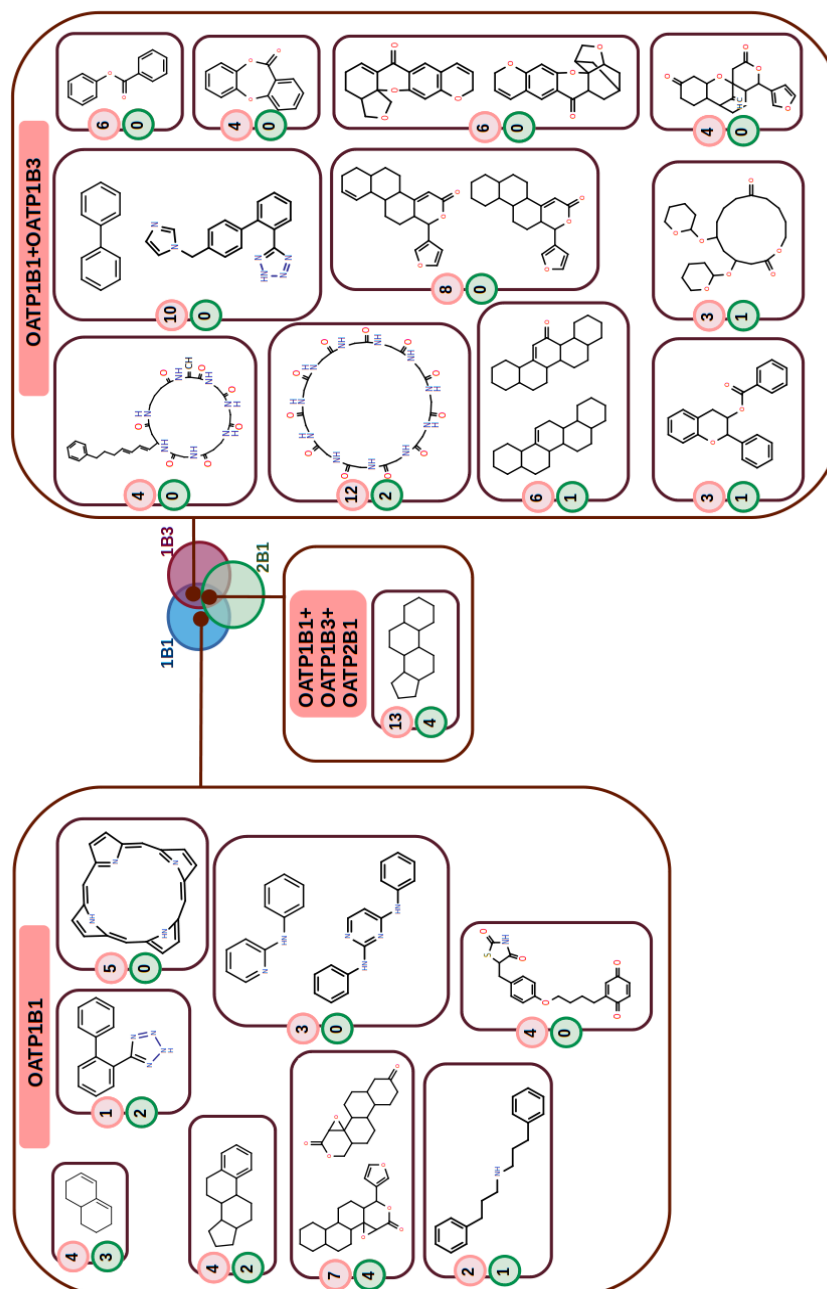


Figure S9. Enriched scaffolds (p-value < 0.1) for hepatic OATP inhibitors. Numbers in pink circles correspond to the number of associated compounds with incomplete pharmacological profile, whereas numbers in green circles correspond to the number of associated compounds with complete pharmacological profile (i.e., compounds from “dense” dataset).

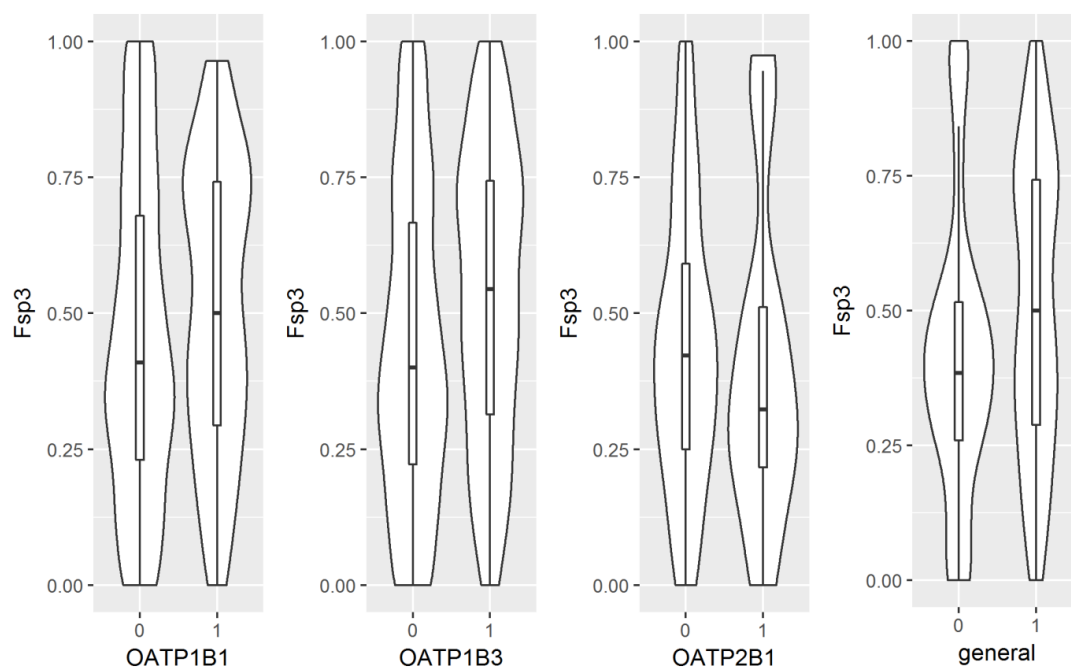


Figure S10. Violin- and boxplots showing the distribution of values for the feature "FractionCSP3" (Fsp3) for inhibitors vs non-inhibitors within four different data sets. Labelling on abscissae: 0....inactives; 1....actives.

Description of Supplementary Data Files:

Data File S1. Csv-file with sparse substrate data set (102 compounds): including CHEMBL_ID (if available), molecule name (if available), InChIKey, canonical SMILES, binary annotations ("1"....active; "0"....inactive; "?"....missing value), and median [-log[activity]] values (if available).

Data File S2. Csv-file with sparse inhibitor data set (1630 compounds): including CHEMBL_ID (if available), molecule name (if available), InChIKey, canonical SMILES, binary annotations ("1"....active; "0"....inactive; "?"....missing value), and median [-log[activity]] values (if available).

Supplementary Information 2

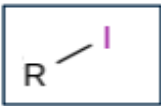
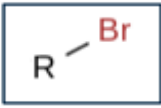

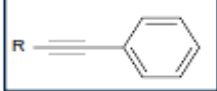
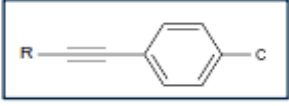
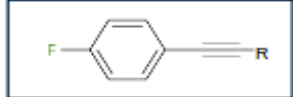
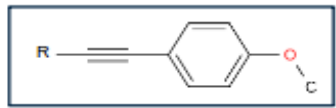
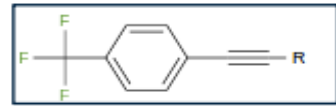
This section includes the supplementary information for Study 2: Réka Laczkó-Rigó, Rebeka Jójárt, Erzsébet Mernyák, Éva Bakos, Alzbeta Tuerkova, Barbara Zdrazil, Csilla Özvegy-Laczka. Structural dissection of 13-epiestrones based on the interaction with human Organic anion-transporting polypeptide, OATP2B1. The Journal of Steroid Biochemistry and Molecular Biology, 2020.

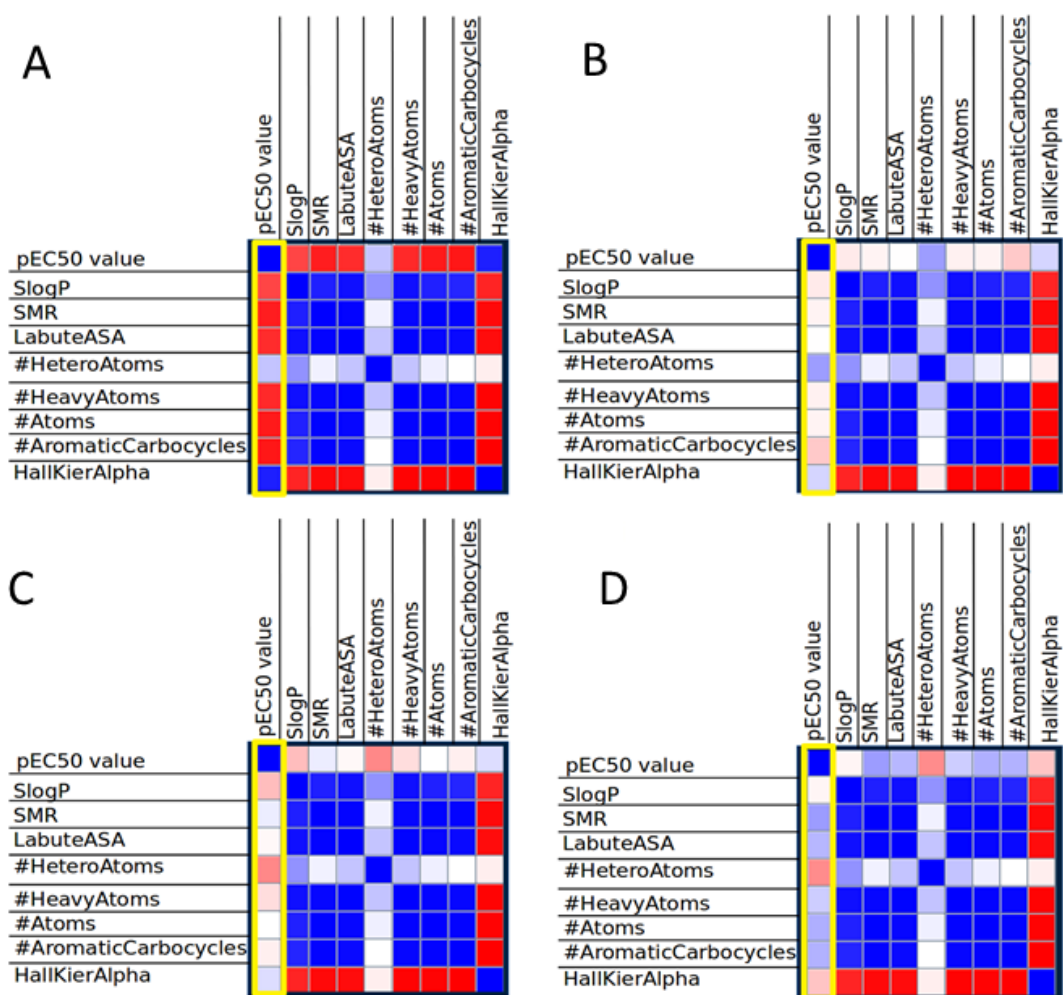
Supplementary Table and Figure for

**Structural dissection of 13-epiestrones based on the interaction with human
Organic anion-transporting polypeptide, OATP2B1**

Réka Laczkó-Rigó, Rebeka Jójárt, Erzsébet Mernyák, Éva Bakos, Alžběta Türková, Barbara
Zdrazil, Csilla Özvegy-Laczka

Supplementary Table 4. HallKier α values for respective substituents at position C-2.

Substituent	Alpha
	0.73
	0.48
	0.29
	-1.22
	-1.22
	-1.29
	-1.42
	-1.43



Supplementary Figure S1: Heat map displaying correlation coefficients of different physico-chemical descriptors with pEC₅₀ values of substituents at position C-2 (A, C) and C-4 (B, D). Heat map displaying correlation coefficients of different physico-chemical descriptors with pEC₅₀ values of substituents at position C-2 (A, C) and C-4 (B, D). Note that 3OH- (A, B) and 3OMe- (C, D) derivatives were treated separately. In addition, the intercorrelation of the individual descriptors is given. The more positive the correlation the more darkness in blue cells, the more negative the more darkness in red cells. White color corresponds to no correlation.

Supplementary Information 3

This section includes the supplementary information for Study 3: Alzbeta Tuerkova and Barbara Zdrazil. A ligand-based computational drug repurposing pipeline Using KNIME and programmatic data access: Case studies for rare diseases and COVID-19. Journal of Cheminformatics, 2020.

A Ligand-based Computational Drug Repurposing Pipeline using KNIME and Programmatic Data Access: Cases Studies for Rare Diseases and COVID-19

Alžbeta Tuerkova and Barbara Zdrazil**

University of Vienna, Department of Pharmaceutical Chemistry, Division of Drug Design and Medicinal Chemistry, Althanstraße 14, A-1090 Vienna, Austria.

Table of Contents

Supplementary Table S1	S2
Supplementary Table S2	S4
Supplementary Table S3	S7
Supplementary Table S4	S9
Supplementary Figure S1.....	S12
Description of Supplementary Data Files.....	S13

Supplementary Table S1: Protein targets with potential interest for treatment of COVID-19 (available from https://covid-19.uniprot.org/uniprotkb?query=*), retrieved August 2020.

UniProt ID	Target Name	Organism	Target Shortcut
O15393	Transmembrane protease serine 2	Homo sapiens	TMPS2_HUMAN
Q92499	ATP-dependent RNA helicase DDX1	Homo sapiens	DDX1_HUMAN
Q9BYF1	Angiotensin-converting enzyme 2	Homo sapiens	ACE2_HUMAN
O43765	Small glutamine-rich tetratricopeptide repeat	Homo sapiens	SGTA_HUMAN
P20701	containing protein alpha	Homo sapiens	ITAL_HUMAN
P35232	Integrin alpha-L	Homo sapiens	PHB_HUMAN
P84022	Prohibitin	Homo sapiens	SMAD3_HUMAN
Q8N3R9	Mothers against decapentaplegic homolog 3	Homo sapiens	MPP5_HUMAN
Q99623	MAGUK p55 subfamily member 5	Homo sapiens	PHB2_HUMAN
P05231	Interleukin-6	Homo sapiens	IL6_HUMAN
P07711	Procathepsin L	Homo sapiens	CATL1_HUMAN
P08887	Interleukin-6 receptor subunit alpha	Homo sapiens	IL6RA_HUMAN
P09958	Furin	Homo sapiens	FURIN_HUMAN
P35613	Basigin	Homo sapiens	BASI_HUMAN
P40189	Interleukin-6 receptor subunit delta	Homo sapiens	IL6RB_HUMAN
P52292	Importin subunit alpha-1	Homo sapiens	IMA1_HUMAN
P62937	Peptidyl-prolyl cis-trans isomerase A	Homo sapiens	PPIA_HUMAN
Q10589	Bone marrow stromal antigen 2	Homo sapiens	BST2_HUMAN
Q16552	Interleukin-17A	Homo sapiens	IL17_HUMAN
Q8NAC3	Interleukin-17 receptor C	Homo sapiens	I17RC_HUMAN
Q8NHX9	Two pore calcium channel protein 2	Homo sapiens	TPC2_HUMAN
Q96F46	Interleukin-17 receptor A	Homo sapiens	I17RA_HUMAN
Q96PD4	Interleukin-17F	Homo sapiens	IL17F_HUMAN
Q9Y2I7	1-phosphatidylinositol 3-phosphate 5-kinase	Homo sapiens	FYV1_HUMAN
Q99623	Prohibitin-2	SARS COV	R1A_CVHSA
P0C6U8	Replicase polyprotein 1a	SARS COV	R1AB_CVHSA

P0C6X7	Replicase polyprotein 1ab	SARS COV-2	R1A_SARS2
P0DTC1	Replicase polyprotein 1a	SARS COV-2	R1AB_SARS2
P0DTD1	Replicase polyprotein 1ab	SARS COV-2	SPIKE_SARS2
P0DTC2	Spike glycoprotein	SARS COV	SPIKE_CVHSA
P59594	Spike glycoprotein	SARS COV	NCAP_CVHSA
P59595	Nucleoprotein	SARS COV	AP3A_CVHSA
P59632	Protein 3a	SARS COV	NS7A_CVHSA
P59635	Protein 7a	SARS COV	VEMP_CVHSA
P59637	Envelope small membrane protein	SARS COV	VME1_CVHSA
P59596	Membrane protein	SARS COV	NS3B_CVHSA
P59633	Non-structural protein 3b	SARS COV-2	AP3A_SARS2
P0DTC3	Protein 3a	SARS COV-2	VME1_SARS2
P0DTC5	Membrane protein	SARS COV-2	NS7A_SARS2
P0DTC7	Protein 7a	SARS COV-2	NCAP_SARS2
P0DTC9	Nucleoprotein	SARS COV	NS6_CVHSA
P59634	Non-structural protein 6	SARS COV	ORF9B_CVHSA
P59636	Protein 9b	SARS COV-2	VEMP_SARS2
P0DTC4	Envelope small membrane protein	SARS COV-2	NS6_SARS2
P0DTC6	Non-structural protein 6	SARS COV-2	ORF9B_SARS2
P0DTD2	Protein 9b	SARS COV	NS7B_CVHSA
Q7TFA1	Protein non-structural 7b	SARS COV	NS8B_CVHSA
Q80H93	Non-structural protein 8b	SARS COV-2	NS8_SARS2
P0DTC8	Non-structural protein 8	SARS COV-2	Y14_SARS2
P0DTD3	Uncharacterized protein 14	SARS COV-2	NS7B_SARS2
P0DTD8	Protein non-structural 7b	SARS COV	NS8A_CVHSA
Q7TFA0	Protein non-structural 8a	SARS COV	Y14_CVHSA
Q7TLC7	Uncharacterized protein 14	SARS COV-2	A0A663DJA2_SARS2
A0A663DJA2	ORF10 protein		

Supplementary Table S2: Protein targets with potential interest for treatment of COVID-19 retrieved from the Open Targets Platform; retrieved in September 2020.

