Heinrich Heine Universität Düsseldorf

Physics- and data science-based computational predictive modeling to investigate molecular mechanisms in cholestatic liver diseases

Inaugural-Dissertation

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

vorgelegt von
Annika Behrendt
aus Witten

Düsseldorf, Mai 2024

Aus dem Institut für Pharmazeutische und Medizinische Chemie
der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Berichterstattende:
1. Prof. Dr. Holger Gohlke
2. Prof. Dr. med. Verena Keitel-Anselmino

Tag der mündlichen Prüfung:
26. August 2024

Eidesstattliche Erklärung

Ich, Annika BEHRENDT, versichere an Eides Statt, dass die Dissertation von mir selbstständig und ohne unzulässige fremde Hilfe unter Beachtung der „Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität Düsseldorf“ erstellt worden ist.

Diese Dissertation wurde in der vorgelegten oder einer ähnlichen Form noch bei keiner anderen Institution eingereicht, und es wurden bisher keine erfolglosen Promotionsversuche von mir unternommen. Diese Dissertation wurde ohne Verwendung von Künstlicher Intelligenz geschrieben.

Düsseldorf, Mai 2024

Table of Contents

List of Figures

List of Abbreviations

ABC	ATP-binding cassette
ABCB4	ABC subfamily B member 4
ABCG5/G8	ABC-transporter G5/G8
ACMG-AMP	American College of Medical Genetics and Association for Molecular Pathology
AF1	Activation function domain 1
AF2	Activation function domain 2
AMBER	Assisted Model Building with Energy Refinement
APL	Aminophospholipid
ASBT	Apical sodium-dependent bile acid transporter
ATP	Adenosine triphosphate
AUC	Area under the curve
BA	Bile acid
BS	Bile salt
BSEP	Bile salt export pump
CA	Cholic acid
CDCA	Chenodeoxycholic acid
CoA	Coenzyme A
CYP7A1	Cytochrome P450 family 7 subfamily A member 1
DBD	DNA-binding domain
DCA	Deoxycholic acid
DILI	Drug-induced liver injury
DR	Direct repeat sequences
ER-2	Everted repeat with 2-base pair spacer
ERAD	ER-associated degradation
ESP	Electrostatic potential
FGF15	Fibroblast growth factor 15
FGF19	Fibroblast growth factor 19
FGFR4	FGF receptor 4
FIC1	Familial intrahepatic cholestasis 1
FN	False negative
FP	False positive
FXR	Farnesoid X receptor
FXRE	FXR response element
GAFF	General AMBER force field
gnomAD	Genome Aggregation Database
GPBAR1	G-protein coupled bile acid receptor
H12	Helix 12
HCC	Hepatocellular carcinoma
HNF4α	Hepatocyte nuclear factor 4α
HSE	Half-sphere exposure
IBABP	Ileal bile acid-binding protein
IBAT	Ileum bile acid transporter
ICP	Intrahepatic cholestasis of pregnancy
IR-1	Inverted repeat with 1-base pair spacer
JNK	c-Jun N-terminal kinase
KIF12	Kinesin family member 12
LBD	Ligand binding domain
LCA	Lithocholic acid
LPAC	Low phospholipid associated cholelithiasis

LRH-1	Liver receptor homolog-1
LXR	Liver X receptor
MCC	Matthew's correlation coefficient
MD	Molecular dynamics
MDR3	Multi-drug response protein 3
Mg	Magnesium
ML	Machine learning
MRP2	Multidrug resistance protein 2
MYO5B	Myosin 5B
NAFLD	Non-alcoholic fatty liver disease
NASH	Non-alcoholic steatohepatitis
NBD	Nucleotide binding domain
NCoA2	Nuclear receptor coactivator 2
NCoR	Nuclear corepressor protein
NF-κB	Nuclear factor-κB
NR	Nuclear Receptor
NR1H4	Nuclear receptor subfamily 1 group H member 4
NTCP	Sodium taurocholate cotransporting polypeptide
OATP1B1	Organic anion transporting polypeptide 1 B1
OCA	Obeticholic acid
OSTα/β	Organic solute transporter alpha/beta
PBC	Primary biliary cholangitis
PC	Phosphatidylcholine
PDB	Protein Data Bank
PFIC	Progressive familial intrahepatic cholestasis
PFIC3 / PFIC5	Progressive familial intrahepatic cholestasis type 3 / type 5
P-gp	P-glycoprotein
PPARγ	Peroxisome proliferator-activated receptor γ
PTEN	Phosphatase and tensin homolog
PTM	Post-translational modification
QM	Quantum mechanics
RESP	Restrained electrostatic potential
RMSD	Root-mean-square deviation
ROC	Receiver operating characteristics
RORγ	Retinoic-acid related-orphan-receptor-C
RSA	Relative solvent accessibility
RXR	Retinoid X receptor
SBARM	Selective bile acid receptor modulator
SHP	Small heterodimer partner
SMOTE	Synthetic minority oversampling technique
SNP	Single nucleotide polymorphism
TGR5	Takeda G protein-coupled receptor 5
TJP2	Tight junction protein 2
TMD	Transmembrane domain
TMH	Transmembrane helix
TN	True negative
TP	True positive
Vasor	Variant assessor of MDR3
WT	Wildtype
XGBoost	Extreme gradient boosting

List of Publications

The presented thesis is based on the following publication and manuscript currently within peer-review:

Behrendt, A., Golchin, P., König, F., Mulnaes, D., Stalke, A., Dröge, C., Keitel, V., Gohlke, H.
Vasor: Accurate prediction of variant effects for amino acid substitutions in multidrug resistance protein 3.
Hepatology Communications (2022), Nov, Vol. 6, Issue 11, p 3098-3111. DOI: 10.1002/hep4.2088
Impact Factor reported in 2022 (Journal Citation Reports): 5.1

Behrendt, A., Stindt, J., Pfister, E.-D., Grau, K., Brands, S., Dröge, C., Stalke, A., Bonus, M., Sgodda, M., Cantz, T., Bastianelli, A., Baumann, U., Keitel, V., Gohlke, H.
Impaired transitioning of the FXR ligand binding domain to an active state underlies a PFIC5 phenotype.
Available as preprint at bioRxiv (2024), Feb. DOI: 10.1101/2024.02.08.579530

During the work for the thesis, these additional publications were co-authored:

Dröge, C., Götze, T., **Behrendt, A.**, Gohlke, H., Keitel, V.
Diagnostic workup of suspected hereditary cholestasis in adults: a case report.
Exploration of Digestive Diseases (2023), Apr, Vol. 2, p 34–43. DOI: 10.37349/edd.2023.00016
Contribution: *in silico* prediction, writing, review, and editing
Impact Factor: not defined yet (status: May 2024)