Uniprot ID	Target Name	Organism	Target Shortcut
P34903	gamma-aminobutyric acid type A receptor subunit alpha3	Homo Sapiens	GABRA3_HUMAN
P14867	gamma-aminobutyric acid type A receptor subunit alpha1	Homo Sapiens	GABRA1_HUMAN
P35354	prostaglandin-endoperoxide synthase 2	Homo Sapiens	PTGS2_HUMAN
O95069	potassium two pore domain channel subfamily K member 2	Homo Sapiens	KCNK2_HUMAN
O00591	gamma-aminobutyric acid type A receptor subunit pi	Homo Sapiens	GABRP_HUMAN
P23219	prostaglandin-endoperoxide synthase 1	Homo Sapiens	PTGS1_HUMAN
P62877	ring-box 1	Homo Sapiens	RBX1_HUMAN
P57789	potassium two pore domain channel subfamily K member 10	Homo Sapiens	KCNK10_HUMAN
Q9H4B7	tubulin beta 1 class VI	Homo Sapiens	TUBB1_HUMAN
P78334	gamma-aminobutyric acid type A receptor subunit epsilon	Homo Sapiens	GABRE_HUMAN
P04350	tubulin beta 4A class IVa	Homo Sapiens	TUBB4A_HUMAN
P48169	gamma-aminobutyric acid type A receptor subunit alpha4	Homo Sapiens	GABRA4_HUMAN
P18507	gamma-aminobutyric acid type A receptor subunit gamma2	Homo Sapiens	GABRG2_HUMAN
P04150	nuclear receptor subfamily 3 group C member 1	Homo Sapiens	NR3C1_HUMAN
Q96SW2	cereblon	Homo Sapiens	CRBN_HUMAN
P14778	interleukin 1 receptor type 1	Homo Sapiens	IL1R1_HUMAN
P01008	serpin family C member 1	Homo Sapiens	SERPINC1_HUMA

N

P22894	matrix metalloproteinase 8	Homo Sapiens	MMP8_HUMAN
Q13885	tubulin beta 2A class IIa	Homo Sapiens	TUBB2A_HUMAN
Q9BVA1	tubulin beta 2B class IIb	Homo Sapiens	TUBB2B_HUMAN
P09237	matrix metalloproteinase 7	Homo Sapiens	MMP7_HUMAN
P45452	matrix metalloproteinase 13	Homo Sapiens	MMP13_HUMAN
Q13619	cullin 4A	Homo Sapiens	CUL4A_HUMAN
P17181	interferon alpha and beta receptor subunit 1	Homo Sapiens	IFNAR1_HUMAN
P30556	angiotensin II receptor type 1	Homo Sapiens	AGTR1_HUMAN
Q16445	gamma-aminobutyric acid type A receptor subunit alpha6	Homo Sapiens	GABRA6_HUMAN
P47870	gamma-aminobutyric acid type A receptor subunit beta2	Homo Sapiens	GABRB2_HUMAN
P23415	glycine receptor alpha 1	Homo Sapiens	GLRA1_HUMAN
P08913	adrenoceptor alpha 2A	Homo Sapiens	ADRA2A_HUMAN
P47869	gamma-aminobutyric acid type A receptor subunit alpha2	Homo Sapiens	GABRA2_HUMAN
P48551	interferon alpha and beta receptor subunit 2	Homo Sapiens	IFNAR2_HUMAN
P08887	interleukin 6 receptor	Homo Sapiens	IL6R_HUMAN
Q8N1C3	gamma-aminobutyric acid type A receptor subunit gamma1	Homo Sapiens	GABRG1_HUMAN
P18505	gamma-aminobutyric acid type A receptor subunit beta1	Homo Sapiens	GABRB1_HUMAN
P28472	gamma-aminobutyric acid type A receptor subunit beta3	Homo Sapiens	GABRB3_HUMAN
Q16531	damage specific DNA binding protein 1	Homo Sapiens	DDB1_HUMAN

Q9NPC2	potassium two pore domain channel subfamily K member 9	Homo Sapiens	KCNK9_HUMAN
P06213	insulin receptor	Homo Sapiens	INSR_HUMAN
O14649	potassium two pore domain channel subfamily K member 3	Homo Sapiens	KCNK3_HUMAN
Q9BUF5	tubulin beta 6 class V	Homo Sapiens	TUBB6_HUMAN
Q99928	gamma-aminobutyric acid type A receptor subunit gamma3	Homo Sapiens	GABRG3_HUMAN
P18825	adrenoceptor alpha 2C	Homo Sapiens	ADRA2C_HUMAN
P31644	gamma-aminobutyric acid type A receptor subunit alpha5	Homo Sapiens	GABRA5_HUMAN
Q7Z418	potassium two pore domain channel subfamily K member 18	Homo Sapiens	KCNK18_HUMAN
O14764	gamma-aminobutyric acid type A receptor subunit delta	Homo Sapiens	GABRD_HUMAN
P68371	tubulin beta 4B class IVb	Homo Sapiens	TUBB4B_HUMAN
P07437	tubulin beta class I	Homo Sapiens	TUBB_HUMAN
P03956	matrix metalloproteinase 1	Homo Sapiens	MMP1_HUMAN
Q9NYK1	toll like receptor 7	Homo Sapiens	TLR7_HUMAN
P27487	dipeptidyl peptidase 4	Homo Sapiens	DPP4_HUMAN
P15509	colony stimulating factor 2 receptor subunit alpha	Homo Sapiens	CSF2RA_HUMAN
Q9NR96	toll like receptor 9	Homo Sapiens	TLR9_HUMAN
Q13509	tubulin beta 3 class III	Homo Sapiens	TUBB3_HUMAN
Q3ZCM7	tubulin beta 8 class VIII	Homo Sapiens	TUBB8_HUMAN
P18089	adrenoceptor alpha 2B	Homo Sapiens	ADRA2B_HUMAN

Supplementary Table S3: Number of unique ligands gathered from PDB, ChEMBL, PubChem, and IUPHAR for COVID-19 targets from UniProt pre-release web page.

Target shortcut	PDB	ChEMBL	IUPHAR	PubChem	# Unique active compounds
PPIA_HUMAN	57	2	1	3123	3183
CATL1_HUMAN	25	38	4	946	1003
ITAL_HUMAN	13	94	2	550	564
FURIN_HUMAN	4	10	1	448	463
R1AB_CVHSA	37	187	0	47	227
ACE2_HUMAN	4	65	3	161	172
R1A_CVHSA	35	92	0	79	141
SMAD3_HUMAN	3	64	0	65	71
IL6_HUMAN	3	0	0	13	16
DDX1_HUMAN	0	7	0	7	14
R1AB_SARS2	14	0	0	0	9
TMPS2_HUMAN	2	3	4	3	7
IL17_HUMAN	7	0	0	0	7
SPIKE_SARS2	5	0	0	0	5
SPIKE_CVHSA	5	0	0	0	5
BST2_HUMAN	5	0	0	0	5
IL6RB_HUMAN	4	0	0	0	4

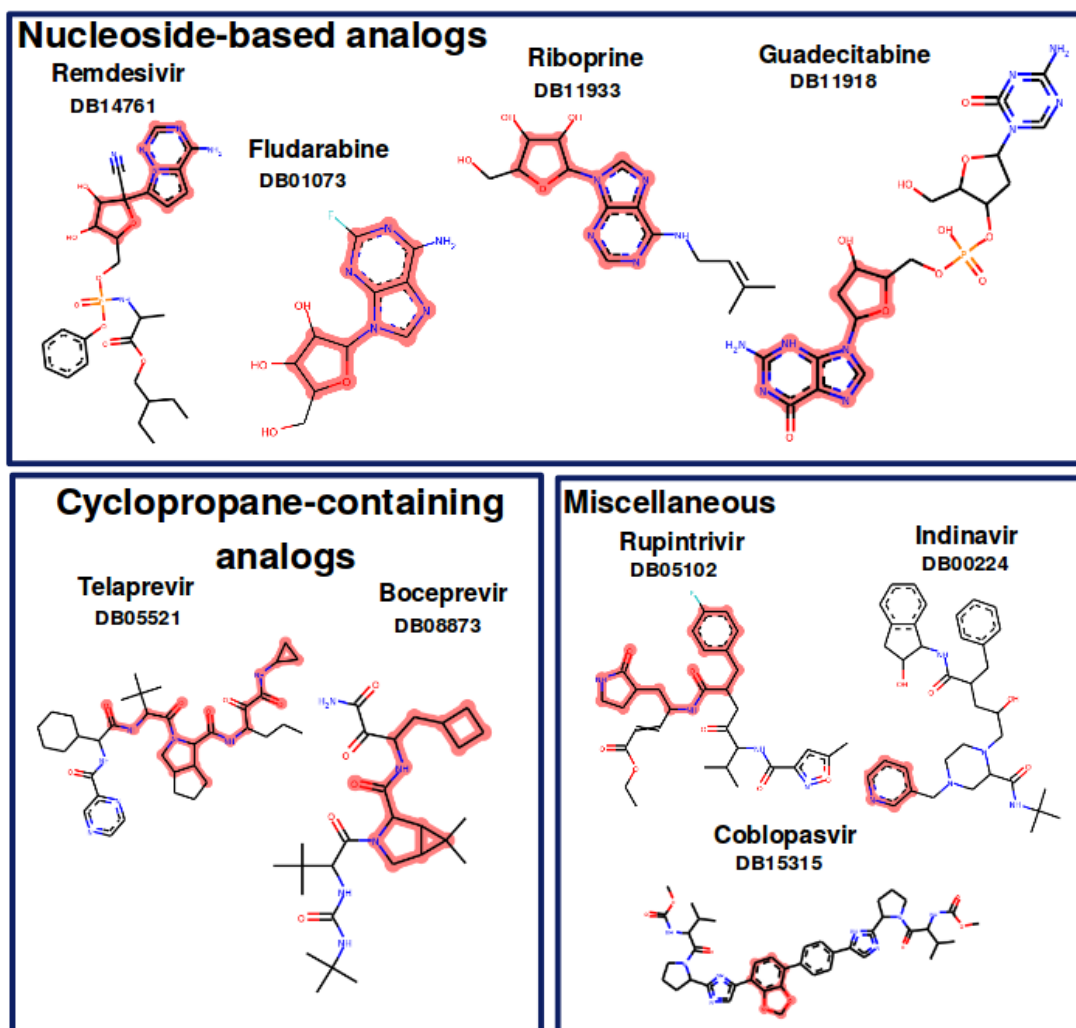
BASI_HUMAN	3	0	0	0	3
IMA1_HUMAN	3	0	0	0	3
IL17F_HUMAN	3	0	0	0	3
SGTA_HUMAN	2	0	0	0	2
R1A_SARS2	2	0	0	0	2
VME1_CVHSA	2	0	0	0	2
IL6RA_HUMAN	2	0	0	0	2
I17RC_HUMAN	1	0	1	0	2
I17RA_HUMAN	2	0	0	0	2
MPP5_HUMAN	1	0	0	0	1
ORF9B_CVHSA	1	0	0	0	1
I17RC_HUMAN	1	0	0	0	1
FYV1_HUMAN	0	0	1	0	1

Supplementary Table S4: Number of unique ligands gathered from PDB, ChEMBL, PubChem, and IUPHAR for COVID-19 targets from the Open Targets Platform.

Target shortcut	ChEMBL	IUPHAR	PubChem	PDB	# Unique active compounds
GABRG2_HUMAN	152	0	677	2	831
MMP13_HUMAN	319	3	80	28	430
GABRB1_HUMAN	121	0	166	0	287
DPP4_HUMAN	140	2	29	14	185
GABRA1_HUMAN	3	4	172	2	181
ADRA2C_HUMAN	60	19	99	0	178
AGTR1_HUMAN	111	14	34	1	160
MMP1_HUMAN	57	1	65	9	132
MMP8_HUMAN	81	2	12	18	113
TUBB2A_HUMAN	53	0	56	0	109
GABRG1_HUMAN	49	0	52	0	101
GABRP_HUMAN	46	0	54	0	100
NR3C1_HUMAN	79	8	0	1	88
GABRE_HUMAN	38	0	46	0	84
ADRA2A_HUMAN	47	13	13	0	73
GABRG3_HUMAN	33	0	35	0	68
GABRB2_HUMAN	0	0	65	1	66
GABRD_HUMAN	0	0	48	0	48

TUBB2B_HUMAN	20	0	20	3	43
ADRA2B_HUMAN	11	14	16	0	41
GABRA5_HUMAN	33	4	0	1	38
TUBB_HUMAN	0	0	33	3	36
TUBB6_HUMAN	15	0	15	0	30
PTGS2_HUMAN	23	0	4	2	29
INSR_HUMAN	4	3	10	9	26
TUBB4A_HUMAN	0	0	23	0	23
TUBB3_HUMAN	0	0	18	4	22
GLRA1_HUMAN	4	7	2	1	14
MMP7_HUMAN	4	2	0	8	14
CRBN_HUMAN	1	7	2	1	11
KCNK9_HUMAN	1	0	5	4	10
GABRA2_HUMAN	4	4	0	0	8
TUBB1_HUMAN	0	0	8	0	8
PTGS1_HUMAN	1	5	1	0	7
TUBB4B_HUMAN	0	0	7	0	7
KCNK2_HUMAN	0	2	0	3	5
SERPINC1_HUMAN	0	2	0	3	5
TLR7_HUMAN	3	1	1	0	5
KCNK10_HUMAN	0	1	0	2	3

KCNK3_HUMAN	0	2	0	1	3
DDB1_HUMAN	0	0	0	2	2
GABRB3_HUMAN	0	0	0	2	2
IL1R1_HUMAN	0	0	0	2	2
IL6R_HUMAN	0	0	0	2	2
CSF2RA_HUMAN	0	0	0	1	1
CUL4A_HUMAN	0	0	0	1	1
GABRA4_HUMAN	1	0	0	0	1
IFNAR1_HUMAN	0	0	0	1	1
IFNAR2_HUMAN	0	0	0	1	1
RBX1_HUMAN	0	0	0	1	1



Supplementary Figure 1: Examples of identified drugs with structural queries highlighted..

Description of Supplementary Data Files:

Supplementary File S1: CSV file with maximum common substructures in SMARTS format (n= 257) detected via hierarchical scaffold clustering for the COVID-19 use case (available at https://github.com/AlzbetaTuerkova/Drug-Repurposing-in-KNIME/blob/master/Supplementary_file1.csv)

Supplementary File S2: CSV file with identified hits for COVID-19 from DrugBank (n=7,836; available at https://github.com/AlzbetaTuerkova/Drug-Repurposing-in-KNIME/blob/master/Supplementary_file2.csv)

Supplementary File S3: CSV file with identified hits for COVID-19 from CAS Dataset (n=36,521; available at https://github.com/AlzbetaTuerkova/Drug-Repurposing-in-KNIME/blob/master/Supplementary_file3.csv)

Supplementary File S4: CSV file with identified for COVID-19 hits by both DrugBank and CAS Dataset (n=228; available at https://github.com/AlzbetaTuerkova/Drug-Repurposing-in-KNIME/blob/master/Supplementary_file4.csv)

Supplementary File S5: CSV file with identified hits for GLUT-1 deficiency syndrome from DrugBank (n=539; available at https://github.com/AlzbetaTuerkova/Drug-Repurposing-in-KNIME/blob/master/Supplementary_file5.csv)

Supplementary File S6: KNIME drug repurposing workflow (KNWF file) where external UniProt dataset is used as an input(available at https://github.com/AlzbetaTuerkova/Drug-Repurposing-in-KNIME/blob/master/DrugRepurposingPipeline_UniProt.knwf)

Supplementary File S7: KNIME drug repurposing workflow (KNWF file) where disease-target associations from the OpenTarget platform are used an input (available at https://github.com/AlzbetaTuerkova/Drug-Repurposing-in-KNIME/blob/master/DrugRepurposingPipeline_OpenTargets.knwf)

Supplementary File S8: A .pdf Tutorial file “Part 1: Programmatic access to UniProt database using KNIME” (available at <https://github.com/AlzbetaTuerkova/Drug-Repurposing-in-KNIME/blob/master/Part1.pdf>)

Supplementary File S9: A .pdf Tutorial file “Part 2: Using cross-references to retrieve structural data from the Protein Data Bank (PDB)” (available at <https://github.com/AlzbetaTuerkova/Drug-Repurposing-in-KNIME/blob/master/Part2.pdf>)

Supplementary File S10: A .pdf Tutorial file “Part 3: Integrative data mining of ligand bioactivity data from ChEMBL and PubChem” (available at <https://github.com/AlzbetaTuerkova/Drug-Repurposing-in-KNIME/blob/master/Part3.pdf>)

Supplementary File S10: A .pdf Tutorial file “Part 4: Substructure searches in DrugBank” (available at <https://github.com/AlzbetaTuerkova/Drug-Repurposing-in-KNIME/blob/master/Part4.pdf>)

Supplementary File S11 A .pdf Tutorial file with the answer sheet (available at <https://github.com/AlzbetaTuerkova/Drug-Repurposing-in-KNIME/blob/master/Answersheet.pdf>)

Supplementary Information 4

This section includes the supplementary information for Study 4: Marleen J. Meyer, Alzbeta Tuerkova, Sarah Römer, Christoph Wenzel, Tina Seitz, Jochen Gaedcke, Stefan Oswald, Jürgen Brockmöller, Barbara Zdrazil, Mladen V. Tzvetkov. Differences in metformin and thiamine uptake between human and mouse organic cation transporter OCT1: structural determinants and potential consequences for intrahepatic concentrations. *Drug Metabolism and Disposition*, 2020.

Supplemental Material

Differences in metformin and thiamine uptake between human and mouse organic cation transporter OCT1: structural determinants and potential consequences for intrahepatic concentrations

Marleen J. Meyer, Alzbeta Tuerkova, Sarah Römer, Christoph Wenzel, Tina Seitz, Jochen Gaedcke,
Stefan Oswald, Jürgen Brockmöller, Barbara Zdrazil, Mladen V. Tzvetkov

Affiliations

Institute of Pharmacology, Center of Drug Absorption and Transport (C_DAT), University Medicine
Greifswald, Greifswald, Germany (MJM, SR, CW, SO, MVT)

Department of Pharmaceutical Chemistry, Division of Drug Design and Medicinal Chemistry,
University of Vienna, Vienna, Austria (AT, BZ)

Department of General, Visceral, and Pediatric Surgery, University Medical Center Göttingen,
Göttingen, Germany (JG)

Institute of Clinical Pharmacology, University Medical Center Göttingen, Göttingen, Germany (TS,
JB)

Table of contents

Supplementary Table S1.....	3
Supplementary Table S2.....	4
Supplementary Table S3.....	5
Supplementary Table S4.....	6
Supplementary Figure S1	7
Supplementary Figure S2	8
Supplementary Figure S3	9
Supplementary Figure S4	10
Supplementary Figure S5	11

Supplementary Tables

Supplementary Table S1. Primers used for the generation of OCT1 constructs

Primer	Sequence (5' → 3')	TMH/amino acid substitution
m/rOCT1_F	GTTGGTGAGGAAGCTTACCCAGCCATGCCCACCGTGA	
human1_for	CGTTTAAACTTAAGCTTGCATGCT	TMH1
hm_1_rev	CCAGGACTCTGGCAGTGGTGGTC	
hm_EC_for	GACCACCACTGCCAGAGTCCTGG	
hm_EC_rev	GGACTGAAAAGAGGTCCAGCTTCCAGG	EC-loop
hm_2_for	CCTGGAAGCTGGACCTCTTTCAGTCC	
hm_2_rev	AGCTTACGGCCAAACCTGTCTGCAA	TMH2
hm_3_for	TTGCAGACAGGTTTGGCCGTAAGCT	
hm_3_rev	TGCAGCAGGCGGAAGAGCAGCATGGA	TMH3
hm_4_for	TCCATGCTGCTCTTCCGCTGCTGCA	
hm_4_rev	CGTTCTTCTGGAGCCCGAGCC	TMH4
hm_5_for	GGCTCGGGCTCCAGAAGAACG	
hm_5_rev	GCCAGCTGCAGCCAGCGCCAGT	TMH5
hm_6_for	ACTGGCGCTGGCTGCAGCTGGC	
hm_6_rev	AACAGCCACCGAGGGGA	TMH6
hm_IC_for	TCCCCTCGGTGGCTGTT	
hm_IC_rev	CAGGATGAAGGTGCGCTTCCTCAG	IC-loop
L155V_hfor	GACCTCTTTCAGTCCTGT G TGAATGCGGGCTTCTTCTTT	Leu155Val in hOCT1
L155V_hrev	AAAGAAGAAGCCCGCATT CAC ACAGGACTGAAAGAGGTC	
V156L_mfor	GACCTTTTTCAGTCCTGT TTG AACTTGGGCTTCTTCCTG	Val156Leu in mOCT1
V156L_mrev	CAGGAAGAAGCCCAAGTT CAA ACAGGACTGAAAAAGGTC	
G165V_hfor	TTCTTCTTTGGCTCTCTC TT GTTGGCTACTTTGCAGAC	
G165V_hrev	GTCTGCAAAGTAGCCAAC AA CGAGAGAGCCAAAGAAGAA	Gly165Val in hOCT1
G181V_hfor	CGTAAGCTGTGTCTCCTG GTG ACTGTGCTGGTCAACGCG	
G181V_hrev	CGCGTTGACCAGCACAGT CACC AGGAGACACAGCTTACG	Gly181Val in hOCT1
V166G_mfor	TTCTTCCTGGGCTCCCTG GGT GTGGGTACATTGCAGAC	
V166G_mfor	GTCTGCAATGTAACCCAC ACC CAGGGAGCCAGGAAGAA	Val166Gly in mOCT1
V182G_mfor	CGTAAGCTCTGCCTCCTG GG AACCACTCTGGTCACCTC	
V182G_mrev	GAGGTGACCAGAGTGGT TCCC AGGAGGCAGAGCTTACG	Val182Gly in mOCT1

Affected codons are underlined, changed bases are boldfaced; TMH, transmembrane helix; EC-loop, large extracellular loop; IC-loop, large intracellular loop; hOCT1, human OCT1; mOCT1, mouse OCT1

Supplementary Table S2. Antibodies and dilutions used for immunocytochemical staining of OCT1-overexpressing cells

	Antibody	Source	Dilution	Applied to cell line			
				hOCT1	hnh	mmh	mOCT1
Primary antibody	Monoclonal mouse anti-hOCT1 2C5 (NBP1-51684)	Novus Biologicals, Abingdon, UK	1:400	X	X	-	-
	Monoclonal rabbit anti-hOCT1 (Orb32029)	Biorbyt Ltd, Cambridge, UK	1:50	X	-	X	-
	Polyclonal rabbit anti-mOCT1	Hermann Koepsell, University of Würzburg (Meyer-Wentrup et al., 1998)	1:50	-	-	-	X
	Monoclonal rabbit anti-Na ⁺ /K ⁺ -ATPase (EP1845Y)	Abcam, Cambridge, UK	1:200	X	X	-	-
	Monoclonal mouse anti-Na ⁺ /K ⁺ -ATPase (sc-48345)	Santa Cruz Biotechnology, Heidelberg, Germany	1:100	X	-	X	X
Secondary antibody	Alexa Fluor® 488 goat anti - mouse IgG (H+L), polyclonal	Thermo Fisher Scientific, Darmstadt, Germany	1:400	X	X	-	-
			1:200	X	-	X	X
	Alexa Fluor® 546 goat anti - rabbit IgG (H+L), polyclonal	Thermo Fisher Scientific, Darmstadt, Germany	1:400	X	X	-	-
			1:100	X	-	X	X

Supplementary Table S3. Proteospecific peptides for quantitative LC-MS/MS analyses of OCT1 protein expression

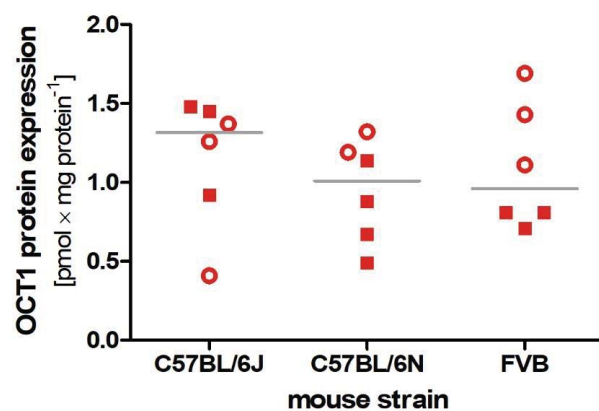
Protein	Peptide	Q1	Q3.1	Q3.2	Q3.3
Human OCT1	ENTIYLK	441	423.5	536.5	-
	ENTIYLK*	445.2	431.5	544.6	-
Human OCT1	LPPADLK	377.3	543.5	211.2	-
	LPPADLK*	381.4	551.5	211.2	-
Human OCT1	DAENLGR	388.1	459.3	588.4	-
	DAENLGR*	393.4	469.3	598.4	-
Mouse OCT1	GVALPETIEEAENLGR	850.1	680	228.2	341.2
	GVALPETIEEAENLGR*	855.2	684.7	228.2	341.3
Na ⁺ /K ⁺ -ATPase	LSLDELHR	328.	435.2	391.7	669.3
	LSLDELHR*	331.5	440.3	396.8	679.4

*stable isotope-labeled internal standard peptide

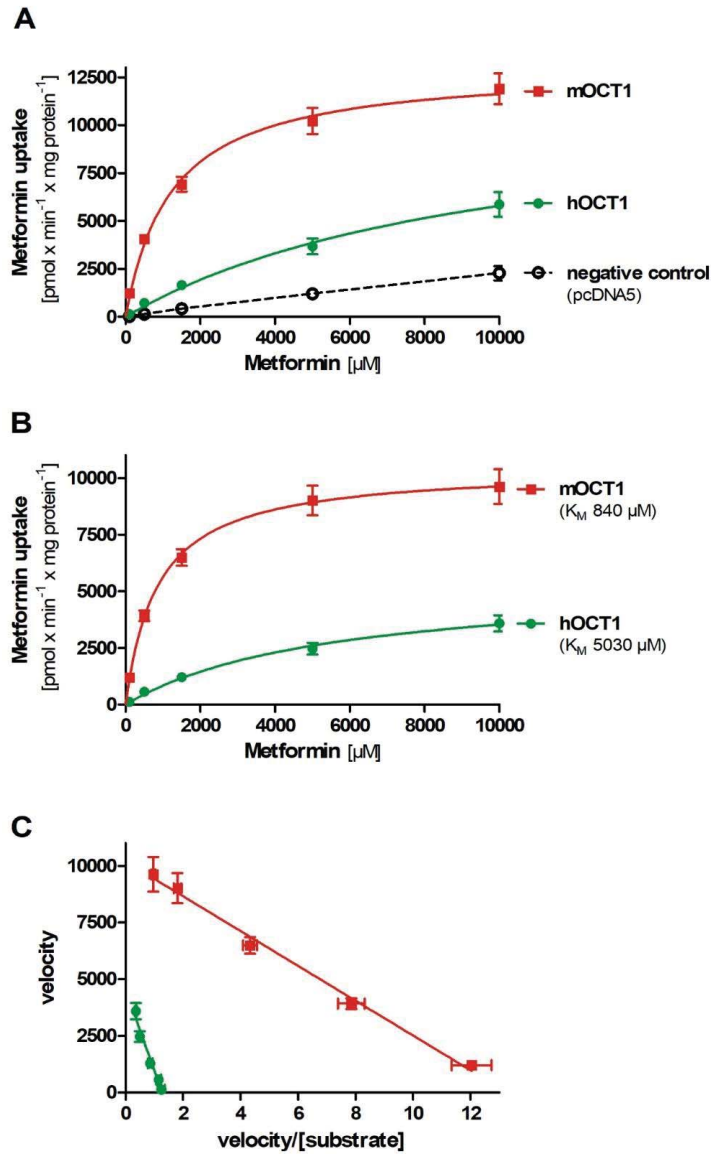
Supplementary Table S4. Comparison of the parameters of transport kinetics between stably and transiently transfected HEK293 cells

Parameter	Mouse					Human				
	Mean	n	SD	95% CI		Mean	n	SD	95% CI	
Metformin (stably transfected cells)										
Affinity for metformin uptake, K_M [μM]	491	11	155	387	595	2197	11	1154	1422	2973
Maximal velocity, v_{max} [$pmol \times min^{-1} \times mg \text{ protein}^{-1}$]	17496	11	7097	12727	22265	14703	11	4346	11783	17623
Maximal velocity, v_{max} [$pmol \times min^{-1} \times pmol \text{ OCT1}^{-1}$]	1353	11	549	985	1722	939	11	278	753	1126
Metformin $CL_{in \text{ vitro}}$ [$\mu L \times min^{-1} \times mg \text{ protein}^{-1}$]	37	11	16.1	26.2	47.9	7.85	11	3.9	5.23	10.5
Metformin (transiently transfected cells)										
Affinity for metformin uptake, K_M [μM]	840	14	225	710	970	5030	14	2101	3817	6243
Maximal velocity, v_{max} [$pmol \times min^{-1} \times mg \text{ protein}^{-1}$]	10187	14	3109	8392	11982	5169	14	2203	3897	6441
Metformin $CL_{in \text{ vitro}}$ [$\mu L \times min^{-1} \times mg \text{ protein}^{-1}$]	12.4	14	2.69	10.8	13.9	1.09	14	0.36	0.88	1.3
Thiamine (stably transfected cells)										
Affinity for thiamine uptake, K_M [μM]	143	5	96.3	22.9	262	1057	5	341	634	1480
Maximal velocity, v_{max} [$pmol \times min^{-1} \times mg \text{ protein}^{-1}$]	3712	5	1031	2432	4992	8261	5	3720	3642	12880
Maximal velocity, v_{max} [$pmol \times min^{-1} \times pmol \text{ OCT1}^{-1}$]	287	5	79.7	188	386	528	5	238	233	823
Thiamine $CL_{in \text{ vitro}}$ [$\mu L \times min^{-1} \times mg \text{ protein}^{-1}$]	34.1	5	18.3	11.3	56.8	7.71	5	1.75	5.54	9.88
Thiamine (transiently transfected cells)										
Affinity for thiamine uptake, K_M [μM]	238	3	103	-19	495	1517	3	491	298	2736
Maximal velocity, v_{max} [$pmol \times min^{-1} \times mg \text{ protein}^{-1}$]	2772	3	823	728	4816	2226	3	404	1222	3230
Thiamine $CL_{in \text{ vitro}}$ [$\mu L \times min^{-1} \times mg \text{ protein}^{-1}$]	12.1	3	1.56	8.20	15.9	1.54	3	0.37	0.62	2.45

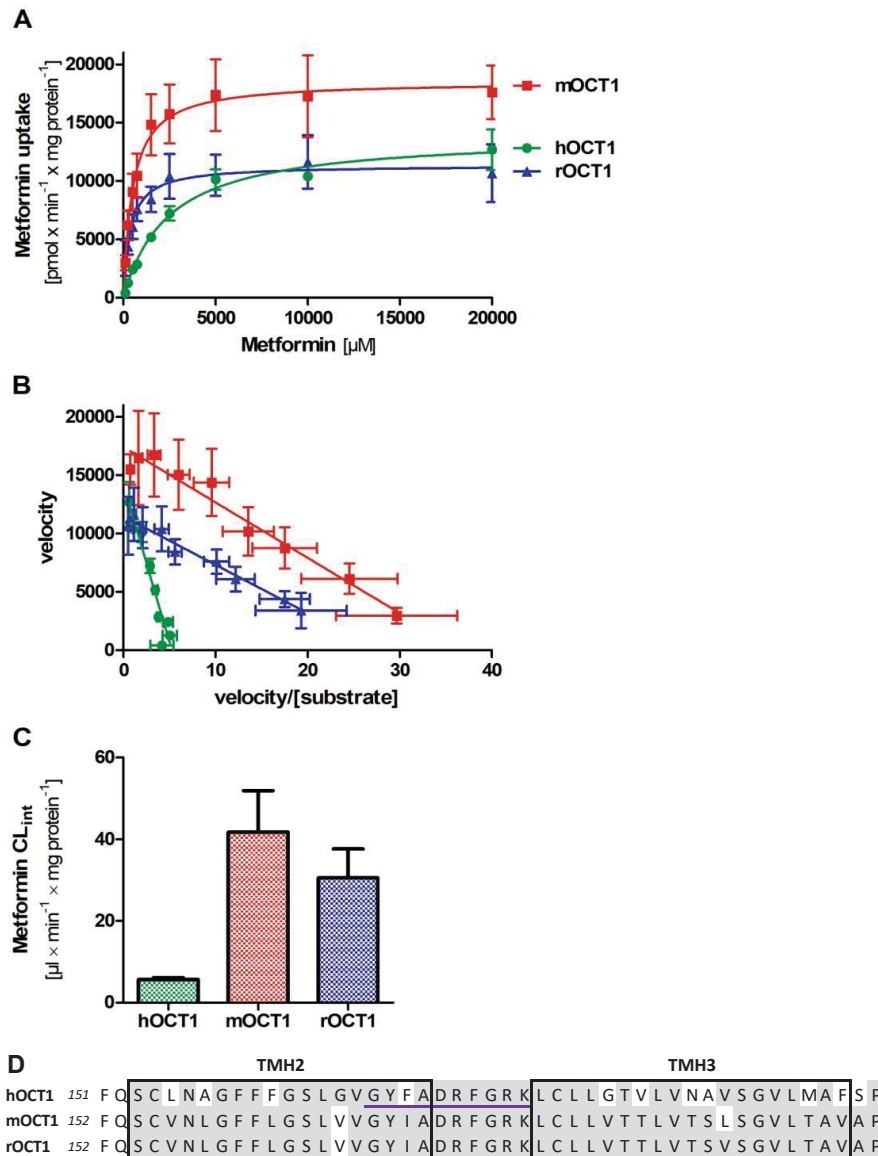
Supplementary Figures



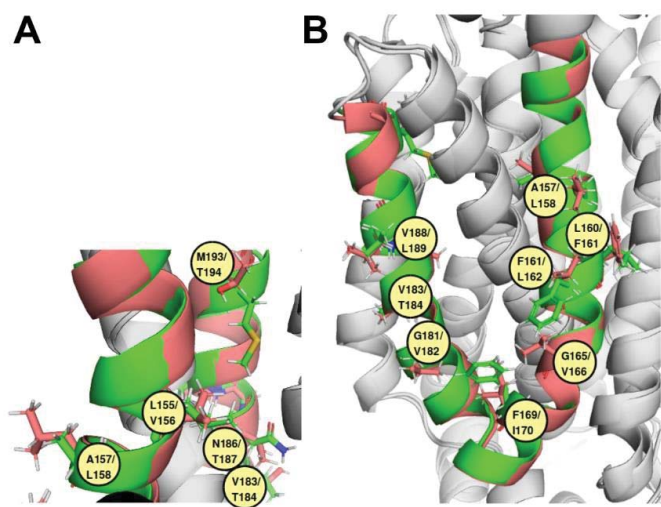
Supplementary Figure S1. Comparison of OCT1 protein expression between different mouse strains OCT1 protein expression in the membrane fraction of mouse liver samples was measured using targeted proteomics. Boxes represent male, open circles represent female animals. Shown are the medians of six animals per mouse strain.



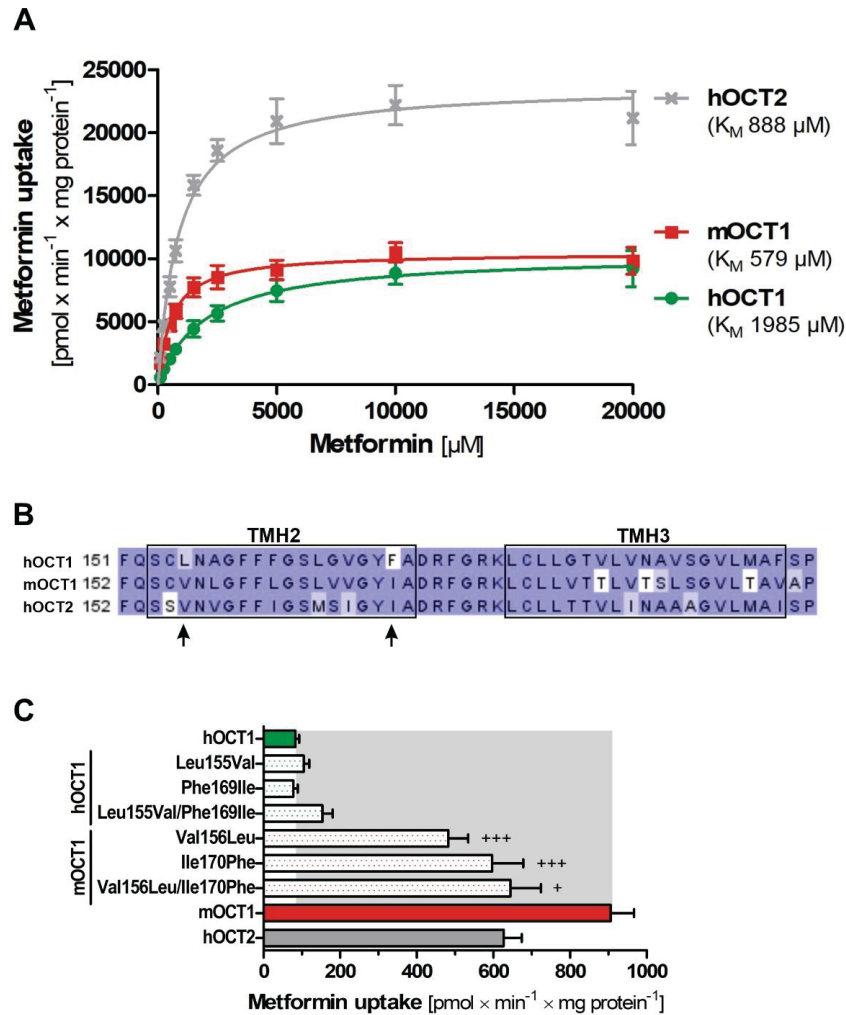
Supplementary Figure S2. Differences in metformin uptake between human and mouse OCT1 using transiently transfected HEK293 cells (A) Concentration-dependent uptake of metformin by human OCT1 (green), mouse OCT1 (red) and the negative control pcDNA5 (empty vector). (B) Concentration-dependent uptake of metformin by human and mouse OCT1. HEK293 cells transiently overexpressing OCT1 were incubated with increasing concentrations of metformin for 2 min. OCT1-mediated uptake was calculated by subtracting the uptake of control cells (pcDNA5) from the uptake of cells overexpressing OCT1. (C) Eadie-Hofstee transformation of the data in B. Shown are means and standard errors of the means of 14 independent experiments.



Supplementary Figure S3. Differences in metformin uptake between human, mouse, and rat OCT1 (A) Concentration-dependent uptake of metformin by human (green), mouse (red), and rat (blue) OCT1. OCT1-overexpressing HEK293 cells were incubated with increasing concentrations of metformin for 2 min. OCT1-mediated uptake was calculated by subtracting the uptake of control cells (pcDNA5) from the uptake of cells overexpressing OCT1. (B) Eadie-Hofstee transformation of the data in A. (C) Comparison of the intrinsic clearance (CL_{int}) between human, mouse, and rat OCT1 calculated using v_{max} and K_M of the data in A. Shown are only measurements of all three OCT1 orthologs in parallel, therefore, data may differ slightly from data shown in Fig. 2A Shown are means and standard errors of the means of four independent experiments. (D) Protein sequence alignment of TMH2 and TMH3 of human, mouse, and rat OCT1 with the conserved A-motif of the MFS underlined in violet. Coloring based on amino acid identity.



Supplementary Figure S4. Structural overview of the TMH2/THM3 region in human and mouse OCT1 *in silico* models Superposition of human and mouse OCT1 structural models with focus on TMH2 and TMH3 and non-conserved residues highlighted and labeled. (A) Inside view and (B) outside view on TMH2 and TMH3.



Supplementary Figure S5. Differences in metformin uptake between human OCT1, mouse OCT1, and human OCT2 (A) Concentration-dependent uptake of metformin by human OCT1 (green), mouse OCT1 (red) and human OCT2 (grey). OCT1 or OCT2-overexpressing HEK293 cells were incubated with increasing concentrations of metformin for 2 min. OCT1 or OCT2-mediated uptake was calculated by subtracting the uptake of control cells (pcDNA5) from the uptake of cells overexpressing human or mouse OCT1 or human OCT2. Shown are only measurements of all three OCT paralogues in parallel, therefore, data may differ slightly from data shown in Fig. 2A. (B) Protein sequence alignment of TMH2 and TMH3 of human OCT1, mouse OCT1, and human OCT2 using the CLUSTAL O(1.2.4) multiple sequence alignment tool (Madeira et al., 2019). Coloring based on the BLOSUM62 matrix. Arrows indicate amino acids mutated in C. (C) Effect of mutations of hLeu155/mVal156 and hPhe169/mIle170 in human/mouse OCT1, respectively, on metformin uptake. HEK293 cells transiently overexpressing OCT1 were incubated with 100 μM metformin for 2 min. Shown are means and standard errors of the means of three independent experiments. +++ $P < 0.001$ compared to mouse OCT1 in a Tukey's post hoc analysis following one-way ANOVA.

Supplementary Information 5

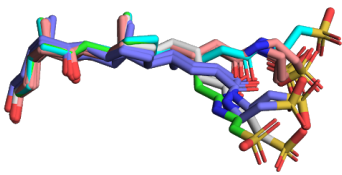
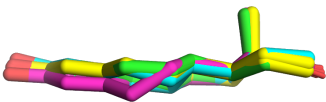
This section includes the supplementary information for Study 5: Alzbeta Tuerkova, Orsolya Ungvári, Erzsébet Mernyák, Gergely Szakács, Csilla Özvegy-Laczka, Barbara Zdrazil. Data-driven Ensemble Docking to Unravel Interactions of Steroid Analogs with Hepatic Organic Anion Transporting Polypeptides, 2020

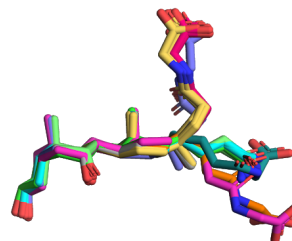
Supporting Information

Data-driven Ensemble Docking to Unravel Interactions of Steroid Analogs with Hepatic Organic Anion Transporting Polypeptides

Supplementary Table S1

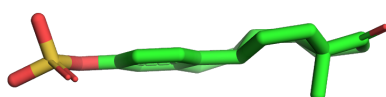
Chemical structures retrieved for steroid analogs from PDB via RESTful web services. Table lists ligand name, stereochemistry type of the steroid nucleus, and visualization of the superimposed structures.