Stalke, A., **Behrendt, A.**, Hennig, F., Gohlke, H., Buhl, N., Reinkens, T., Baumann, U., Schlegelberger, B., Illig, T., Pfister, E. D., Skawran, B.
Functional characterization of novel or yet uncharacterized ATP7B missense variants detected in patients with clinical Wilson's disease.
Clinical Genetics (2023), Aug, Vol. 104, Issue 2, p 174-185. DOI: 10.1111/cge.14352
Contribution: *in silico* analyses, free folding energy calculations, writing, review, and editing.
Impact Factor reported in 2022 (Journal Citation Reports): 3.5

Zusammenfassung

Eine der zentralen Funktionen der Leber ist die Produktion und Erhaltung der Gallenflüssigkeit. Zuzüglich zu der seit langem etablierten Rolle in der Fettabsorption wurden Gallensäuren vor Kurzem als Signalmoleküle identifiziert, die das Darmmikrobiom und wichtige zelluläre Signalwege beeinflussen. Fehlregulierte Gallenhomöostase aufgrund von genetischen Änderungen in Schlüsselproteinen in Hepatozyten ist ein Kennzeichen von cholestatischen Erkrankungen wie der progressiven familiären intrahepatischen Cholestase (PFIC). Die Phospholipid-Floppase MDR3, die sich in der kanalikulären Membran befindet und für den Transport von Phosphatidylcholin und damit für die Aufrechterhaltung eines nicht-toxischen Verhältnisses von Lipiden zu Gallensäuren in den Mischmizellen der Galle verantwortlich ist, ist bei PFIC Typ 3 betroffen. Aminosäuresubstitutionen stellen den größten Teil der bei PFIC3-Patienten identifizierten ursächlichen Veränderungen innerhalb des *ABCB4* Gens (das für das MDR3 Protein kodiert) dar. In der Publikation I habe ich ein maschinelles Lernen-basiertes Programm entwickelt, welches Varianten als benigne oder pathogen klassifizieren kann, um Kliniker und Wissenschaftler bei der Einschätzung von neuen Varianten für weitere Testungen zu unterstützen. Da MDR3 an einer Reihe von Lebererkrankungen beteiligt ist, lässt sich das Programm auf jede MDR3-Variante anwenden. MDR3 wird, wie viele andere Proteine innerhalb des komplexen Netzwerks zur Regulierung der Gallenhomöostase, durch den Nuklearen Rezeptor FXR transkriptionell reguliert. Eine klinisch identifizierte homozygote Missense-Variante, die mit PFIC Typ 5 assoziiert ist, wurde innerhalb der Publikation II mittels einer Kombination von *in vitro* und *in silico* Ansätzen analysiert, um den molekularen Mechanismus zu entschlüsseln. Die Variante, lokalisiert innerhalb der Ligandenbindungs-domäne, beeinflusst die Positionierung von Helix 12, welche entscheidend für die Proteinaktivität ist. Die Variante zeigte einen reduzierten Übergang vom inaktiven zum aktiven Zustand, passend zur verringerten Transkriptionsaktivität in zellulären Assays. Darüber hinaus könnte der enthüllte Übergang zwischen den Konformationszuständen des Wildtyp-FXR Proteins eine Grundlage für zukünftige neue Erkenntnisse im Bereich der spezifischen Targeting-Strategien bieten. FXR hat vielfältige Funktionen innerhalb des menschlichen Körpers, und isoform-, gewebe-, und ligandenspezifische Effekte legen nachgeschaltete Ziele auf der Gen-Ebene fest. Trotz der Komplexität und des inhärenten Risikos von Nebenwirkungen bleibt die pharmakologische Intervention mittels FXR von hohem Interesse

aufgrund der Beteiligung im Fett- und Glukosestoffwechsel, Entzündungen und Immunität, sowie der Gallenhomöostase und der Verbindung zum Mikrobiom. Dementsprechend muss die sichere Beeinflussung von FXR auf einem detaillierten Verständnis der Protein-Dynamik basieren. In der vorliegenden Arbeit stelle ich ein neues, verlässliches Vorhersageprogramm für die Pathogenität von MDR3-Varianten und Erkenntnisse in die Regulierung der FXR-Aktivität vor.

Abstract

One of the central functions of the liver is the production and maintenance of bile. In addition to their long-established role in fat absorption, bile acids have recently been identified as signaling molecules able to influence the gut microbiome and major cellular pathways. Dysregulated bile homeostasis due to genetic alterations in key protein players within hepatocytes is a hallmark of cholestatic diseases such as progressive familial intrahepatic cholestasis (PFIC). The phospholipid floppase MDR3, located at the canalicular membrane and responsible for transporting phosphatidylcholine and thus maintaining non-toxic lipid to bile salt ratios within bile mixed micelles, is impacted within PFIC type 3. Missense amino acid substitutions represent the majority of causative alterations within the *ABCB4* gene (encoding for MDR3 protein) identified in PFIC3 patients. In Publication I, I developed a machine learning program to classify variants as benign or pathogenic, thus assisting clinicians and researchers in the assessment of novel variants for further testing. Due to the involvement of MDR3 in a range of liver diseases, the tool is applicable to any MDR3 variant. MDR3, like many other proteins involved in the complex network maintaining bile homeostasis, is transcriptionally regulated by the nuclear receptor FXR. A clinically identified homozygous missense variant associated with PFIC type 5 was analyzed within Publication II using a combination of *in vitro* and *in silico* approaches to unravel the molecular mechanism. Located within the ligand binding domain, the variant impacts the positioning of helix 12, which is critical for protein activity. The variant showed reduced transitioning from the inactive to active state, in line with reduced transcriptional activity in cellular assays. Additionally, the uncovered transitioning between conformational states in the wildtype FXR protein may provide a basis for novel insights into specific targeting strategies. FXR has broad functions within the human body, and isoform-, tissue-, and ligand-specific effects determine downstream gene targets. Despite the complexity and inherent risk of side effects, pharmacological targeting of FXR remains of high interest due to its involvement in lipid and glucose metabolism, inflammation, and immunity, as well as bile homeostasis and its microbiome linkage. Accordingly, safely targeting FXR must be grounded in a thorough understanding of its protein dynamics. Within the presented work, I provide a novel, reliable prediction tool for the pathogenicity of MDR3 variants and insights into FXR activity regulation.