Ligand	Steroid Stereochemistry	Superimposed structures
Taurochenodeoxycholic acid	cis-trans-trans	
(9beta,13alpha)-3-hydroxyestra-1,3,5(10)-trien-17-one	trans-trans-trans	
Glycochenodeoxycholic acid	cis-trans-trans	



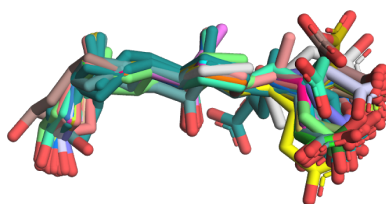
Estrone-3-sulfate

trans-trans-trans



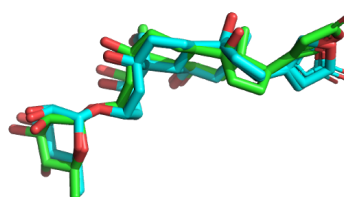
(3alpha,5beta,12alpha)-3,12-dihydroxycholestan-24-oic acid

cis-trans-trans



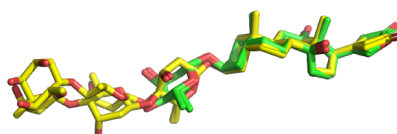
Ouabain

cis-trans-cis



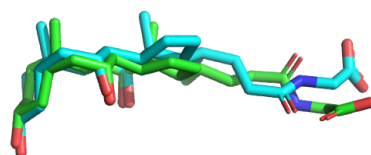
Digoxin

cis-trans-cis



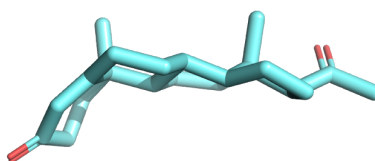
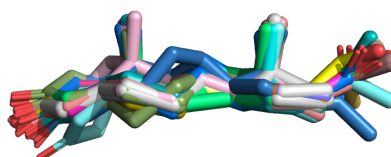
Glycocholic acid

cis-trans-trans



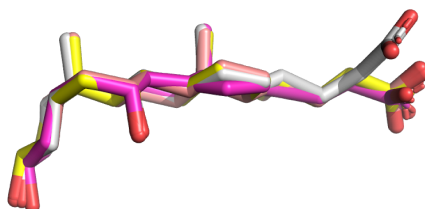
Progesterone

cis/trans-trans-trans



Iso-ursodeoxycholic acid

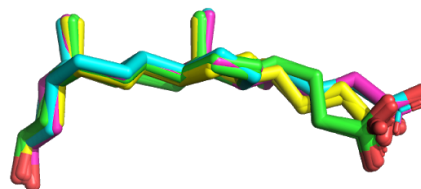
cis-trans-trans



(3beta,5beta,14beta,
17alpha)-3-hydroxy
cholan-24-oic acid

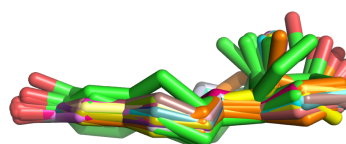
cis-trans-trans





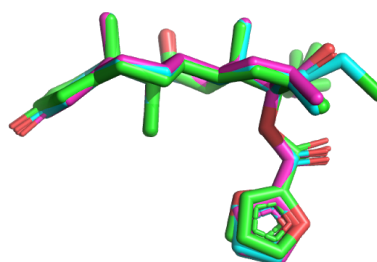
Estradiol

trans-trans-trans



Mometasone furoate

cis-trans-trans



Supplementary Table S2

Crystal structures of MFS transporters identified on the basis of signature dynamics investigations. PDB ID and the corresponding title are listed here. If a certain PDB structure was identified as a suitable template for one of the hepatic OATPs, the prediction score from pGenThreader algorithm is listed in this table.

PDB ID	Title	Score OATP1B1	Score OATP1B3	Score OATP2B1
3o7q	Crystal structure of a Major Facilitator Superfamily (MFS) transporter, FucP, in the outward conformation	75.504	74.551	93.148
4m64	3S crystal structure of na ⁺ /melibiose symporter of salmonella typhimurium	79.538	NA	NA
3wdo	Structure of e. coli YATJ transporter	79.538	69.415	86.484
4gby	The structure of the MFS (Major Facilitator Superfamily) proton:xylose symporter XYLI bound to d-xylose	NA		NA
4gbz	The structure of the MFS (Major Facilitator Superfamily) proton:xylose symporter XYLe bound to d-glucose	NA	NA	NA
4gc0	The structure of the MFS (major facilitator superfamily) proton:xylose symporter XYLe bound to 6-bromo-6-deoxy-d-glucose	57.105	NA	NA
4zwc	Crystal structure of maltose-bound human GLUT3 in the outward-open conformation at 2.6 Angstrom	NA	NA	NA
4zw9	Crystal structure of human GLUT3 bound to d-glucose in the outward-occluded conformation at 1.5 Angstrom	NA	NA	NA
5c65	Structure of the human glucose transporter GLUT3 / slc2a3	NA	NA	NA

4jre	Crystal structure of nitrate/nitrite exchanger nark with nitrite bound	NA	NA	NA
4zwb	Crystal structure of maltose-bound human GLUT3 in the outward-occluded conformation at 2.4 Angstrom	NA	NA	NA
4oaa	Crystal structure of e. coli lactose permease g46w, g262w bound to sugar	NA	NA	NA
4zyr	Crystal structure of e. coli lactose permease g46w/g262w bound to p-nitrophenyl alpha-d-galactopyranoside (alpha-npg)	NA	NA	NA
4gxb	Structure of the snx17 atypical ferm domain bound to the npxy motif of p-selectin	NA	NA	NA
3o7p	Crystal structure of the e.coli fucose:proton symporter, FucP (n162a)	NA	NA	NA
5gxb	Crystal structure of a LacY/nanobody complex	NA	NA	NA

Supplementary Table S3

AUC values for top five models per transporter.

Transporter	AUC
OATP1B1	0.830
OATP1B1	0.821
OATP1B1	0.810
OATP1B1	0.810
OATP1B1	0.810
OATP1B3	0.940
OATP1B3	0.930
OATP1B3	0.930
OATP1B3	0.917

OATP1B3	0.909
OATP2B1	0.694
OATP2B1	0.680
OATP2B1	0.672
OATP2B1	0.672
OATP2B1	0.661

Supplementary Table S4

Cluster analysis for OATP1B1 (total number of actives: 20). Selected clusters are highlighted in bold.

ID	# unique compounds	# poses/compound in a cluster			
		Mean	Median	Min	Max
1	7 (35%)	1.714	1	1	4
2	11 (55%)	1.727	1	1	4
3	10 (50%)	1.3	1	1	3
4	7 (35%)	1.714	1	1	4
5	13 (65%)	1.692	1	1	3
6	17 (85%)	2.000	1	1	5
7	9 (45%)	1.444	1	1	3
8	3 (15%)	1.333	1	1	2
9	4 (20%)	1.0	1	1	1
10	3 (15%)	1.333	1	1	2
11	3 (15%)	1.0	1	1	1
12	3 (15%)	1.333	1	1	2
13	6 (30%)	1.0	1	1	1
14	4 (20%)	1.0	1	1	1

15	5 (25%)	1.4	1	1	2
----	---------	-----	---	---	---

Supplementary Table S5

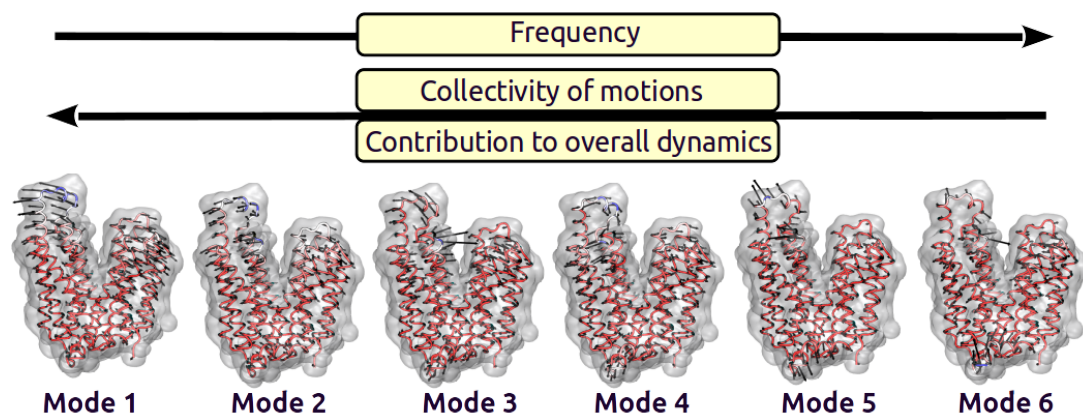
Cluster analysis for OATP1B3 (total number of actives: 12). Selected clusters are highlighted in bold.

ID	# unique compounds	# poses/compound in a cluster			
		Mean	Median	Min	Max
1	12 (100%)	4.333	4	1	7
2	5 (42%)	1.6	2	1	2
3	2 (16%)	1.5	1.5	1	2
4	10 (83%)	1.4	1	1	3
5	2 (16%)	2.5	2.5	2	3
6	5 (42%)	1.6	1	1	3
7	3 (25%)	1.333	1	1	2
8	2 (16%)	2	2	1	3
9	5 (42%)	1	1	1	1

Supplementary Table S6

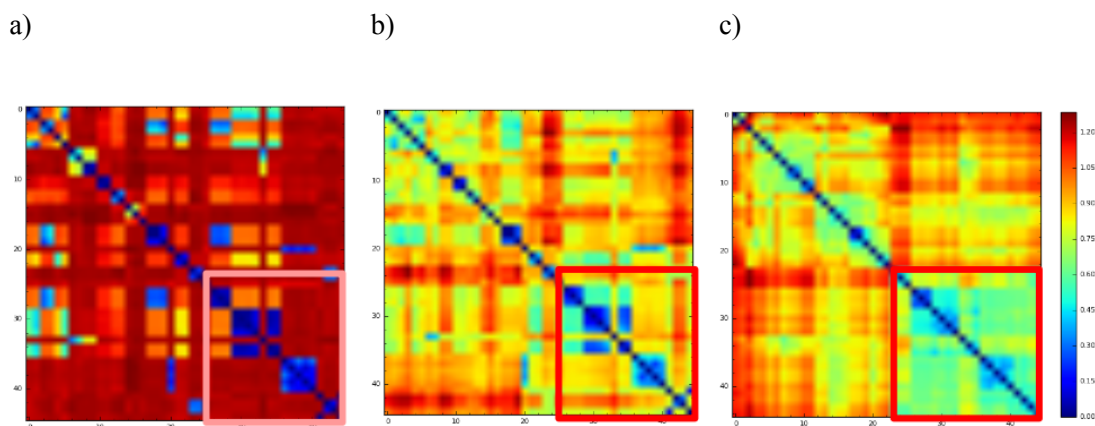
Cluster analysis for OATP2B1 (total number of actives: 16). Selected cluster is highlighted in bold.

ID	# unique compounds	# poses/compound in a cluster			
		Mean	Median	Min	Max
1	15 (94%)	5.066	5	3	8
2	5 (33%)	1.0	1	1	1
3	3 (20%)	1.333	1	1	2
4	5 (33%)	1.800	2	1	3
5	6 (40%)	1.0	1	1	1
6	4 (26%)	1.25	1	1	2
7	3 (20%)	1.666	2	1	2
8	4 (26%)	1.25	1	1	2
9	9 (60%)	1.666	2	1	2



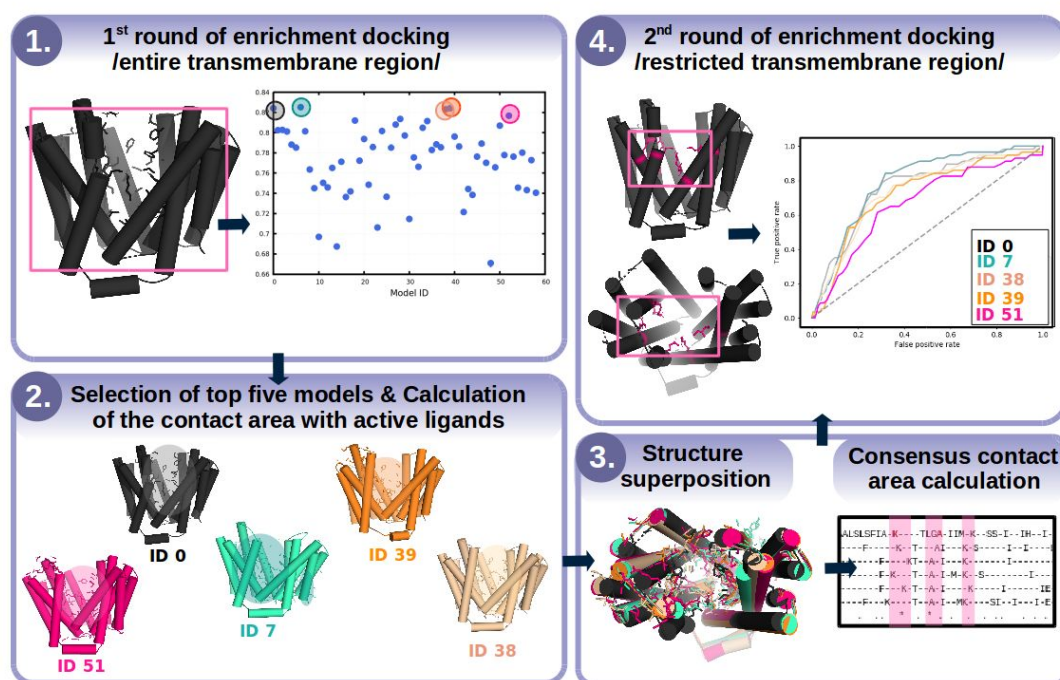
Supplementary Figure S1

Six lowest frequency modes calculated for the selected template (Fucose transporter, PDB ID 3o7q). Black arrows in the structures indicate the magnitude and directionality of the fluctuation vectors.



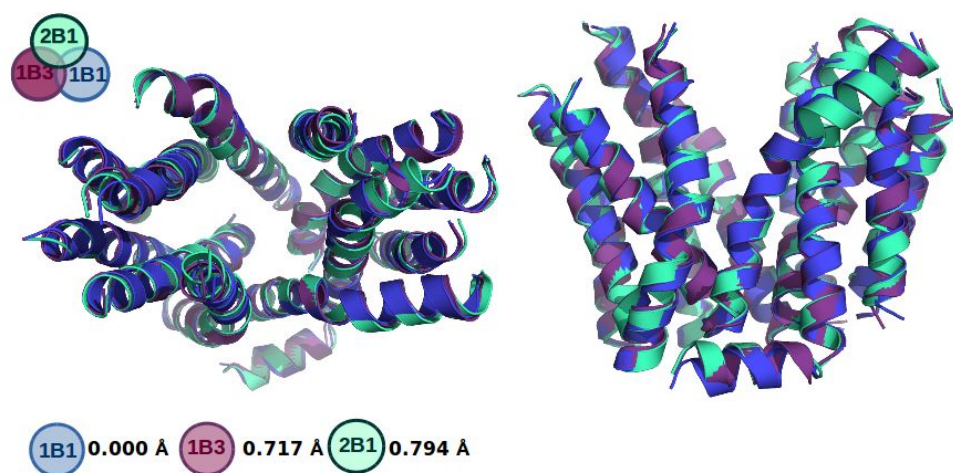
Supplementary Figure S2

Covariance matrices showing a) sequence (Hamming distance), b) secondary structure (RMSD), and c) dynamic similarity between retrieved MFS transporters. The cluster of dynamically analogous transporters is shown by the red square in the matrices. Blue color indicates a strong similarity (identity), whereas red corresponds to complete dissimilarity. The sequence- (a) and secondary structure- (b) based matrices are reordered according to the dynamics- (c) matrix to demonstrate that the sequentially and/or structurally divergent proteins might still share comparable intrinsic dynamics. PDB IDs of the structures unravelled upon the similarity in their intrinsic dynamics are the following: 3o7q (reference structures), 4m64, 3wdo, 4gby, 4gbz, 4gc0, 4zwc, 4zw9, 5c65, 4jre, 4zwb, 4oaa, 4zyr, 4gxb, 3o7qp, 5



Supplementary Figure S3

Schematic overview of the enrichment docking procedure.



Supplementary Figure S4

Comparison of the top prioritized structural models for OATP1B1 (the blue structure), OATP1B3 (the magenta structure), and OATP2B1 (the green structure). An average RMSD was calculated (OATP1B1 was defined as a reference structure for RMSD calculation).

Table listing amino acid residues spanning the different transmembrane regions for the three hepatic transporters.

TMH#	amino acid sequence
1	OATP1B1: K----MFLAALSLSFIAKTLGAIIMKSSIIHIERR OATP1B3: K----MFLAALSFSYIAKALGGIIMKISITQIERR OATP2B1: KL FVL----CHSL LQLAQLMISGYLKSSISTVEKR Cons: *.....*.....*.***...*.*
2	OATP1B1: SLVGFIDGSFEIGNLLVIVFVS YFGSKLHR OATP1B3: SLAGLIDGSFEIGNLLVIVFVS YFGSKLHR OATP2B1: QTSGLLASFNEVGNTALIVFVS YFGSRVHR Cons: *.....*.**...*****..**
3	OATP1B1: PKLIGIGCFIMGIGVLTALPHFF OATP1B3: PKLIGIGCLLMGTGSILTSLPHFF OATP2B1: PRMIGYGAILVALAGLLMTLPHF I Cons: *..**.*.....*..****.
4	OATP1B1: WIYVF MGNMLRGIGETPIVPLGLSYI OATP1B3: WIYVF MGNMLRGIGETPIVPLGISYI OATP2B1: VGIMFVAQTLLGVGGVPIQPFGISYI Cons: *.....*.*.*..**.*.*.***
5	OATP1B1: HSSLYL GILNAIAMIGPIIGFTLGSLF OATP1B3: HSSLYL GSLN AIGMIGPVIGFALGSLF OATP2B1: NSPLYL GILFAVTMMGPGLAFGLGSLM Cons: .*.****.*.*.....*.****.
6	OATP1B1: WWLNFLVSGLF SI ISSIPFFF OATP1B3: WWLGFLVSGLF SI ISSIPFFF OATP2B1: WWLGFL IAAGA VALAA IPYFF Cons: ***.*.....**.*
7	OATP1B1: GFFQSFKSILTNP LYVMFVLL TLLQVSSYIGAFTY VFKYVEQQ OATP1B3: GFFQSLKSILTNP LYVIFLLL TLLQVSSFIGSFTY VFKYMEQQ OATP2B1: VFPRVLLQTLRHPI FL LVLSQVCLSSMAAGMAIFLPKFLE RQ

	Cons:	. * * . . * * * * * *
8	OATP1B1:	KANILLGVITIPIFASGMFLGGYIIK
	OATP1B3:	HANFLLGIITIP TVATGMFLGGFIIK
	OATP2B1:	YANLLIGCLSFP SVIVGIVVGGVLVK
	Cons:	. ** . * . * * * *
9	OATP1B1:	VGIAKFSCFTAVMSLSFYLLY
	OATP1B3:	VGIAKFSFLTSMISFLFQLLY
	OATP2B1:	VGCGALCLLGMLLCLFFSLPL
	Cons:	* * * . * . .
10	OATP1B1:	FFYFFVAIQVLNLFFSALGGTSHVML
	OATP1B3:	FFFIYVAIQVINSLFSATGGTTFILL
	OATP2B1:	FVVPFLLL VSLGSALACLTHTPSFML
	Cons:	* * *
11	OATP1B1:	LALGFHSMVIRALGGILAPIYFGALIDT
	OATP1B3:	LAMGFQSMVIRTLGGILAPIYFGALIDK
	OATP2B1:	LAVGIQFMFLRILAWMPSPVIHGSAIDT
	Cons:	* * . * . . . * . . * * . . . * . * .
12	OATP1B1:	TFSRVYLGLSSMLR-----VSSLVL
	OATP1B3:	TFGRVYLGLSIALRFPALVL-----
	OATP2B1:	TLRNRFIGLQFFFK-----TGSVIC
	Conf:	* * *

Supplementary Table S8

Residues known to be implicated in the transport function. Transmembrane helix number, amino acid, probe substrate, and reference to the primary source (PubMed ID) are indicated.

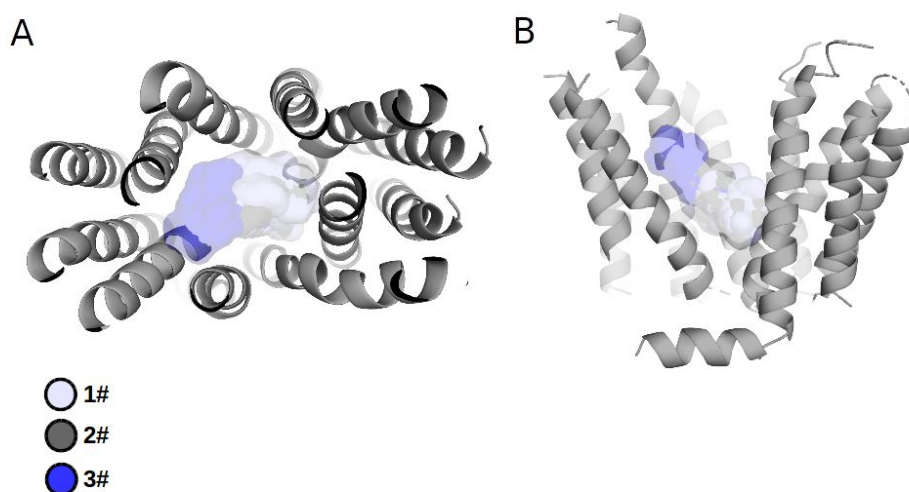
Residues which were observed in the docking poses of steroidal compounds are marked in color (blue for OATP1B1, pink for OATP1B3, and green for OATP2B1, respectively).

TMH	Amino acid	Probe substrate	OATP1B1	OATP1B3	OATP2B1
1	L31/L31/V52	estrone-3-sulfate, taurocholic acid			29871943
1	L34/L34/H55	estrone-3-sulfate, taurocholic acid			29871943
1	F38/Y38/Q59	estrone-3-sulfate, taurocholic acid			29871943
1	S37/S37/L58	estrone-3-sulfate, taurocholic acid			29871943
1	K41/K41/Q62	estrone-3-sulfate	31254566		29871943
1	A45/G45/S66	cholecystokinin-8, estrone-3-sulfate, taurocholic acid, epigallocatechin gallate	22352740	22352740 ,31353905	29871943
1	M48/M48/L69	estrone-3-sulfate, taurocholic acid,			29871943
1	K49/K49/K70	estrone-3-sulfate	31254566		
1	R57/R57/K78	estradiol-17 β -glucu	20821001		

		ronide, estrone-3-sulfate , bromosulfophthalei n			
2	D70 /D70/ A91	estrone-3-sulfate	22574206		
2	F73 /F73/ N94	estrone-3-sulfate, estradiol 17beta-d-glucuroni de	22574206, 11477075		
2	E74 / E74 / E95	estrone-3-sulfate	22574206		
2	G76/G76/G97	estrone-3-sulfate	22574206		
4	V174/V174/M192	estrone-3-sulfate, SN-38, pravastatin, estradiol-17beta-glu curonide	15608127		
5	R181 /R181/L199	estradiol - 17β - gl ucuronide	https://doi.o rg/10.1096/ fasebj.21.5. A196-d		
6	W258/W258/W276	taurocholic acid, estrone-3-sulfate	23858103		
6	W259/W259/W277	taurocholic acid,	23858103		

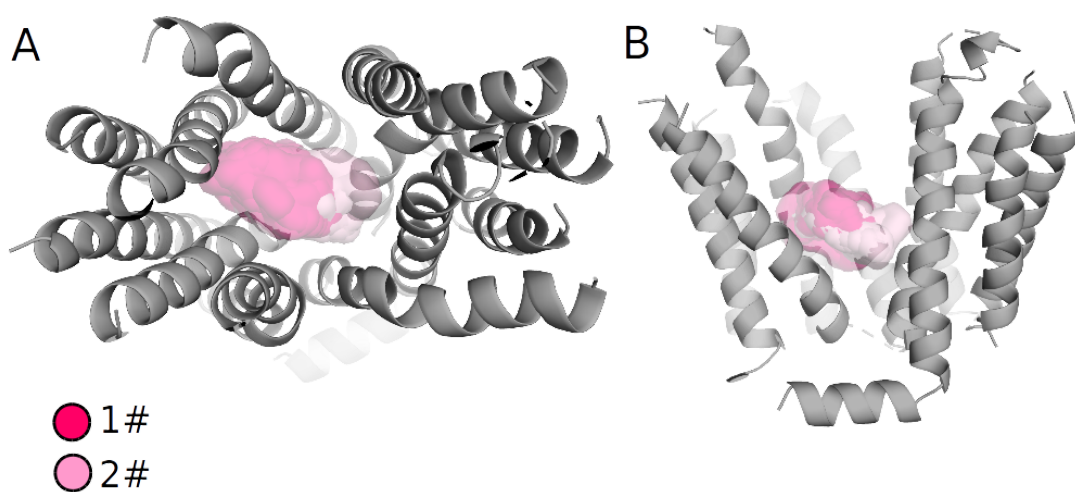
		estrone-3-sulfate			
7	T345/T345 /Q380	estrone-3-sulfate, estradiol 17beta-glucuronide	15632119		
7	I353/I353/A388	estrone-3-sulfate, estradiol 17beta-d-glucuronide	11477075		
7	K361/K361/K386	estrone-3-sulfate, estradiol-17β-glucuronide, bromosulphophthalen	20821001	21642393	
7	T357/T357/I392	estrone-3-sulfate			12130747
10	F537/Y537/F564	cholecystokinin-8		18690707	
10	L545/S545/S572	estrone-3-sulfate	19760661, 22352740	22352740, 18690707	
10	F546/L546/A573	estrone-3-sulfate	19760661		
10	L550/T550/L577	estrone-3-sulfate	19760661	18690707	
10	G552/G552 /H579	estrone-3-sulfate, pravastatin, rosuvastatin,			https://doi.org/10.1039/C6MD00235

		sulfasalazine, naringin, progesterone			<u>H</u>
10	S554/T554/P581	estrone-3-sulfate	19760661		
10	H555/F555/S582	epigallocatechin gallate		31353905	
11	R580/ R580 /R607	estrone-3-sulfate,est radiol-17 β -glucuron ide,bromosulfophth alein	20821001		
12	F591/F591/H618	estrone-3-sulfate, pravastatin, rosuvastatin, sulfasalazine,naring in, progesterone			https://doi.org/10.1039/C6MD00235 <u>H</u>



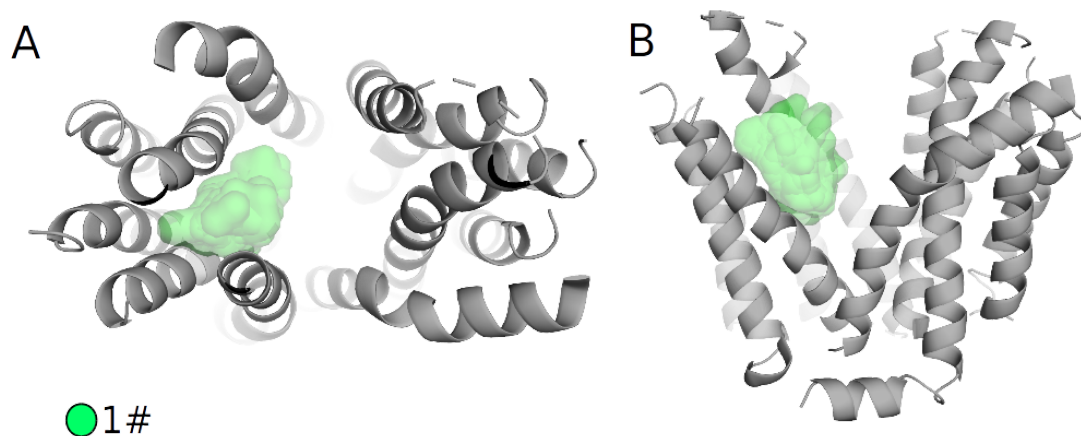
Supplementary Figure S5

Top three enriched clusters in OATP1B1. For sake of simplicity, maximum common substructure of the docked steroids is displayed in the transparent surface representation. Enriched clusters (cluster 1#, 2#, and 3#) are colored as indicated in the Figure. In Figure (B) TMH2 and TMH11 are excluded from the visualization.



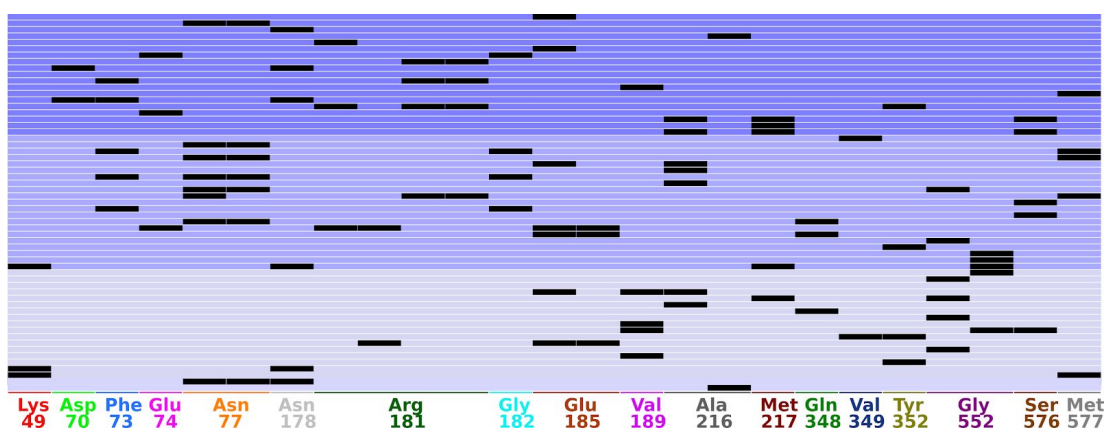
Supplementary Figure S6

Top two enriched clusters in OATP1B3. For sake of simplicity, maximum common substructure of the docked steroids is displayed in the transparent surface representation. Enriched clusters (cluster 1# and 2#) are colored as indicated in the Figure. In Figure (B) TMH2 and TMH11 are excluded from the visualization.



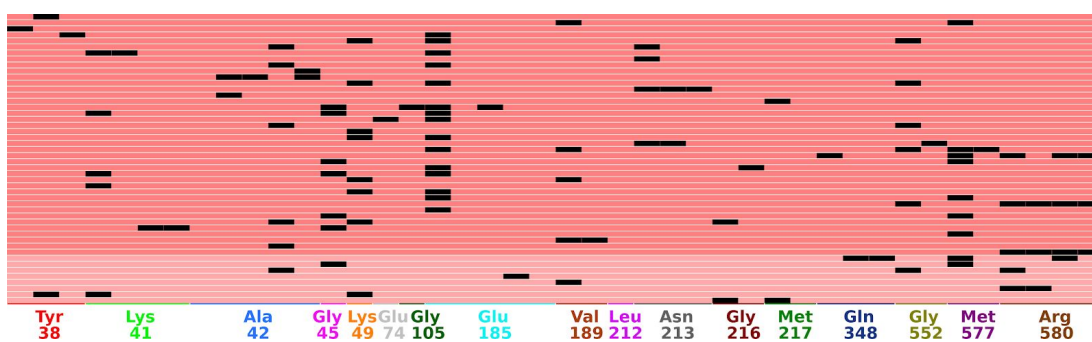
Supplementary Figure S7

Top enriched cluster in OATP2B1. For sake of simplicity, maximum common substructure of the docked steroids is displayed in the transparent surface representation. Enriched cluster (cluster 1#) is colored as indicated in the Figure. In Figure (B) TMH2 is excluded from the visualization.



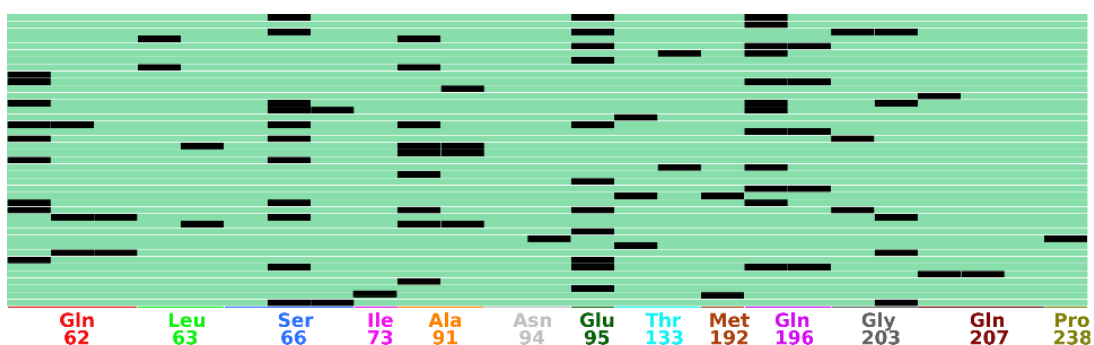
Supplementary Figure S8

Protein-Ligand Interactions Fingerprints for top three enriched OATP1B1 clusters. Distinct clusters are highlighted in different colors.



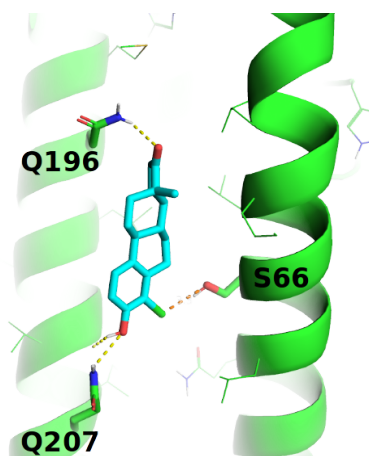
Supplementary Figure S9

Protein-Ligand Interactions Fingerprints for top two enriched OATP1B3 clusters. Distinct clusters are highlighted in different colors.



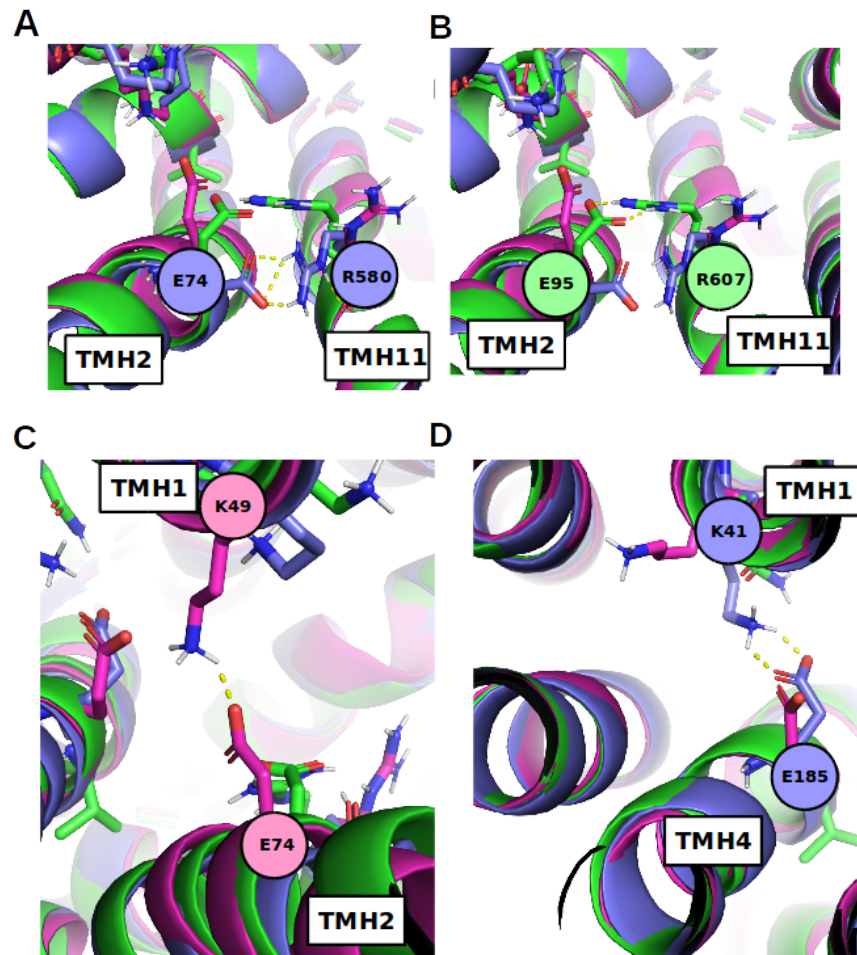
Supplementary Figure S10

Protein-Ligand Interactions Fingerprints for the top enriched OATP2B1 cluster.



Supplementary Figure S11

Chlorinated 13-epiestrone (compound 27) at the R-4 position binds to a close proximity of SER66 (distance 3.3 Å). However, the interaction angle C19-C11---OG(S66) is approximately 110°, thus the likelihood for the halogen bond formation is decreased. Color coding: docked compounds = cyan (carbon), red (oxygen), green (chloride), OATP2B1 = green.



Supplementary Figure S12

Identified intramolecular salt bridges in hepatic OATPs. (A) GLU74(TM2)-ARG580(TM11) bridge aims to stabilize transporter structure. The same scenario has been observed with the corresponding residues in OATP2B1 (as shown in Figure B). In OATP1B3 (C), a salt bridge is formed between residues at TMH1 (LYS49) and TMH2 (GLU74), which impacts the N-terminal binding site. (D) An additional salt bridge is formed in OATP1B1 between TMH1 (LYS41) and TMH4 (GLU185). Color coding: OATP1B1 = the blue structure, OATP1B3 = the magenta structure, OATP2B1 = the green structure

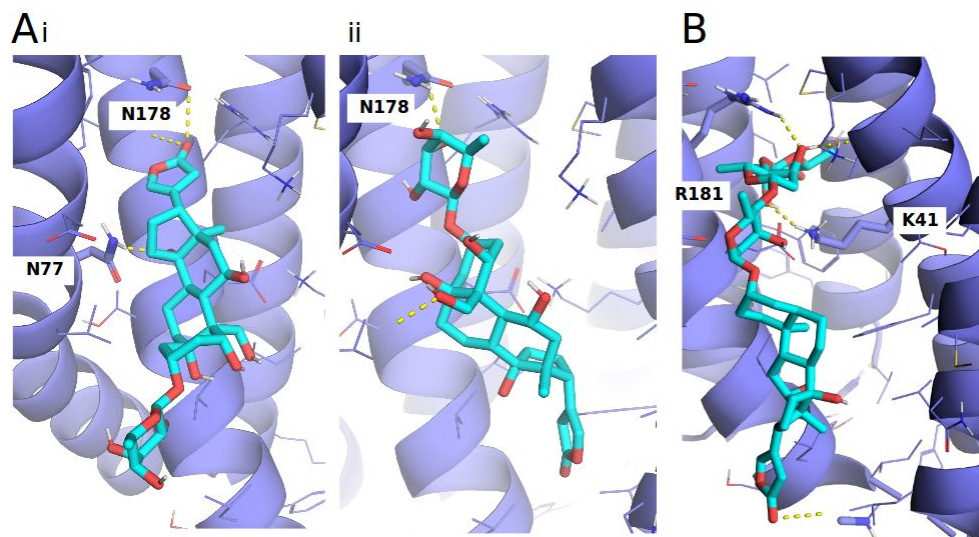


Figure S13

The N-terminal region in OATP1B1 is capable of accommodating ring moieties. (A) Ouabain with the (i) R-17 and (ii) R-3 ring substituent. (B) Digoxin. Color coding: docked compounds = cyan (carbon), red (oxygen), OATP1B1 = blue, OATP1B3=magenta, OATP2B1 = green.

Additional supplementary files

Supplementary file S1: Alignment file used for the comparative modeling on OATP1B1 ('OATP1B1.ali').

This file is freely available on GitHub: <https://github.com/AlzbetaTuerkova/EnsembleDocking>

Supplementary file S2: Alignment file used for the comparative modeling on OATP1B3 ('OATP1B3.ali')

This file is freely available on GitHub: <https://github.com/AlzbetaTuerkova/EnsembleDocking>

Supplementary file S3: Alignment file used for the comparative modeling on OATP2B1 ('OATP2B1.ali')

This file is freely available on GitHub: <https://github.com/AlzbetaTuerkova/EnsembleDocking>

Supplementary file S4: A .csv table listing the steroid analogs gathered from the open domain. Table lists the target (OATP1B1, OATP1B3, OATP2B1), InChIKey, all available bioactivity values per compound in MicroM range, canonical smiles, and compound name.