Chapter 1 Introduction

Guided by evolutionary processes leading to an astounding plethora of cell diversity and cellular mechanisms, the human body functions as a deeply connected network of specialized organs and tissues (Alberts et al., 2007; Asada et al., 2019; Bartsch et al., 2015). Derived from a single cell, epigenetic changes and signaling networks enable cells to differentiate into specialized cell types within organs performing carefully adjusted and regulated functions (Alberts et al., 2007). Due to the high interconnectivity between cells and, on the higher level, organs, a misfunction can lead to imbalances in connected systems. As such, many diseases, while potentially originating in a specific location within the body, have implications for other organs and show debilitating effects beyond the direct reach of the affected cell area. The liver is the main site of impairment in progressive familial intrahepatic cholestasis (PFIC), a heterogenous group of rare disorders (Clayton, 1969; Davit-Spraul et al., 2009; Prescher et al., 2019). These genetic disorders impact the ability of hepatocytes to properly form and secrete bile, leading to early-onset progressive liver disease (Davit-Spraul et al., 2009; Gomez-Ospina et al., 2016; Gonzales et al., 2017; Sambrotta et al., 2014). Beyond the liver, the expression of functionally impaired proteins in other organs, as well as altered bile properties, which in turn affect the microbiome interactions, can lead to the involvement of other organs, particularly the intestinal system, exemplifying a strong gut-liver axis (Pfister et al., 2022; Yu et al., 2023).

Within affected patients, genetic analysis often identifies alterations leading to amino acid missense variants within relevant hepatocyte proteins. Cellular assays to unravel variant impact are time- and cost-intensive, and accordingly, computational methods such as machine learning (ML) or molecular dynamics (MD) simulations have been increasingly employed to aid the evaluation of variant effects. ML, a computational technique to extract underlying patterns from datasets and extrapolate to novel data, is increasingly impacting science from basic research to clinical applications (Greener et al., 2022; Iqbal et al., 2021; Stormo et al., 1982; Yip et al., 2017). In the context of protein missense variants, in which single amino acid positions are exchanged, ML approaches resulted in a range of prediction tools for mutational impact and are routinely used to guide researcher efforts (Adzhubei et al., 2010; Choudhury et al., 2022; Frazer et al., 2021; Livesey & Marsh, 2023). MD simulations allow the study of protein motions on an atomic level over time in a controlled computational model system and have proven useful in deciphering protein dynamics (Latorraca et al., 2017; Prescher et al.,

2021), protein-protein (Koch et al., 2019), protein-ligand (Bonus et al., 2020), or protein-nucleic acid interactions (Yoo et al., 2020).

Within the presented thesis, I established an ML tool and employed MD simulations to analyze PFIC-relevant proteins, namely multidrug resistance protein 3 (MDR3) and farnesoid X receptor (FXR). Based on a unique MDR3 dataset, I derived a protein-specific ML prediction tool for MDR3 missense variants to classify variants into the categories of benign or pathogenic. To ensure easy access, the tool is available as a webserver as well as a standalone version (Publication I). Further, I uncovered conformational transitioning from the inactive to active state for FXR using MD simulations and revealed a decreased transitioning for a clinically identified variant, explaining its reduced protein activity (Publication II). Publication II is currently in the peer-review process (status: May 2024).

1.1 Bile homeostasis as a cornerstone of liver health

The human liver is responsible for a variety of functions within the body, including lipid uptake and secretion, cholesterol homeostasis, generation of signaling molecules, glucose metabolism, bile formation and secretion (Knell, 1980; Ma et al., 2006; Trefts et al., 2017). Bile is necessary for the emulsification and absorption of fat and fat-soluble vitamins within the digestive system (Di Gregorio et al., 2021). Furthermore, it is needed in the elimination of potentially harmful exogenous toxins and endogenous lipophilic substances such as bilirubin, while it is also the main elimination route for cholesterol (Boyer, 2013; Hofmann, Alan, 2009). Bile consists of water, bile salts, phospholipids, cholesterol, bilirubin, cations and anions, and other proteins, amino acids, and vitamins in smaller traces (Boyer, 2013). Within the enterohepatic circulation, bile acids are actively absorbed and transported back to the liver (**Figure 1**) (Hofmann, Alan, 2009; Hofmann, 1976). Upon ingested food reaching the stomach (**Figure 1**, I), the gallbladder starts to release stored bile into the primary part of the small intestinal tract, the duodenum (**Figure 1**, II). Here, the bile acts on the ingested lipids and fat-soluble vitamins, preventing the formation of large fat droplets while facilitating easier enzymatic attack on lipids based on an increased surface available for the enzymes in smaller fat droplets (Di Gregorio et al., 2021). Over the course of the small intestinal tract, the bile acids are absorbed, mostly in the ileal part, via the ileum bile acid transporter (IBAT) expressed in enterocytes (Dawson et al., 2003). Transported via the portal vein and mixed in the liver

with oxygen-rich blood from the hepatic artery within the sinusoids, the bile acids reach their original site of production, the hepatocytes (**Figure 1**, III). The enterohepatic circulation recycles in healthy state around 95% of the bile acids (Halilbasic et al., 2013; Hofmann, Alan, 2009). While this is in general beneficial in terms of energetic costs, it also allows a direct feedback mechanism, in which bile acid levels can be sensed and re-adjusted if necessary. This feature is essential to control functioning fat digestion as well as to guard against critically elevated bile acid levels, as high levels of bile acids can act on cell membranes, leading to cell toxicity (Ikeda et al., 2017; Oude Elferink & Paulusma, 2007). Their unique features justify taking a closer look at these molecules.

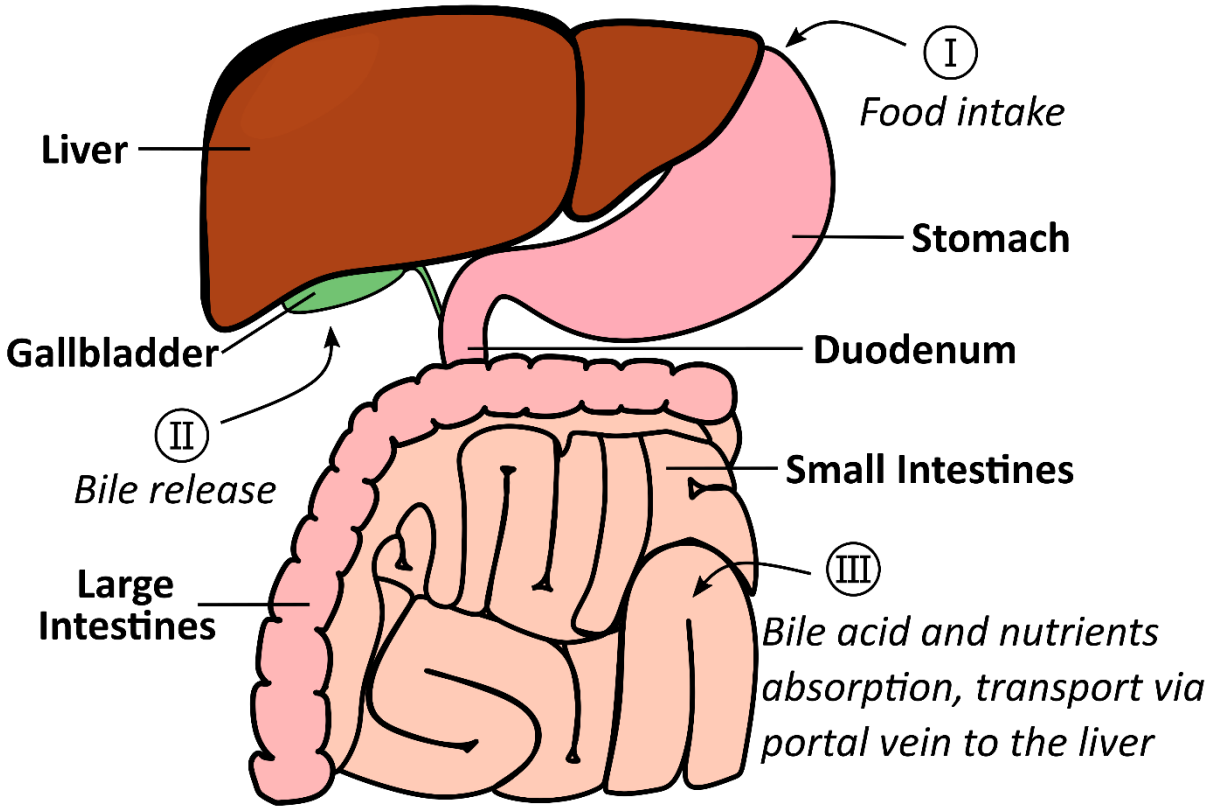


Figure 1: Overview of the enterohepatic circulation. Bile is produced within the liver and stored via bile canaliculi in the gallbladder. After food intake (I), bile is released from the gallbladder (II), entering the intestinal system at the Duodenum, the first of the three sections of the small intestines. In healthy state, 95% of liver-secreted bile acids are absorbed within the small intestines (III), especially from the distally located ileal epithelial cells. The portal vein transports absorbed bile acids and nutrients via the liver to other body parts, where bile acids get taken up by hepatocytes, returning to their production site.

1.2 The unique structure of bile acids

Bile acids are generated from cholesterol (A. E. C. Wen & Campbell, 1977), and thus, all bile acids share the common structure of the cholane skeleton with three six-membered and one five-membered carbon rings (**Figure 2**, A). Within humans, chenodeoxycholic acid (CDCA) and cholic acid (CA) are the two main bile acids synthesized from cholesterol (Vlahcevic et al., 1991), which are conjugated with taurine or glycine before secretion into bile (Chiang, 2013). Conjugation, performed by the enzymes bile acid-Coenzyme A (CoA) ligase or bile acid-CoA:amino acid N-acyltransferase, occurs at the side chain and reduces the molecules' pKa (Di Gregorio et al., 2021). Consequently, conjugated bile acids will be present in their ionized salt form (accordingly termed bile salts), thus lowering their passive absorption through cellular membranes within the intestinal tract. These molecules, synthesized within hepatocytes, are referred to as primary bile acids. Overall, their structure results in a hydrophilic and a lipophilic molecule side (**Figure 2**, B). This aids in fat emulsification, preventing fat droplets from converging and enabling easier access for attacking enzymes to break down the fats (**Figure 2**, C). Thus, bile salts have an important role in our digestive system, facilitating the absorption of lipids and lipophilic vitamins. Furthermore, the gut microbiome actively contributes to the diversity by converting the primary bile acids into a variety of secondary bile acids (S. L. Collins et al., 2023; Guzior & Quinn, 2021; Quinn et al., 2020).

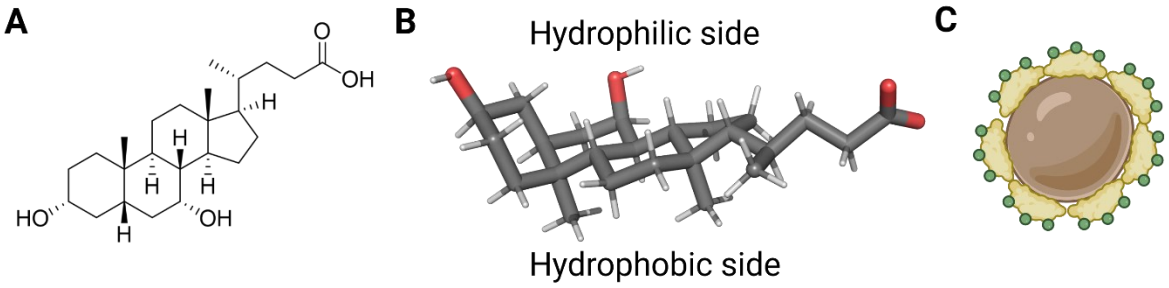


Figure 2: Structure of the bile acid CDCA. [A] Structure of CDCA. [B] Three-dimensional view of a conformation of CDCA in its ionized form, extracted from a crystal structure of FXR bound with agonistic CDCA (PDB ID: 6HL1 (Merk et al., 2019)) and depicted with PyMOL. Oxygen atoms are depicted in red, carbon atoms in grey and hydrogens in white. [C] Emulsifying effect of bile acids with a central fat droplet (brown sphere). The structure of bile acids, with hydrophilic charged groups (green points) pointing towards the surrounding hydrophilic environment and its hydrophobic side (yellow area) interacting with lipids from the fat droplet, preventing the coalescence of droplets. Panel C was created using BioRender.

Modifications range from deconjugation, dehydroxylation, oxidation and epimerization to reconjugation, and the levels of specific secondary bile acids differ not only from person to person based on the individual microbiome, but also along the intestinal tract (Guzior & Quinn, 2021; Shalon et al., 2023). In general, this metabolism by the microbiota has been long established (Gustafsson et al., 1966). Given the abundance and diversity of the microbial community, researchers have long been striving to unravel the composition and interaction patterns in greater detail. Recent advances both in molecular (Blaut et al., 2002; Hillman et al., 2017) and sampling techniques (Shalon et al., 2023) are enabling a new era, introduced with the breakthrough finding of novel microbially conjugated bile acids, namely the conjugation of amino acids phenylalanine, tyrosine and leucine to bile acids (Quinn et al., 2020). As such, the field of microbiome research and the interplay with bile acids and their role in healthy and disease states is currently under intense investigation, and a series of novel important insights are expected (S. L. Collins et al., 2023; Guzior & Quinn, 2021; Shalon et al., 2023). Additionally, bile acids have been identified as important signaling molecules for lipid and glucose metabolism (de Aguiar Vallim et al., 2013; Ma et al., 2006), inflammation (M. Li et al., 2017), and immunity (Fiorucci et al., 2018; Godlewska et al., 2022). Another key aspect of bile acids, and one that will be further discussed in the next chapter, is the bile acid feedback loop regulating its own homeostasis.