This file is freely available on GitHub: <https://github.com/AlzbetaTuerkova/EnsembleDocking>

Supplementary file S5: A python script for retrieving a maximum common substructure out of the group of 3D ligand structures.

```
import glob
import sys
from rdkit import Chem
from rdkit.Chem import rdFMCS

# Find all .mol files in a current directory

mol_files = glob.glob('*.mol')
mol_list= []

# create a list of input molecules

for mol in mol_files:
    single = Chem.MolFromMolFile(mol)
    mol_list.append(single)

# find maximum common substructure (bond order is set to be flexible)

res = rdFMCS.FindMCS(mol_list,
bondCompare=rdFMCS.BondCompare.CompareAny).smartsString
pattern = Chem.MolFromSmarts(res)

# read a molecule in mol format

m = Chem.MolFromMolFile(sys.argv[1])
conf = m.GetConformer()
sub = m.GetSubstructMatch(pattern)

# get coordinates of a maximum common substructure of a molecule

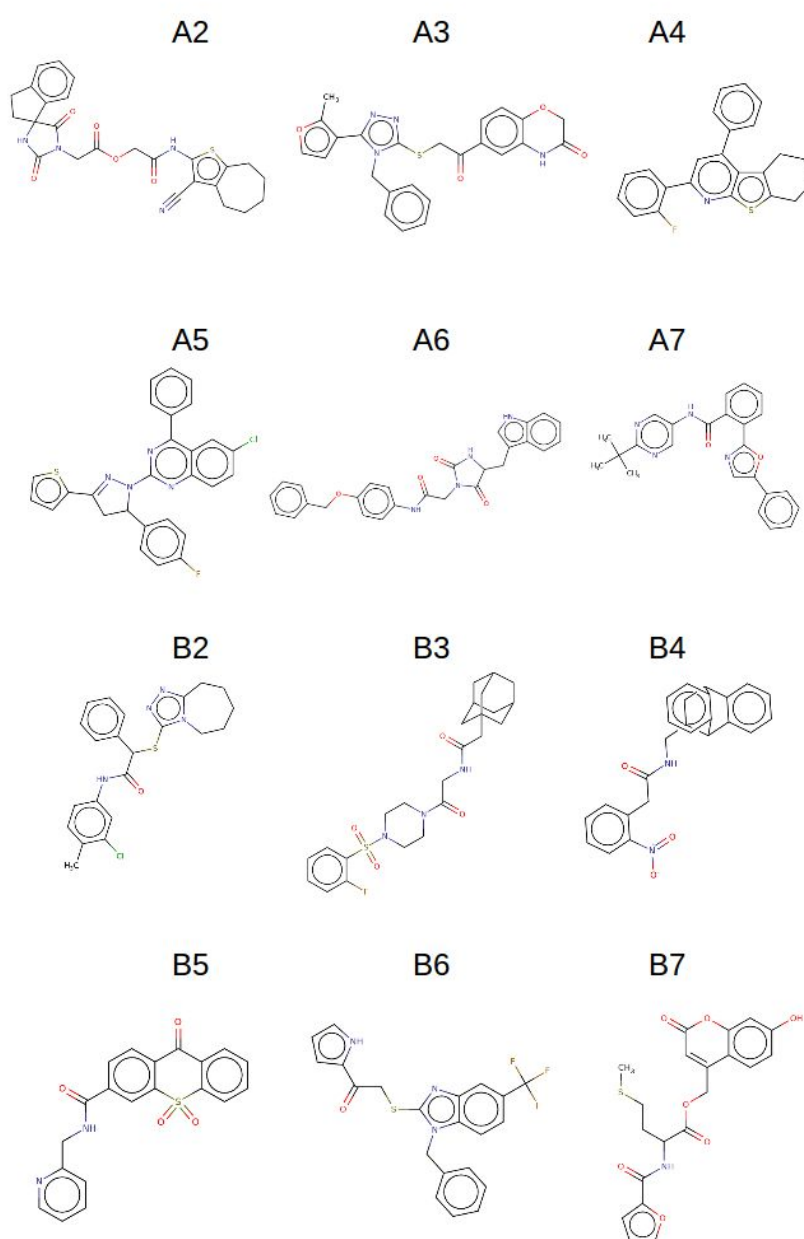
print(len(sub))
print('')
for s in sub:
    coordinates=conf.GetAtomPosition(s)
    print(str(m.GetAtoms()[s].GetSymbol()) + " " + str(coordinates.x) + " "
+ str(coordinates.y) + " " + str(coordinates.z))
```

Supplementary Information 6

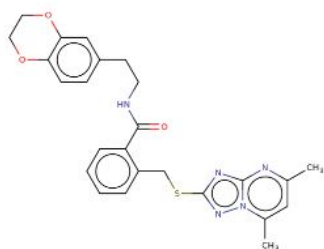
This section includes the supplementary information for Study 6: Alzbeta Tuerkova, Brandon J. Bongers, Ulf Norinder, Csilla Özvegy-Laczka, Gerard JP van Westen, Barbara Zdrazil. Combining AI-driven and structure-based approaches to identify novel inhibitors of hepatic organic anion transporting polypeptides (OATPs), 2020

Supporting Information

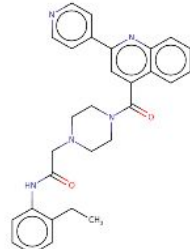
Combining AI-driven and structure-based approaches to identify novel inhibitors of hepatic organic anion transporting polypeptides (OATPs)



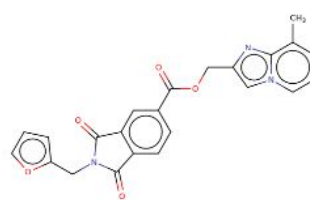
C2



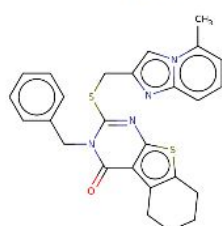
C3



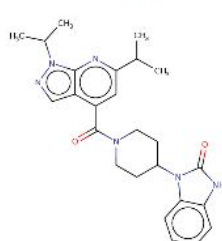
C4



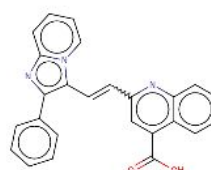
C5



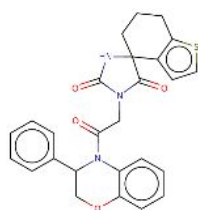
C6



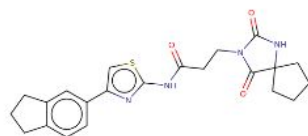
C7



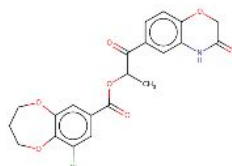
D2



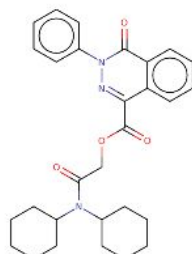
D3



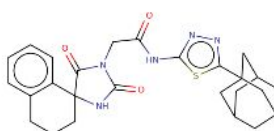
D4



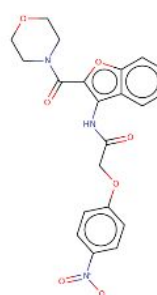
D5

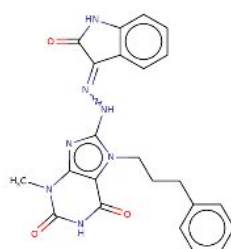
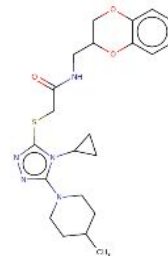
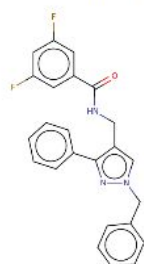
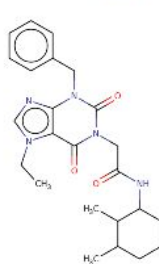
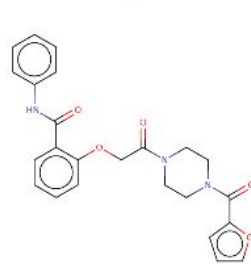
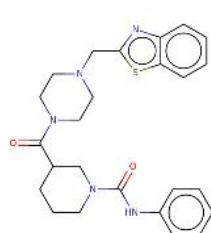
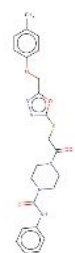
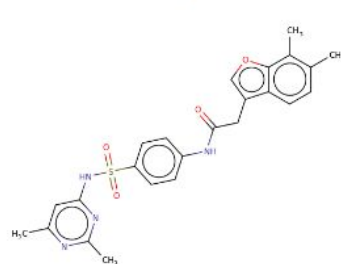
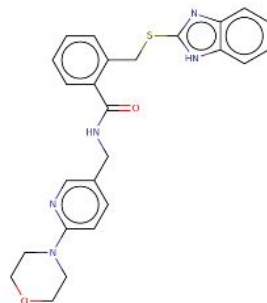
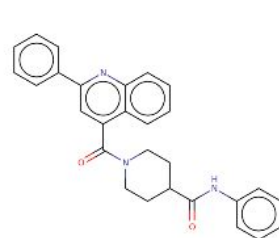
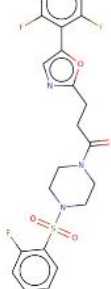


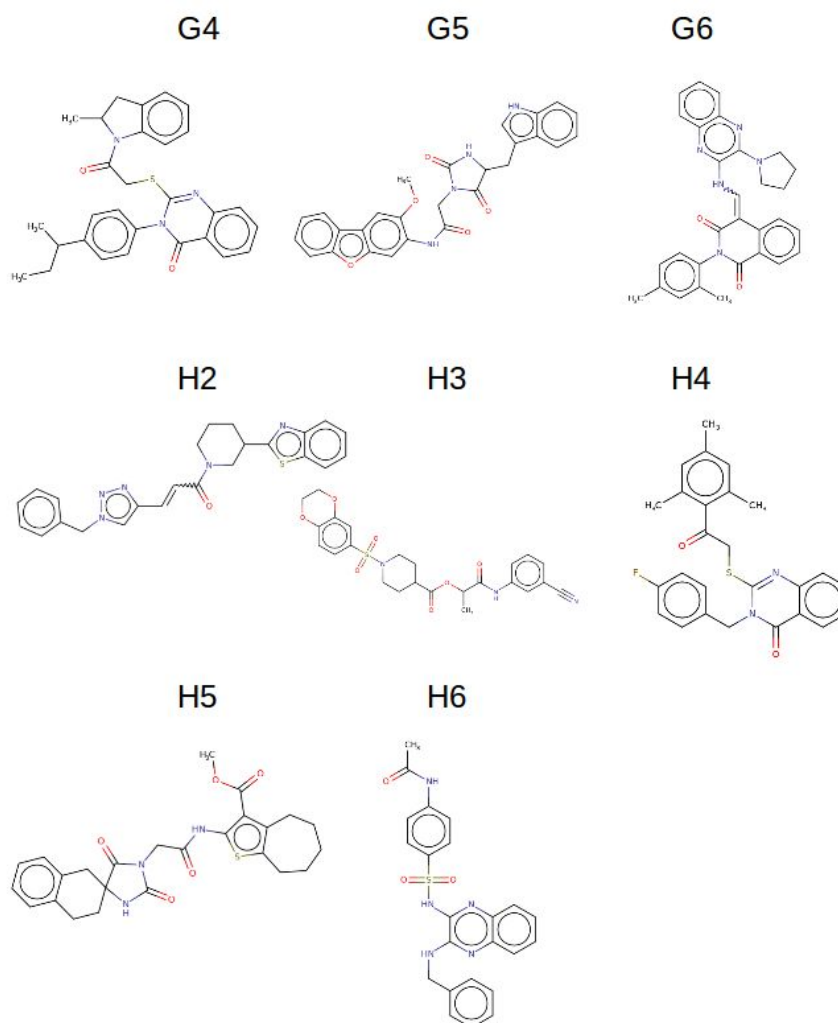
D6



D7



E2**E3****E4****E5****E6****F2****F3****F4****F5****F6****G2****G3**



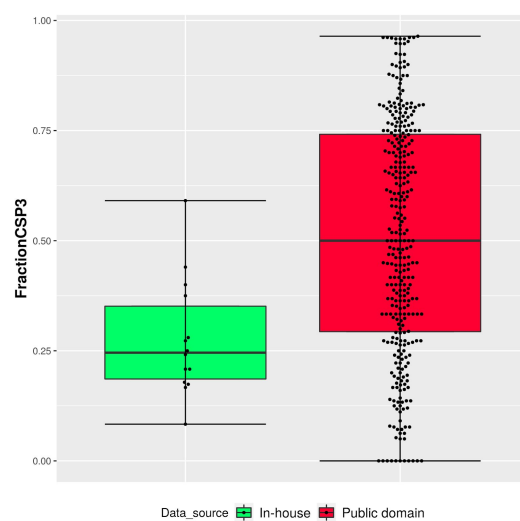
Supplementary Figure S1: Chemical structures (n=44) and associated codes from Table 1 prioritized by structure-based virtual screening.

	IC ₅₀ (μM)		
	OATP1B1	OATP1B3	OATP2B1
H6	no effect	no effect	> 10
H5	~25	6.95	0.39
G4	no effect	no effect	~6
B2	no effect	> 10	> 10
E5	7.61	7.50	1.48
C7	> 10	5.4	0.04
E3	2.69	1.53	1.32
B4	~10	~10	2.37

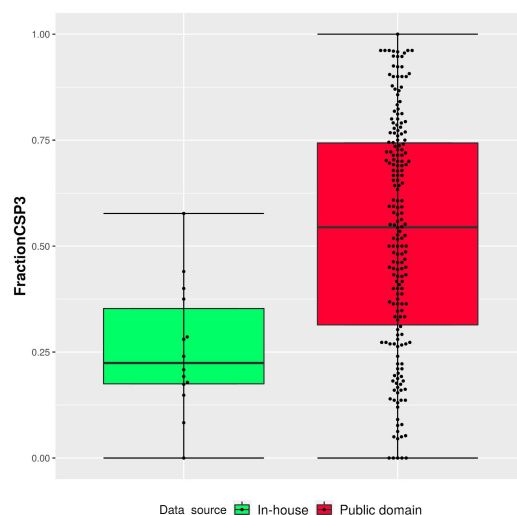
Supplementary Table S1: Table listing IC₅₀ values for eight selected inhibitors.

Compound code	Canonical SMILES
H6	<chem>CC(=O)Nc1ccc(cc1)S(=O)(=O)Nc2nc3cccc3nc2NCc4cccc4</chem>
H5	<chem>COC(=O)c1c2CCCCC2sc1NC(=O)CN3C(=O)NC4(CCc5cccc5C4)C3=O</chem>
G4	<chem>CCC(C)c1ccc(cc1)N2C(=Nc3cccc3C2=O)SCC(=O)N4C(C)Cc5cccc45</chem>
B2	<chem>[O-][N+](=O)c1cccc1CC(=O)NCC2CC3c4cccc4C2c5cccc35</chem>
E5	<chem>Fc1cc(F)cc(c1)C(=O)NCc2cn(Cc3cccc3)nc2c4cccc4</chem>
C7	<chem>OC(=O)c1cc(\C=C\c2c(nc3cccn23)c4cccc4)nc5cccc15</chem>
E3	<chem>CN1C(=O)NC(=O)c2c1nc(N\N=C/3\C(=O)Nc4cccc34)n2CCCc5cccc5</chem>
B4	<chem>[O-][N+](=O)c1cccc1CC(=O)NCC2CC3c4cccc4C2c5cccc35</chem>

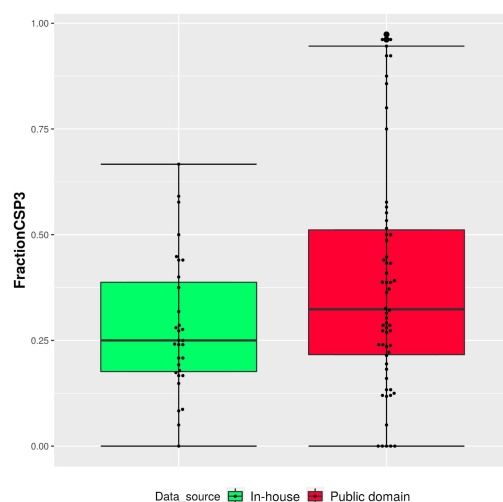
Supplementary Table S2: Canonical SMILES and for eight selected inhibitors, as indicated in Supplementary Table S1.



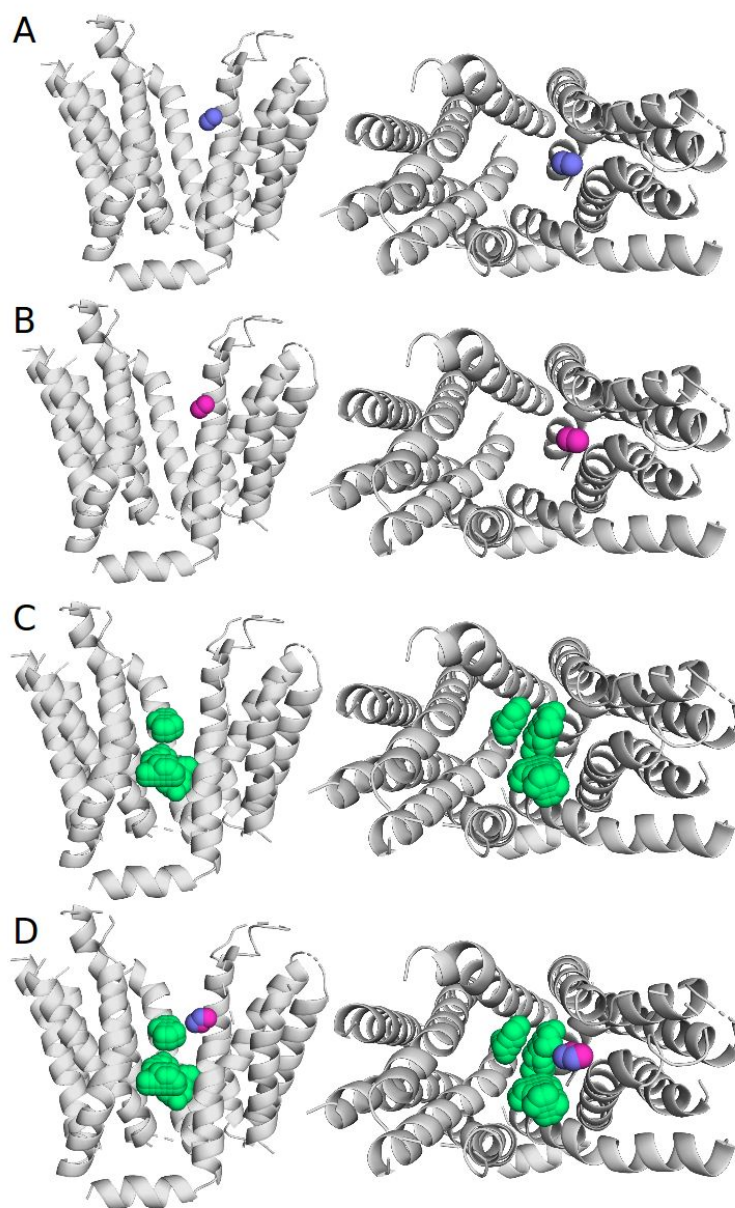
Supplementary Figure S2: Distribution of the FractionCSP3 in the newly measured OATP1B1 inhibitors (“in-house”, activity threshold $\leq 10 \mu\text{M}$), versus OATP1B1 inhibitors originating from the public domain.



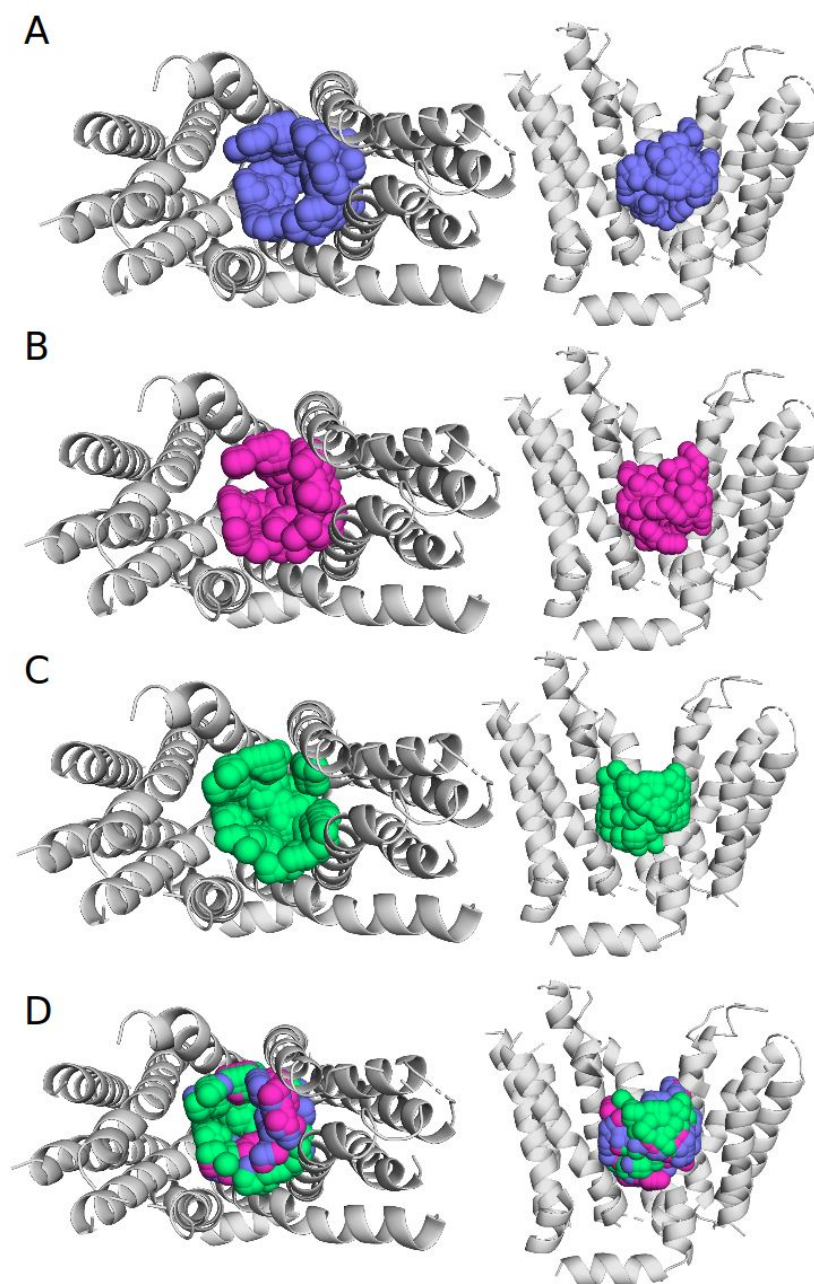
Supplementary Figure S3: Distribution of the FractionCSP3 in the newly measured OATP1B3 inhibitors (“in-house”, activity threshold $\leq 10 \mu\text{M}$), versus OATP1B3 inhibitors originating from the public domain.



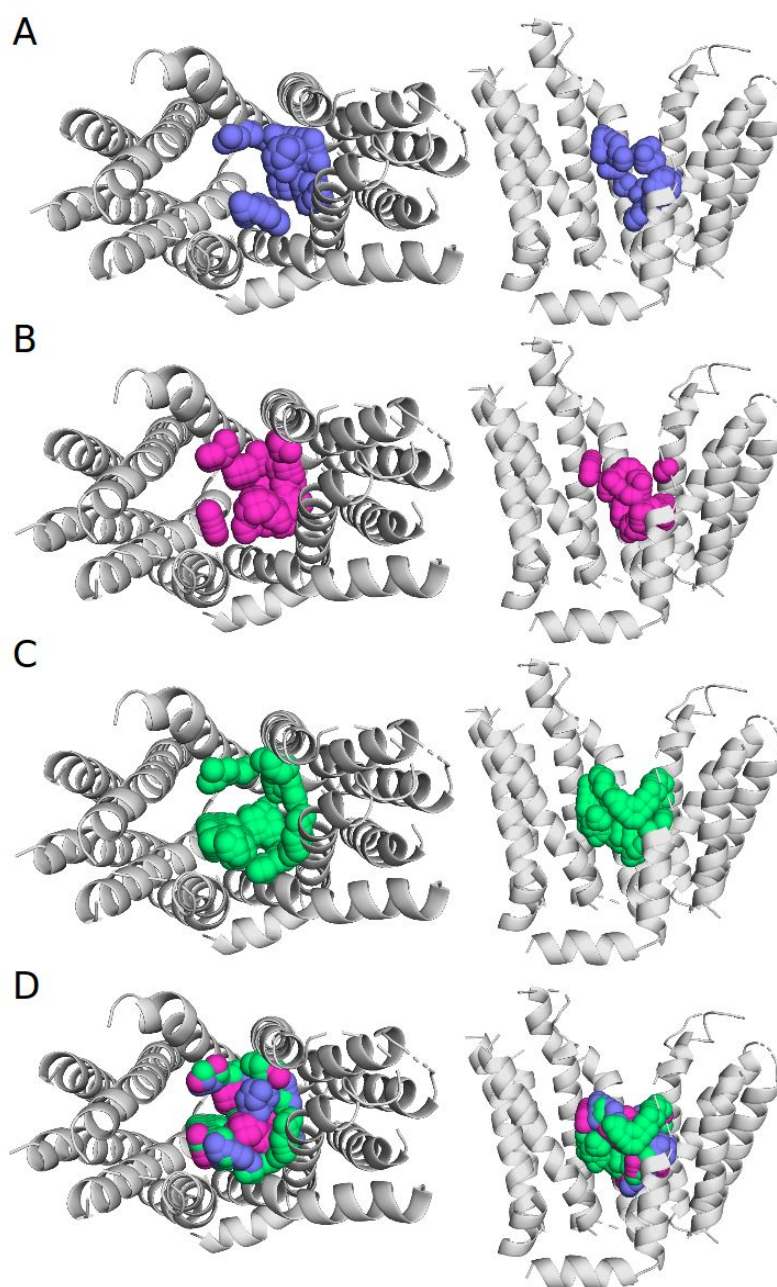
Supplementary Figure S4: Distribution of the FractionCSP3 in the newly measured OATP2B1 inhibitors (“in-house”, activity threshold $\leq 10 \mu\text{M}$), versus OATP2B1 inhibitors originating from the public domain.



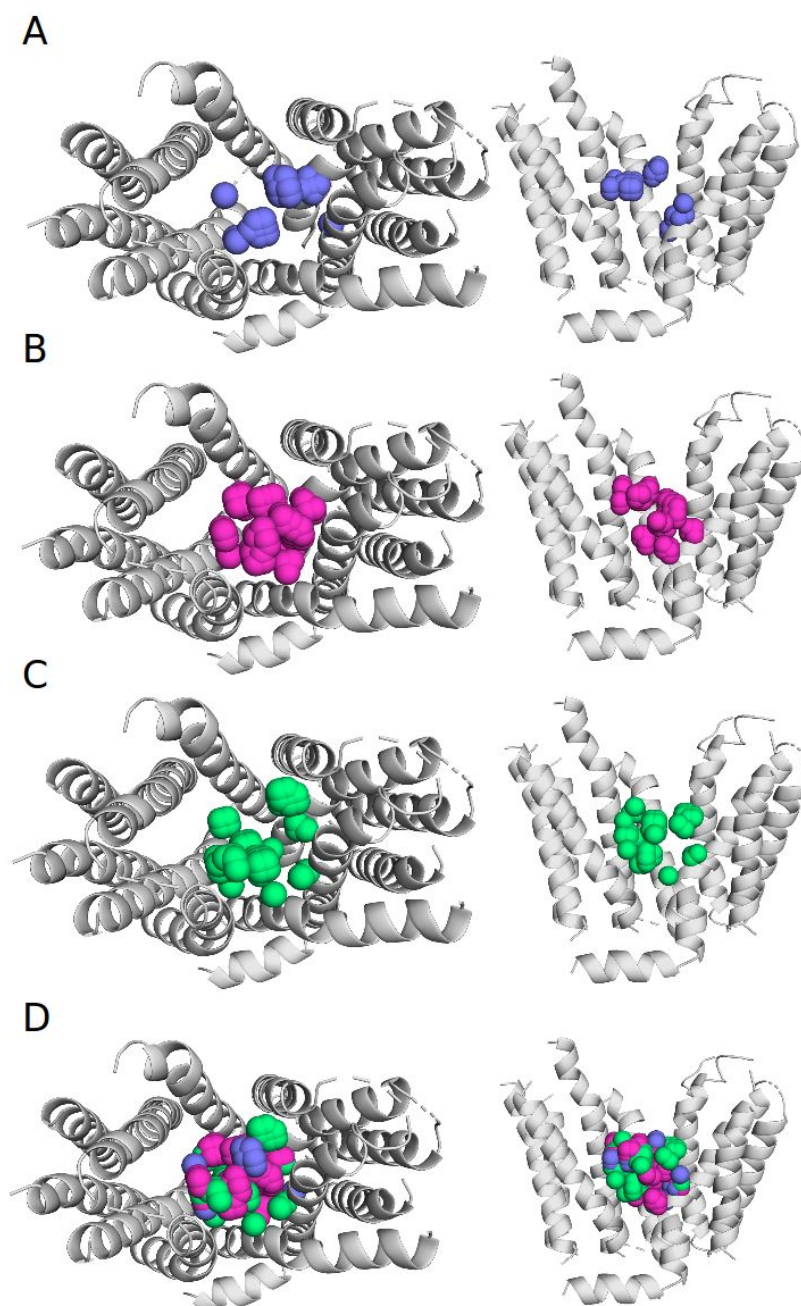
Supplementary Figure S5: Volumetric map showing the distribution of **aromatic** residues in the binding site of (A) OATP1B1 (blue coloring), (B) OATP1B3 (magenta coloring), (C) OATP2B1 (green coloring), and (D) superimposed transporters.



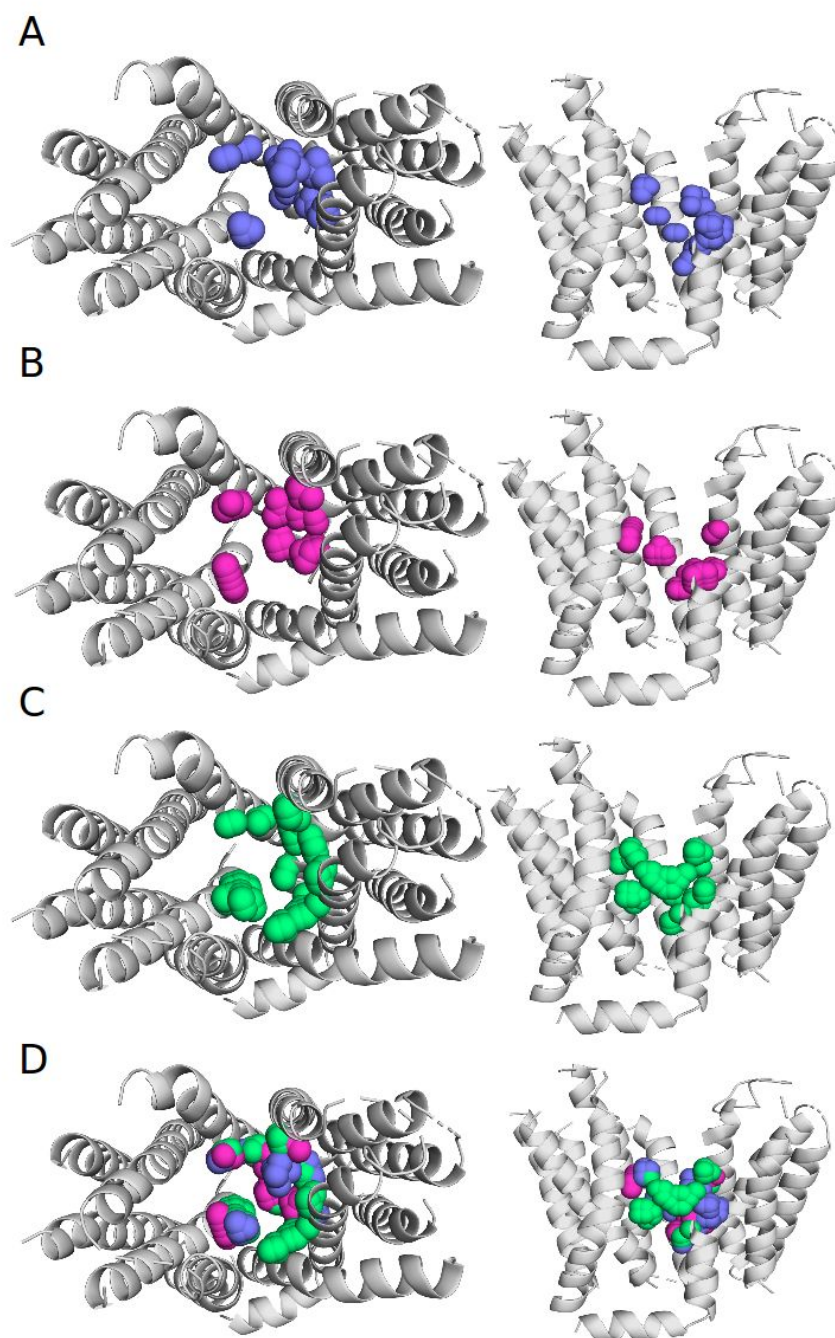
Supplementary Figure S6: Volumetric map showing the distribution of **hydrophobic** residues in the binding site of (A) OATP1B1 (blue coloring), (B) OATP1B3 (magenta coloring), (C) OATP2B1 (green coloring), and (D) superimposed transporters.



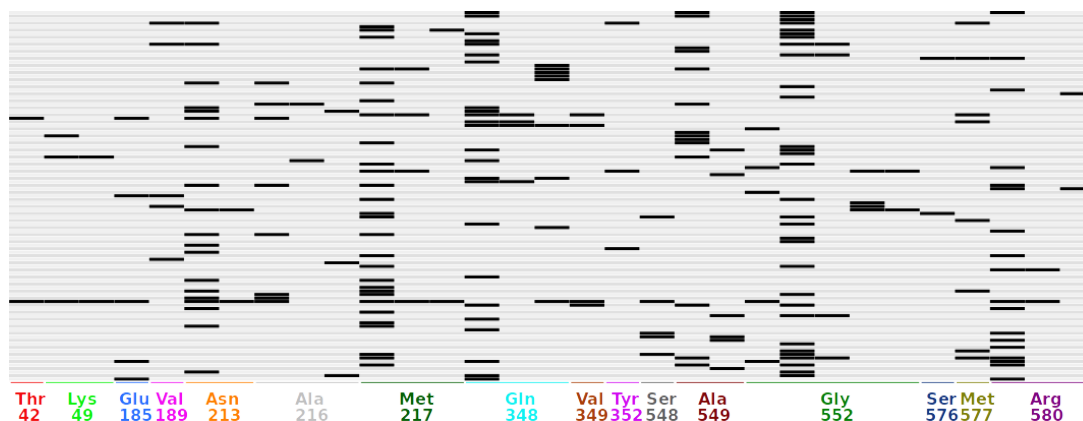
Supplementary Figure S7: Volumetric map showing the distribution of **hydrophilic** residues in the binding site of (A) OATP1B1 (blue coloring), (B) OATP1B3 (magenta coloring), (C) OATP2B1 (green coloring), and (D) superimposed transporters.



Supplementary Figure S8: Volumetric map showing the distribution of **H-bond donors** in the binding site of (A) OATP1B1 (blue coloring), (B) OATP1B3 (magenta coloring), (C) OATP2B1 (green coloring), and (D) superimposed transporters.



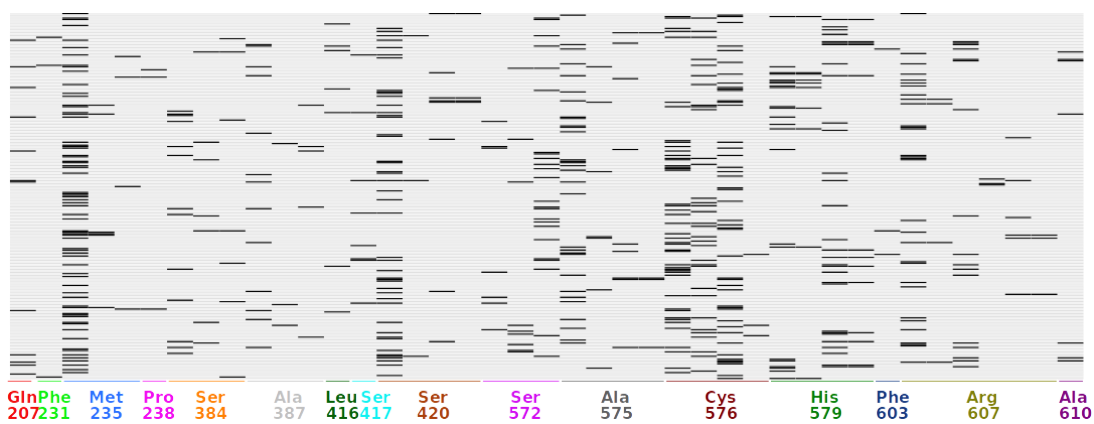
Supplementary Figure S9: Volumetric map showing the distribution of **H-bond acceptors** in the binding site of (A) OATP1B1 (blue coloring), (B) OATP1B3 (magenta coloring), (C) OATP2B1 (green coloring), and (D) superimposed transporters.



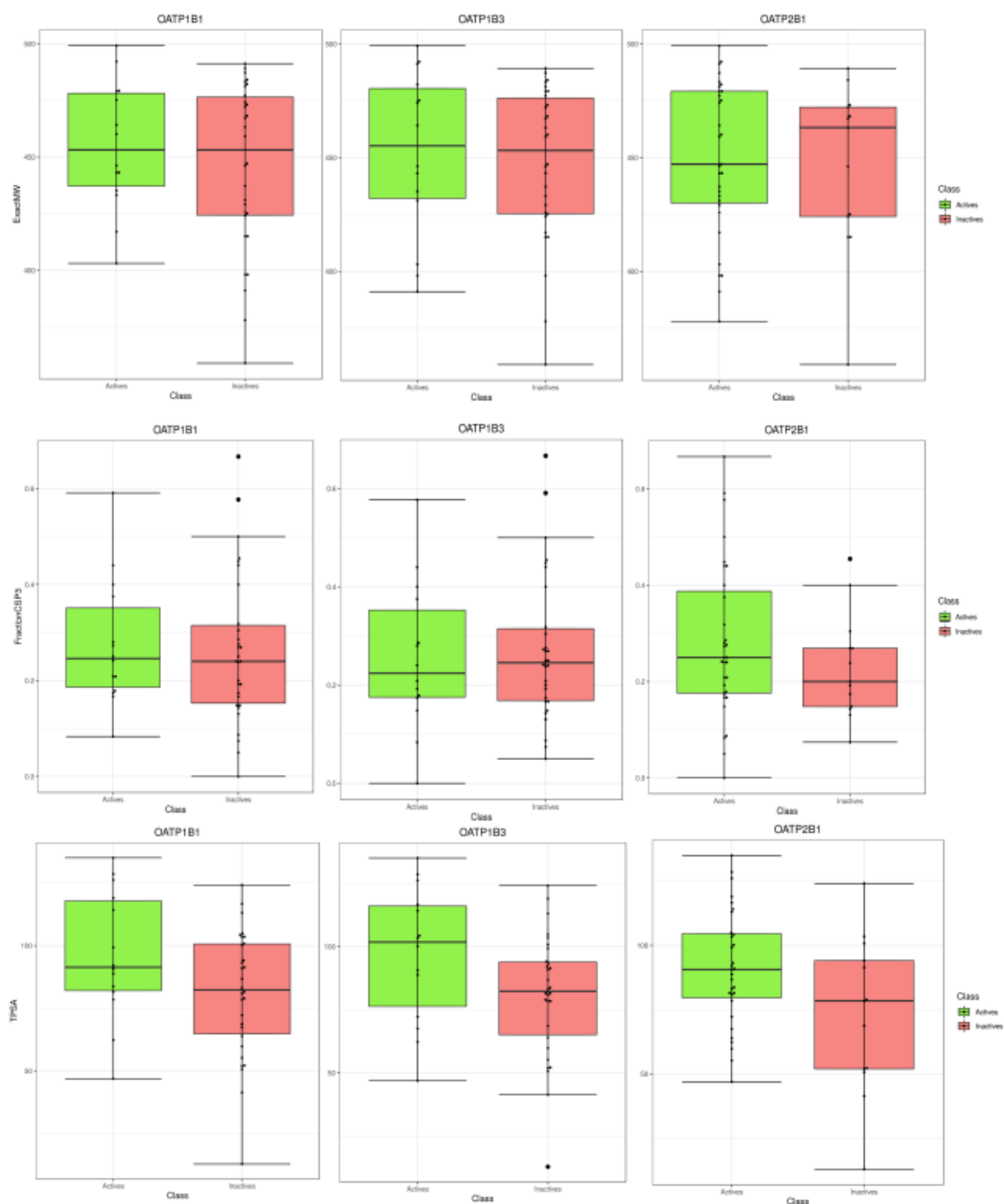
Supplementary Figure S10: Protein-ligand interaction fingerprints for OATP1B1 inhibitors (threshold $\leq 10 \mu\text{M}$).