1.3 Hepatocytes control the bile formation

For a well-functioning fat digestion, the ability to sense and adjust bile formation is important. Bile salts produced in hepatocytes and released into the enterohepatic circulation enable this feedback mechanism (**Figure 1** and **Figure 3**). Another reason for tight control of bile salt levels is their cytotoxic effects due to their detergent nature and their potential to induce proinflammatory stimuli at higher concentrations (Claudel & Trauner, 2020; Ikeda et al., 2017; M. Li et al., 2017). Bile acids entering from the portal vein can act as agonists for FXR, the central transcription factor of the bile formation network (Jiang et al., 2021; H. Wang et al., 1999). Additionally, FXR activation by elevated bile acid levels in enterocytes can influence hepatocyte metabolism, for example, via fibroblast growth factor 19 (FGF19), expressed within the enterocyte and traveling via the portal vein to suppress the production of bile acids within hepatocytes (Katafuchi & Makishima, 2022). Upon bile acid binding, FXR translocates into the nucleus, dimerizes with the transcription factor retinoid X receptor (RXR) and binds

to its specific DNA response element (Claudel et al., 2002; Forman et al., 1995; Laffitte et al., 2000). In this way, FXR can exert control over the synthesis of bile acids from cholesterol via the rate-limiting enzyme cytochrome P450 family 7 subfamily A member 1 (CYP7A1), inhibit further uptake of bile acids from the portal vein via inhibition of the sodium taurocholate cotransporting polypeptide (NTCP) and increase the efflux of bile acids from hepatocytes into the portal vein via the heterodimer organic solute transporter alpha/beta (OSTα/β) in order to avoid reaching toxic levels of bile acids within the cell (Chiang et al., 2000; Claudel et al., 2002; Dash et al., 2017; Hoeke et al., 2009). The expression of organic anion transporting polypeptide 1 B1 (OATP1B1) has been found to be induced by FXR and liver X receptor (LXR) in a hepatoma-derived cell line (Meyer zu Schwabedissen et al., 2010). Previous studies have established downregulation of OATP1B1 within PFIC type 2 and type 3 (Keitel et al., 2005) and repression of both OATP1B1 and OATP1B3 in CDCA-treated human liver slices (Jung et al., 2007), potentially to protect hepatocytes from toxic intracellular bile acid levels. Additionally, FXR drives protein transporter expression necessary for the transport of bile components into the canaliculi. The most prominent and well-studied example is the bile salt export pump (BSEP), which is under FXR-regulated promotor control (Ananthanarayanan et al., 2001; Dash et al., 2017). This ATP-binding cassette (ABC) transporter translocates bile salts against a concentration gradient from within the hepatocytes through the canalicular membrane into the canaliculi (Gerloff et al., 1998; Strautnieks et al., 1998). Mixed micelles with lipids such as phosphatidylcholine (PC) are formed in the bile canaliculi, preventing detergent effects of the bile salts on the cell membranes (Ikeda et al., 2017; Oude Elferink & Paulusma, 2007). PC is flopped from the inner canalicular membrane for extraction into bile micelles by another ABC transporter, the multidrug resistance protein 3 (MDR3, gene name *ABCB4*) (Olsen et al., 2020; Prescher et al., 2021; A. J. Smith et al., 1994). Like BSEP, MDR3 is a FXR-regulated target (Dash et al., 2017; L. Huang et al., 2003; Ijssennagger et al., 2016). However, MDR3 expression was not fully abrogated in patients with loss of function FXR variants (Gomez-Ospina et al., 2016), indicating a more complex transcriptional regulation.

For a healthy bile formation, other key proteins have been identified (**Figure 3**). Due to ongoing research efforts, which are often guided by clinical screening of patients and rigorous sequencing, further proteins are regularly identified and could be extending this list of involved proteins (e.g., the microtubule motor protein Kinesin family member 12 (KIF12) (Maddirevula et al., 2019; Stalke et al., 2022)). Within the intracellular transportation

machinery, apical targeting of membrane proteins such as BSEP and MDR3 is a prerequisite for bile formation. The cytoskeleton motor protein myosin 5B (MYO5B), important for epithelial cell polarization and vesicular trafficking, has been shown to be required for proper BSEP localization (Müller et al., 2008). Mutations in MYO5B have been associated with microvillus inclusion disease, characterized by loss of microvilli on enterocytes' surface, and with PFIC type 6 (Gonzales et al., 2017; Müller et al., 2008; Qiu et al., 2017). The aminophospholipid flippase familial intrahepatic cholestasis 1 (FIC1) is responsible for maintaining the membrane asymmetry at the canalicular membrane, and mutations have been associated with liver diseases such as PFIC type 1 (Eppens et al., 2001; Paulusma et al., 2006). Variants leading to dysfunction of the tight junction protein 2 (TJP2) have been associated with PFIC type 4 (Sambrotta et al., 2014). TJP2 is a scaffolding protein involved in establishing tight junctions through interaction with cytoskeletal proteins and integral membrane proteins such as Claudin proteins (Carlton et al., 2003) and despite widespread expression, TJP2 dysfunction might impact mainly the liver due to the specific environment with high exposure of tight junctions to detergent bile salts (Sambrotta et al., 2014; Sambrotta & Thompson, 2015). Within the liver, the heterodimer ABCG5/G8 is responsible for cholesterol secretion (Graf et al., 2003) and the ATP transporter multidrug resistance protein 2 (MRP2), besides its function in detoxification, transports bilirubin into the bile canaliculi (Jedlitschky et al. 1997; Gabriele Jedlitschky, Hoffmann, and Kroemer 2006).

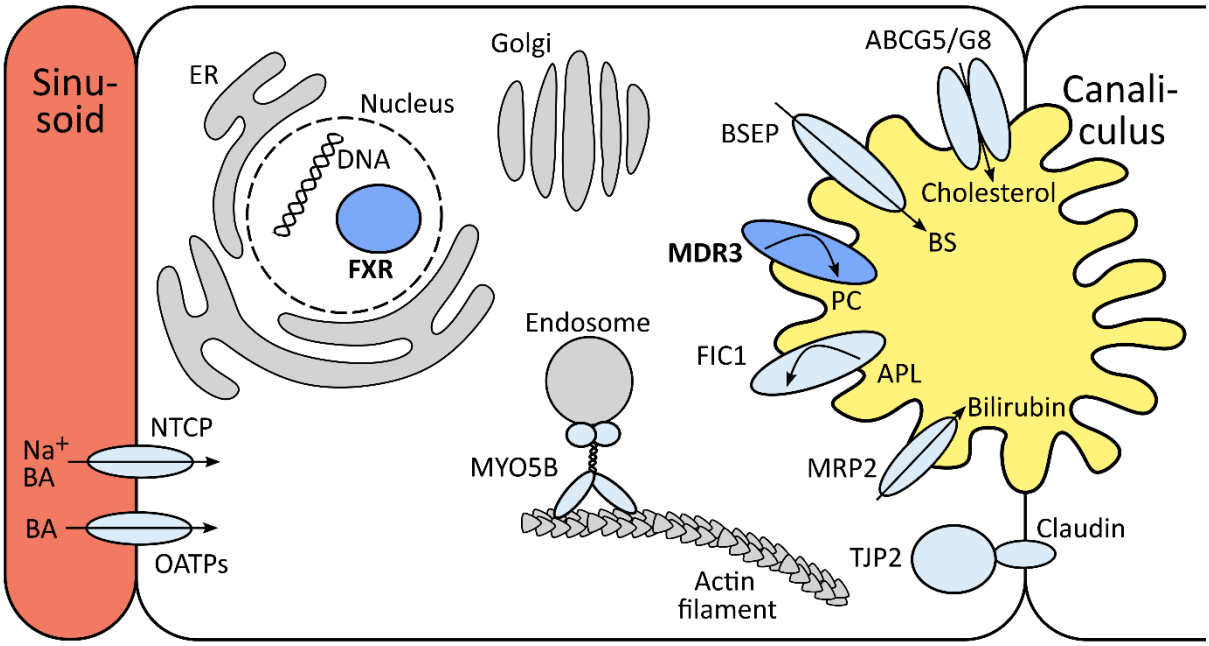


Figure 3: Overview of important proteins within a hepatocyte involved in the bile formation and enterohepatic cycle. FXR and MDR3 proteins (dark blue) are marked in bold as they are the focus of this thesis. Other key players (light blue) are the bile salt export pump (BSEP), familial intrahepatic cholestasis 1 (FIC1), the heterodimer ABC transporter G5/G8 (ABCG5/G8), multidrug resistance protein 2 (MRP2) [all located within the canalicular membrane], as well as tight junction protein 2 (TJP2), myosin 5B (MYO5B) [cytosolic proteins] and sodium taurocholate cotransporting polypeptide (NTCP) and organic anion transporting polypeptide (OATP) [located in the basal membrane]. Straight arrows indicate molecule transport directions and curved arrows indicate molecule flipping or flopping within the membrane bilayer. ER: endoplasmatic reticulum, BA: bile acids, BS: bile salts, PC: phosphatidylcholine, APL: aminophospholipid. Distantly adapted from Pfister et al., 2022, Figure 1 and Dröge et al., 2017, Figure 1. Further, Figure 7 and Figure 10 in this dissertation are based on Figure 3.

Destabilization of this tightly regulated system can lead to pathological effects and liver disease, from less severe impacts like intrahepatic cholestasis of pregnancy (ICP) to severe diseases like PFIC. The affected patient numbers are low in accordance with its classification as rare diseases, but disease severity often necessitates liver transplantation (Srivastava, 2014). Further, the different PFIC subtypes highlight the interplay of proteins involved within normal bile formation. The BMBF-funded consortium HiChol follows a multi-disciplinary approach to study phenotypes, molecular causes, effects, and treatment options of PFIC diseases based on the genetic analyses of individual patients. My work was performed as part of the HiChol consortium. The combination of *in vitro*, *in vivo* and *in silico* studies, together with patient and clinical data, enables a holistic view with the goal of advancing the basic understanding of molecular mechanisms and improving patients' quality of life. In this context, I have focused on providing classification guidance for novel variants of the MDR3

protein (Publication I) and on elucidating the molecular mechanism of a missense variant of the FXR protein (Publication II).

1.3.1 Variant classification in MDR3

The adenosine triphosphate (ATP)-binding cassette (ABC) subfamily B member 4 (ABCB4, also known as MDR3) is almost exclusively expressed within the liver (Sticova & Jirsa, 2020; Uhlén et al., 2015; Van der Bliek et al., 1987). MDR3 dysfunction has been associated with a wide range of liver diseases with varying severities, such as ICP, drug-induced liver injury (DILI), low phospholipid-associated cholelithiasis (LPAC), liver fibrosis, liver cirrhosis as well as hepatobiliary malignancy and progressive familial intrahepatic cholestasis type 3 (PFIC3) (Deleuze et al., 1996; Dixon et al., 2000; C. Dong et al., 2020; Dröge et al., 2017; Gudbjartsson et al., 2015; Lang et al., 2007; Pauli-Magnus et al., 2004). Genetically, about 70% of the disease-causing variants are missense variants (Delaunay et al., 2016), in which one amino acid residue within the MDR3 protein sequence is exchanged for a different amino acid. Upon gene sequencing, the effect of identified variants within the *ABCB4* gene is hard to predict, as single nucleotide polymorphisms (SNPs) without pathogenic association can also occur. Since *in vitro* studies to analyze mutational effects are lengthy and time-consuming, I developed an ML prediction tool to accurately predict novel variants into the categories of benign or pathogenic (Publication I, Chapter 4). The project was published in Hepatology Communications (2022) and intended to assist clinicians in the initial assessment of novel variants identified in patients.

1.3.2 Impacted transitioning of a variant in FXR

The farnesoid X receptor (FXR), also called nuclear receptor subfamily 1 group H member 4 (NR1H4), is a transcription factor that controls the network of bile homeostasis by acting as a master regulator (Makishima et al., 1999; Parks et al., 1999; H. Wang et al., 1999). A homozygous variant within the FXR protein was identified in a patient suffering from PFIC subtype 5, leading to an amino acid exchange from threonine to isoleucine at position 296 (p.(Thr296Ile), identifier NM_001206979.2: c.887C>T in the FXR-encoding *NR1H4* gene) (Pfister et al., 2022). *In vitro* assays (performed by Dr. Jan Stindt, Heinrich Heine University Düsseldorf, Germany), *in vivo* patient sample analysis (performed by Dr. Carola Dröge and Prof. Dr. Verena Keitel-Anselmino) and *in silico* studies (performed by me) were performed to elucidate the underlying molecular mechanisms. Within cellular assays, the variant protein

showed significantly reduced transcriptional activity while presenting a normal protein localization and similar overall protein levels. Agonist binding occurs within the ligand binding domain (LBD) of FXR and favors a conformational state of the nearby helix 12 that creates an interaction surface for nuclear coactivators (Mi et al., 2003). Using MD simulations, I investigated the LBD of FXR and compared the wildtype to the variant protein. The variant showed a significant destabilization of the active conformation and a reduced ability to reach the active conformation from an inactive starting position. To further exclude that the variant might impact ligand binding, I analyzed the melting temperature of FXR *in vitro*, revealing no significant differences between wildtype and variant protein. The project is available as preprint (DOI 10.1101/2024.02.08.579530) (Publication II, Chapter 5).