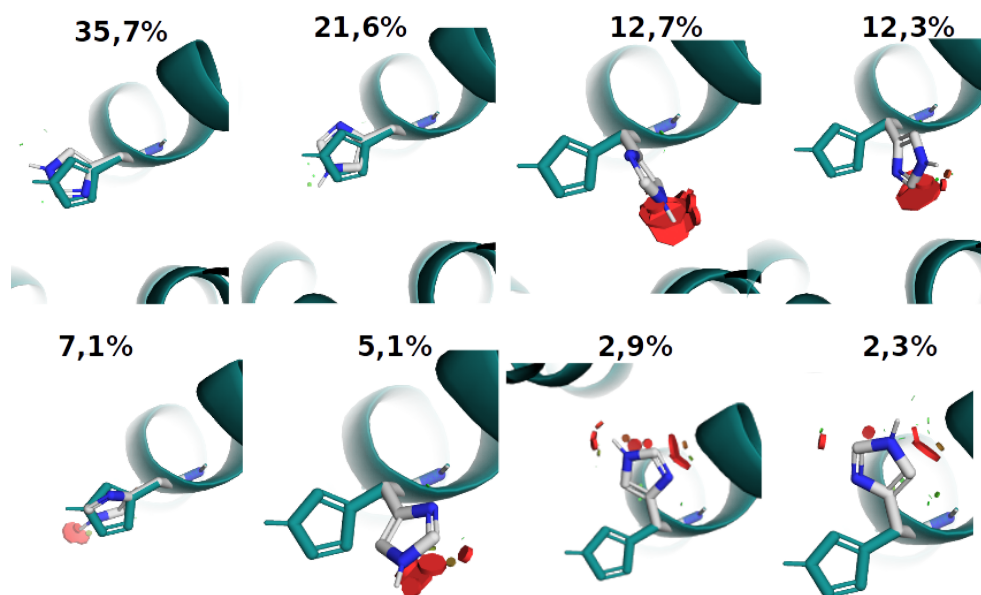
Supplementary Figure S11: Protein-ligand interaction fingerprints for OATP1B3 inhibitors (threshold $\leq 10 \mu\text{M}$).



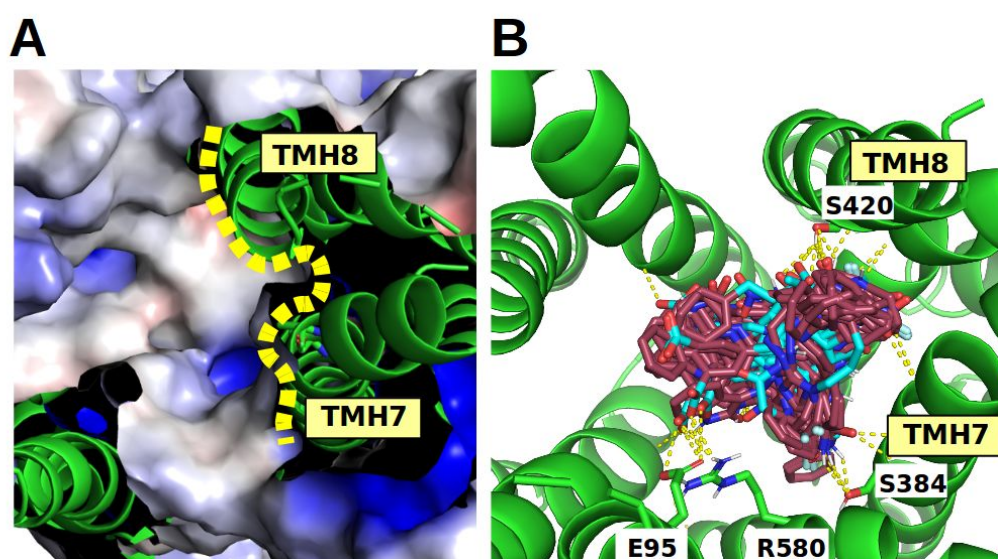
Supplementary Figure S12: Protein-ligand interaction fingerprints for OATP2B1 inhibitors (threshold $\leq 10 \mu\text{M}$).



Supplementary Figure S13: Distribution of molecular weight, FractionCSP3, and TPSA for OATP1B1, OATP1B3, OATP2B1 (non)inhibitors (threshold < 10 μ M).



Supplementary Figure S14: Possible rotamers for HIS579 in OATP2B1. Red regions show steric clashes with the surrounding residues. Percentage values indicate the occupancy of a certain rotamers. Rotamer analysis shows that HIS579 is pointing inside of the OATP2B1 binding site with a high probability, thus having an impact on the pocket geometry.



Supplementary Figure S15: (A) Mapped electrostatic potential onto the OATP2B1 surface shows an accessible cavity between TMH7 and TMH8. (B) Docked poses for the most potent OATP2B1 actives (codes: B4, C7, E3, E5) shows the shape complementarity with the OATP2B1 binding site. Several residues (SER420 or SER384) form H-bonds with the docked ligands (indicated by the yellow dashed line).

Abstract

Uptake transporters of the solute carrier (SLC) superfamily expressed in hepatocytes help maintain cellular homeostasis by regulating the transport of both endogenous substrates and xenobiotic compounds. An impaired function of these transporters might lead to clinically relevant drug-drug interactions with a potential development of drug-induced liver injury. In this thesis, we followed a holistic *in silico* approach, connecting structure-based modeling with data-science (cheminformatics) approaches to gain an in-depth understanding of transporter-ligand interactions. The biological focus of the thesis lied on hepatic organic anion transporting polypeptides belonging to the SLC family - OATP1B1, OATP1B3, and OATP2B1. In addition, we studied structural determinants of organic cation transporter 1 (OCT1) as another representative of the pharmaceutically relevant hepatic transporters. The research done in the framework of this thesis is structured into six individual studies (Study 1-6).

In Study 1, integrative data mining of OATP bioactivities from public databases was performed. Our intention was to analyze the substrate and inhibitor data with respect to data coverage, distribution of bioactivities, and to uncover enriched chemical substructures with pronounced selectivity profiles. Further, binary classification models were developed to identify important molecular features which were used to study commonalities and differences across the three OATPs. In Study 2, R-group decomposition of a congeneric series of analogs derived from 13-epiestrones was done to study the effect of different substituents on OATP2B1 inhibition. Presence of halogenated substituents at the R-2 position was identified as an important molecular determinant of OATP2B1 inhibition. In Study 3, we focused more on the methodological aspects of the thesis and developed an automated modeling pipeline for performing another emerging cheminformatic technique - ligand-based drug repurposing. The usefulness of the workflow was demonstrated for two case studies (GLUT1-deficiency syndrome and COVID-19).

In Study 4 we used structure-based modeling to shed light on differences in uptake of clinical substrates between human and mouse hepatic OCT1. Computational modeling attributed the differences between human and mouse OCT1 to hydrophobic packing interactions between TMH1 and TMH2. Study 5 involved signature dynamics of major facilitator superfamily proteins by normal mode analysis, generation of structural models for OATP1B1, OATP1B3, and OATP2B1 by ensemble docking, and the elucidation of binding mode hypotheses for compound derivatives possessing a steroidal scaffold. Differences in binding of steroids to the three transporters were attributed to different electrostatics and shape complementarity. In addition, several non-conserved residues in the N-terminal region provided structural insights into selectivity switches across the three transporters. Study 6 presents novel OATP inhibitors identified upon a combination of different computational approaches (structure-based virtual screening, conformational prediction, proteochemometric and deep learning models, respectively), which were subsequently validated by a transporter inhibition assay. By investigating binding modes of newly identified inhibitors we showed that the differences in the inner cavity across the three transporters were affected by different localization of aromatic residues.

In a biological context, the presented thesis ultimately contributed to the elucidation of molecular determinants of hepatic uptake transporters with a special focus on hepatic OATP-ligand interactions and selectivity. Last but not least, leveraging open data using (semi-)automated workflows was found to be a useful approach to increase the confidence of structure-based modeling approaches applied herein.

Zusammenfassung

In Hepatozyten exprimierte Aufnahmetransporter der SLC (sog. Solute Carrier) Superfamilie tragen zur Aufrechterhaltung der zellulären Homöostase bei, indem sie den Transport sowohl endogener Substrate als auch xenobiotischer Verbindungen regulieren. Eine beeinträchtigte Funktion dieser Transporter könnte zu klinisch relevanten Arzneimittel-Wechselwirkungen mit einer möglichen Entwicklung einer Arzneimittel-induzierten Leberschädigung führen. In dieser Arbeit verfolgten wir einen ganzheitlichen *in silico* Ansatz, der strukturbasierte Modellierung mit datenwissenschaftlichen Methoden (Cheminformatik) verbindet, um ein tiefgreifendes Verständnis der Transporter-Ligand-Wechselwirkungen zu erlangen. Der biologische Schwerpunkt der Arbeit lag auf hepatischen organischen Anionen transportierenden Polypeptiden der SLCO-Familie - OATP1B1, OATP1B3 und OATP2B1. Darüber hinaus untersuchten wir strukturelle Determinanten des organischen Kationentransporters 1 (OCT1) als weiteren Vertreter der pharmazeutisch relevanten Lebertransporter. Die im Rahmen dieser Arbeit durchgeführten Forschungsarbeiten gliedern sich in sechs Einzelstudien (Studie 1-6).

In Studie 1 wurde ein integratives Data Mining von OATP-Bioaktivitätsdaten aus öffentlichen Datenbanken durchgeführt. Unser Ziel war es, die Substrat- und Inhibitor-Daten hinsichtlich der Datenabdeckung und Verteilung der Bioaktivitätsmessungen pro Verbindung zu analysieren und häufig vorkommende Substrukturen mit ausgeprägten Selektivitätsprofilen aufzudecken. Des Weiteren wurden binäre Klassifizierungsmodelle entwickelt, um wichtige molekulare Eigenschaften zu identifizieren, die zur Untersuchung von Gemeinsamkeiten und Unterschiede zwischen den drei OATPs verwendet wurden. In Studie 2 wurde die R-Gruppen-Zerlegung einer Reihe von 13-Epiestronen abgeleitet

Analoga durchgeführt, um die Wirkung verschiedener Substituenten auf die OATP2B1-Aktivität zu untersuchen. Das Vorhandensein halogener Substituenten an der R-2-Position wurde als wichtige molekulare Determinante der OATP2B1-Hemmung identifiziert. In Studie 3 setzten wir den Schwerpunkt auf die methodischen Aspekte der Arbeit und entwickelten eine automatisierte Modellierungspipeline für die Durchführung einer weiteren neuen cheminformatischen Technik - einer auf Liganden-basierenden Wirkstoff Umwidmung (Drug Repurposing). Der Nutzen des Workflows wurde in zwei Fällen (GLUT1-Mangel-Syndrom und COVID-19) demonstriert. In Studie 4 verwendeten wir Strukturbasierte Modelle, um die unterschiedliche Aufnahme klinischer Substrate zwischen hepatischem OCT1 von Mensch und Maus zu beleuchten. Computermodele haben gezeigt, dass Unterschiede zwischen OCT1 von Mensch und Maus auf die hydrophoben Packungs-Wechselwirkungen zwischen transmembranärer Helix 1 und transmembranärer Helix 2 zurückzuführen wurden. Studie 5 befasste sich mit Signature Dynamics von Proteinen der Major Facilitator Superfamilie durch Normal Mode Analysis (NMA), die Erzeugung von Strukturmodellen für OATP1B1, OATP1B3 und OATP2B1 durch Ensemble-Docking und die Aufklärung von Bindungsmodushypothesen für Derivate mit Steroidgerüst. Unterschiede in der Steroidbindung zwischen den drei Transportern wurden auf unterschiedliche Elektrostatik und Formkomplementarität zurückgeführt. Darüber hinaus lieferten mehrere nicht konservierte Reste in der N-terminalen Region strukturelle Einblicke in Selektivitätsschalter zwischen den drei Transportern. In Studie 6 wurden neuartige OATP-Inhibitoren vorgestellt, die anhand einer Kombination verschiedener Berechnungsansätze (strukturbasiertes virtuelles Screening, Conformal Prediction, proteochemometrische- bzw. Deep-Learning-Modelle) identifiziert wurden, die anschließend durch den Transporter-Inhibitions-Assay validiert wurden. Durch die Untersuchung der Bindungsmodi neu gemessener Inhibitoren zeigten wir, dass die Unterschiede im inneren Hohlraum zwischen den drei Transportern auf die unterschiedliche Lokalisierung der aromatischen Reste zurückzuführen wurden.

In einem biologischen Kontext trug die vorgestellte Arbeit zur Aufklärung von strukturellen Aspekten von Lebertransportern bei, wobei ein besonderer Schwerpunkt auf den Wechselwirkungen und der Selektivität von hepatischen OATP-Liganden lag. Des Weiteren hat sich die Verwendung frei zugänglicher Daten (open data) mithilfe von (halb-) automatisierten Workflows als nützlicher Ansatz erwiesen, um das Vertrauen in die hier verwendeten strukturbasierten Modellierungsansätze zu erhöhen.

Part VI

SCIENTIFIC OUTPUT

3.7 Publications

Alzbeta Tuerkova, Brandon J. Bongers, Ulf Norinder, Orsolya Ungvári, Virág Székely, Csilla Özvegy-Laczka, Gergely Szakács, Gerard JP van Westen, Barbara Zdrazil. Combining AI-driven and structure-based approaches to identify novel inhibitors of hepatic organic anion transporting polypeptides (OATPs). **2020** *In preparation*

Alzbeta Tuerkova, Orsolya Ungvári, Erzsébet Mernyák, Gergely Szakács, Csilla Özvegy-Laczka, Barbara Zdrazil. Data-driven Ensemble Docking to Unravel Interactions of Steroid Analogs with Hepatic Organic Anion Transporting Polypeptides. **2020** *In preparation*

Alzbeta Tuerkova and Barbara Zdrazil. A ligand-based computational drug repurposing pipeline Using KNIME and programmatic data access: Case studies for rare diseases and COVID-19. *Journal of Cheminformatics* **2020**.

Available at <https://doi.org/10.1186/s13321-020-00474-z>

Marleen J. Meyer, Alzbeta Tuerkova, Sarah Römer, Christoph Wenzel, Tina Seitz, Jochen Gaedcke, Stefan Oswald, Jürgen Brockmöller, Barbara Zdrazil, Mladen V. Tzvetkov. Differences in metformin and thiamine uptake between human and mouse organic cation transporter OCT1: structural determinants and potential consequences for intrahepatic concentrations. *Drug Metabolism and Disposition*, **2020**.

Available at <https://doi.org/10.1124/dmd.120.000170>

Réka Laczkó-Rigó, Rebeka Jójárt, Erzsébet Mernyák, Éva Bakos, Alzbeta Tuerkova, Barbara Zdrazil, Csilla Özvegy-Laczka. Structural dissection of 13-epiestrones based on the interaction with human Organic anion-transporting polypeptide, OATP2B1. *The Journal of Steroid Biochemistry and Molecular Biology*, **2020**, 105652.

Available at <https://doi.org/10.1016/j.jsbmb.2020.105652>

Alžběta Türková and Barbara Zdrazil. Current advances in studying clinically relevant transporters of the solute carrier (slc) family by connecting computational modeling and data science. *Computational and Structural Biotechnology Journal*, **2019**, 17: 390-405.

Available at <https://doi.org/10.1016/j.csbj.2019.03.002>

Alžběta Türková, Sankalp Jain, Barbara Zdrazil. Integrative data mining, scaffold analysis, and sequential binary classification models for exploring ligand profiles of hepatic organic anion transporting polypeptides. *Journal of chemical information and modeling*, **2018**, 59.5: 1811-1825.

Available at <https://doi.org/10.1021/acs.jcim.8b00466>

Other publications

Alzbeta Tuerkova*, Ivo Kabelka*, Tereza Králová, Lukáš Sukeník, Šárka Pokorná, Martin Hof, Robert Vácha. Effect of helical kink in antimicrobial peptides on membrane pore formation. *eLife*, **2020**, 9.

Available at <https://doi.org/10.7554/eLife.47946>

* *Contributed equally.*

H M Meldal, Hema Bye-A-Jee, Lukáš Gajdoš, Zuzana Hammerová, Aneta Horáčková, Filip Melicher, Livia Perfetto, Daniel Pokorný, Milagros Rodriguez Lopez, Alžběta Türková, Edith D Wong, Zengyan Xie, Elisabeth Barrera Casanova, Noemi del-Toro, Maximilian Koch, Pablo Porras, Henning Hermjakob, Sandra Orchard. Complex Portal 2018: extended content and enhanced visualization tools for macromolecular complexes. *Nucleic acids research*, **2019**, 47.D1: D550-D558.

Available at <https://doi.org/10.1093/nar/gky1001>

3.8 Selected Talks

Alzbeta Tuerkova, Sankalp Jain, Csilla Özvegy-Laczka, Gergely Szakács, Ulf Norinder, Gerard JP van Westen, Brandon J Bongers, Barbara Zdrazil. Multiscale Computer-based Studies to Identify Ligand Interactions and Selectivity Among Hepatic Organic Anion Transporting Polypeptides. In *EUROPIN Summer School on Drug Design*, **15-20 September 2019**, Vienna, Austria.

Alzbeta Tuerkova, Sankalp Jain, Virág Székely, Csilla Özvegy-Laczka, Gergely Szakács, Ulf Norinder, Barbara Zdrazil. Linking multiscale data analyses, ligand- and structure-based modeling to explore ligand interactions with hepatic organic anion transporting polypeptides. In *ACS National Meeting & Expo*, **25-29 August 2019**, San Diego, CA, USA.

Alzbeta Tuerkova, Barbara Zdrazil. Exploring Conformational Space of Hepatic Uptake Transporters via Normal Mode Simulations In *Gordon Research Seminar on Computer Aided Drug Design - Combining Artificial Intelligence and Physics-Based Modeling for Small- and Macromolecular Drug Design*, **13-14 July 2019**, West Dover, VT, USA.

3.9 Selected Posters

Alzbeta Tuerkova, Barbara Zdrazil. Systematic Pipeline for Automated Structure-based Molecular Design on a Large Scale: Beyond the Static Picture of Hepatic Organic Anion Transporting Polypeptides. In *ACS National Meeting & Expo*, **25-29 August 2019**, San Diego, CA, USA.

Alzbeta Tuerkova, Barbara Zdrazil. Interconnecting Large-scale Data Analyses and Ensemble Docking for Elucidating Hepatic-OATP Ligand Interactions and Selectivity. In *Gordon Research Conference on Computer Aided Drug Design - Integrating Big Data and Macromolecular Protein Structures into Small Molecule Design*, **14-19 July 2019**, West Dover, VT, USA.

Alzbeta Tuerkova, Barbara Zdrazil. Application of Normal Mode Simulations to Enhance Conformational Sampling of Hepatic Organic Anion Transporting Polypeptides. In *Molecular Dynamics Today*, **14-15 March 2019**, Bologna, Italy.

Alzbeta Tuerkova, Barbara Zdrazil. Combining Ligand-based and Structure-based Approaches for Exploring Ligand Selectivity among Hepatic Organic Anion Transporting Polypeptides. *In 14th German Conference on Cheminformatics, 11-13 November 2018*, Mainz, Germany.

Alzbeta Tuerkova, Sankalp Jain, Barbara Zdrazil. Exploring Ligand Selectivity Among Hepatic Organic Anion Transporting Polypeptides via Data Fusion, Scaffold Analysis, and Sequential Binary Classification Modelling. *In 22nd European Symposium on Quantitative Structure-Activity Relationships, 16-20 September 2018*, Thessaloniki, Greece.

3.10 Awards

CINF Scholarship for Scientific Excellence. The scholarship program of the Division of Chemical Information (CINF) of the American Chemical Society (ACS) is designed to reward graduate and postdoctoral students in chemical information and related sciences for scientific excellence and to foster their involvement in CINF. *Award received on 25 August, 2019*

3.11 Teaching

Lecturer of the course about science approaches for drug discovery, a part of "Experimental Methods in Drug Discovery and Preclinical Drug Development" course based at University of Vienna. Topics covered by the course: Introduction to KNIME workflows; programmatic access to life-science databases; database mining and cross-referencing through API requests; extraction of Murcko Scaffolds; cheminformatics analysis; clustering of ligands on basis of similarity; substructure screening of DrugBank; integrative data mining of ligand bioactivities from ChEMBL and PubChem. Department of Pharmaceutical Chemistry, *Summer semester 2020*

Teaching materials (input data, KNIME workflows, step-by-step tutorials) available at:

<https://github.com/AlzbetaTuerkova/Drug-Repurposing-in-KNIME>