Chapter 2 Background

2.1 Computational biology

Computational biology has developed into an indispensable research area to tackle many biological questions despite it being a rather new field in the context of biological research, with research beginning in the 1950s with the pioneering work of Margaret Oakley Dayhoff (Dayhoff, 1966; Gauthier et al., 2019). The rapidly growing data collection through advances in biological techniques (be it in genomics (Ansorge et al., 1986; F. S. Collins & Fink, 1995; L. M. Smith et al., 1986), proteomics (Reel et al., 2021), structural biology (Unwin & Henderson, 1975), or cellular imaging (Klar et al., 2000) to only name a few) demands for tools to analyze and visualize the data, as well as to extract patterns, draw conclusions and use generated data for simulations and predictions. Accordingly, computational biology nowadays is an extensive research area with numerous subfields, for example, machine learning (Chapter 2.1.1) or molecular dynamics simulations (Chapter 2.1.2). While the field is rapidly advancing and yields increasingly accurate predictions, studies benefit from collaborative efforts of researchers providing *in vitro* or *in vivo* data to corroborate computational results and vice versa.

2.1.1 Machine learning

Machine learning (ML) describes the process of identifying a model that can describe or predict data to a sufficiently accurate level (Lo Vercio et al., 2020). In itself, this process is not something unique – animals learn and interact with the world in a learning process where decisions are made based on previous information and derived causative patterns (Greener et al., 2022). A lion cub might learn about different classes of animals, learning to differentiate which ones are potential food sources (e.g., zebras) and which ones to better stay clear off (e.g., porcupines). Much of this classification process that the lion is undertaking is likely image-based, i.e., the visual cue of spikes presented by the porcupines will result in the behavioral output of being more careful and not attacking. Translated into the world of ML, the lion has learned, based on previous data, from the available information (image-based, smell-based, touch-based) to classify new objects into potential food sources or danger

sources by applying and testing different behavioral models and continuously correcting them. The lion cub will probably have tried more than once to attack a porcupine, either with his claws or jaw, and adjust his behavior based on good results – e.g., gaining food – or bad results – e.g., pain. However, one must be wary of comparing how computer programs and animals learn. Here, I am describing associative learning, meaning associating data to a particular outcome, employing a reward and/or punishment system. Animal traits such as curiosity and play, huge driving factors for efficient learning for future situations, are unknown in ML and even artificial intelligence. ML models will perform their training as often as the researcher designing them wants them to; they will not spontaneously decide to train more or search for more data out of curiosity (with the exception of reinforcement learning, see next paragraph). While this can be seen as an inherent limitation, it provides the huge advantage of reproducibility in the context of research. Using the same underlying data and the same training conditions, the model will always result in the same output – a feat that will never be reached in animal learning. One might argue that other factors contribute to this variability within animals (genetics, environmental factors, previous learning experiences). Machine-based learning, in contrast, can start from a "blank slate" state (see Essay "Tabula Rasa" (2019) by David Young, including a foreword by Jason Bailey, published on Artnome [www.artnome.com]). Try as we might, animals cannot reset themselves to a blank canvas, while resetting machines is a common procedure. Overall, this provides ML with the advantage of standardization and reproducibility (Heil et al., 2021).

ML has three major forms: supervised, unsupervised, and reinforcement learning (Morales & Escalante, 2022). In supervised machine learning approaches, the ML model is built on a dataset with a known output, a so-called label, and trained to find a function to map dataset features to the label (Lo Vercio et al., 2020). Such a label does not exist in unsupervised learning, and the machine is expected to derive meaningful patterns from the datasets (Greener et al., 2022). This is extremely valuable in situations where data is too complex for a human to process, and ML techniques can aid, for example, in reducing the dimensionality of the problem with principal component analysis (Salem & Hussein, 2019). In reinforcement learning, the ML is an agent interacting with its environment. The agent learns from feedback from the environment after performing an action (Morales & Escalante, 2022). Reinforcement learning mimics more closely the process of natural learning similar to human or animal learning, resulting in adaptability based on continuous interaction and feedback with a given

environment (Sutton & Barto, 2018). Currently, reinforcement learning harbors great potential and has even outperformed human performance at complex computer games (Mnih et al., 2015). Applications in real life healthcare scenarios remain challenging (Dulac-Arnold et al., 2021) but advances are evident especially in areas of sequential decision making (Böck et al., 2022; Coronato et al., 2020). Nonetheless, in the field of biomedicine, supervised ML techniques are most common since models are often sought to associate specific human features (weight, smoking status, age, imaging data, etc.) with a disease outcome (e.g., chronic obstructive pulmonary disease development (X. Wang et al., 2023), skin cancer detection (Esteva et al., 2017), (Jovel & Greiner, 2021)).

Underfitted and overfitted models

In a supervised ML approach, a model will learn to associate specific feature patterns for data points with an output (Lo Vercio et al., 2020). Its learning system is based on a loss function that is calculated at each learning iteration and indicates an improvement or worsening of the model (Kamatani et al., 2017; Morales & Escalante, 2022). Additionally, the designing researcher has the opportunity and responsibility to survey the performance of the ML model and thread the line between underfitting and overfitting. An overfitted model follows the underlying training data too closely and believes that the inherent noise contains valuable information. It has thus memorized the training data instead of learning its underlying trends (Jovel & Greiner, 2021). Overfitted models fail to draw appropriate conclusions for future observations, severely limiting their predictive power (Lo Vercio et al., 2020). Underfitted models, however, fail to capture the connection between training data and labels, indicating that the model does not capture the complexity of the analyzed system. Underfitted models will perform poorly on the training data and have poor predictive power (Lo Vercio et al., 2020). While poor performance at predicting the training data labels is an easy way to detect underfitting, overfitting is a more common problem in ML due to its good performance on training data (but with the major drawback of poor performance on unseen or future data). To counteract overfitting, it is common practice to withhold a part of the available data (the so-called test set) (Liu et al., 2019; Michelucci, 2018). The ML model will be trained on a large part of the available data and finally predict the output for the test set, thus imitating how the model will behave on future unseen data. Overfitted models will show reduced performance in the performance comparison between the training and final test set. Another technique to evaluate and limit overfitting is the use of resampling techniques for the training dataset

(Charilaou & Battat, 2022). Making use of the same underlying principle of withholding a part of the available data to use as an interim test, the dataset – without the final test set – can be split into training data and internal validation set. The algorithm will train on the training data, evaluate its performance on the internal validation set, shuffle the data again and start training again with a new split of training data and internal validation set (**Figure 4**). Of note, there is some ambiguity in the field regarding the naming of the different datasets (e.g., the internal validation set is sometimes referred to as dev dataset (Michelucci, 2018) or test set and the external set sometimes referred to as validation set (Cabitza et al., 2021)). Here, in line with the naming in Publication I, I will follow the introduced naming of validation set as the internal validation subset used within the training of the algorithm and test set as the final external set to evaluate the model performance on unseen data.

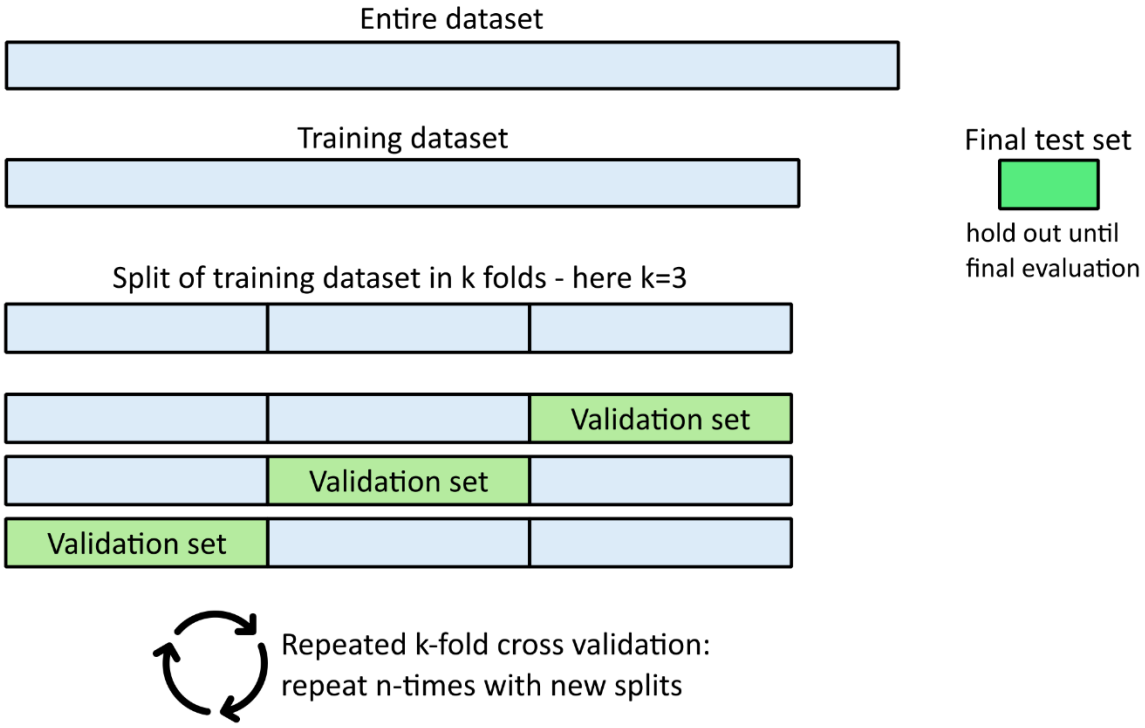


Figure 4: Schematic overview of dataset handling for ML models with repeated k-fold cross-validation. Performance evaluation scores are retained after each iteration to compare to the performance on the final test set.

A popular resampling technique is k-fold cross-validation, in which the training dataset is split into equally sized subsets (so-called folds) where k indicates the number of subsets. Each one of those subsets will now, in turn, be used as an internal validation set, meaning there will be k-times iteration rounds where one specific subset serves as a validation dataset (Refaeilzadeh et al., 2009). After each iteration, the evaluation scores are saved, but the model itself is

discarded, meaning the ML model faces entirely new data in every iteration from its point of view. The saved evaluation scores aid in evaluating the ML model and testing its performance. A variation of the k-fold cross-validation is the repeated k-fold cross-validation (**Figure 4**), where another parameter defines the number of repeats, and within each repeat, the folds are differently split, resulting in a more robust model assessment (Rodriguez et al., 2010; Rodríguez et al., 2013).

Popular algorithms for ML models – Decision Trees

The choice of algorithm to employ for an ML model depends on the specific question the ML model is trying to answer. In general, ML models create an objective function to map the input to the output variable. The objective function consists of a loss function and a regularization term (Equation 1).

$$Objective\ function = \sum l(y_i, y_{i\_pred}) + \sum \Omega(f) \qquad \text{Equation 1}$$

The first term describes the loss function measuring the difference between true output (y) and predicted output (y_{pred}) at the instance i. Ω(f) represents the regularization term that is applied to each tree (f) of the ensemble. The loss function evaluates the error the model makes during training and iteratively tries to minimize it, while the regularization term acts to control overfitting.

In supervised ML, predictions are made based on a learning period with a training dataset containing established examples with known output (Lo Vercio et al., 2020; Rokach & Maimon, 2005). This output can be a categorical (e.g., image classification into category dog or cat) or a continuous (e.g., estimation of house prices) variable. Common ML algorithms to employ for such tasks are Decision Trees (Pedregosa et al., 2011; Quinlan, 1993), Naïve Bayes (John & Langley, 1995), Support Vector Machines (Keerthi et al., 2001), Random Forest (Breiman, 2001), Linear Regression (J. Han et al., 2011), Logistic Regression (Cessie & Houwelingen, 1992) and Neural Networks (J. Han et al., 2011; Lo Vercio et al., 2020; Sarker, 2021). Decision Trees are a popular choice for classification problems, where the data is recursively split based on the most significant attributes or features (Rivera-Lopez et al., 2022; Rokach & Maimon, 2005). The general idea of a Decision Tree is commonly used in daily life decisions, even though we are mostly unaware of it, and, in contrast to an ML model, we do not have to follow the rules of the tree strictly. In a simplified way, I might ask myself the question if and what to eat

(**Figure 5**). Collecting data about previous times I have asked myself that question I can create a dataset and identify important features, for example my hunger level, finances, and the current situation I am in. Those features carry information on why I reached a specific decision (which output class I chose) and I can reconstruct a basic Decision Tree from it. The output can be a multi-class classification (e.g., eat in a restaurant, eat Apple, cook food, do not eat), a binary classification (e.g., eat / do not eat) or a regression (e.g., how much to eat).

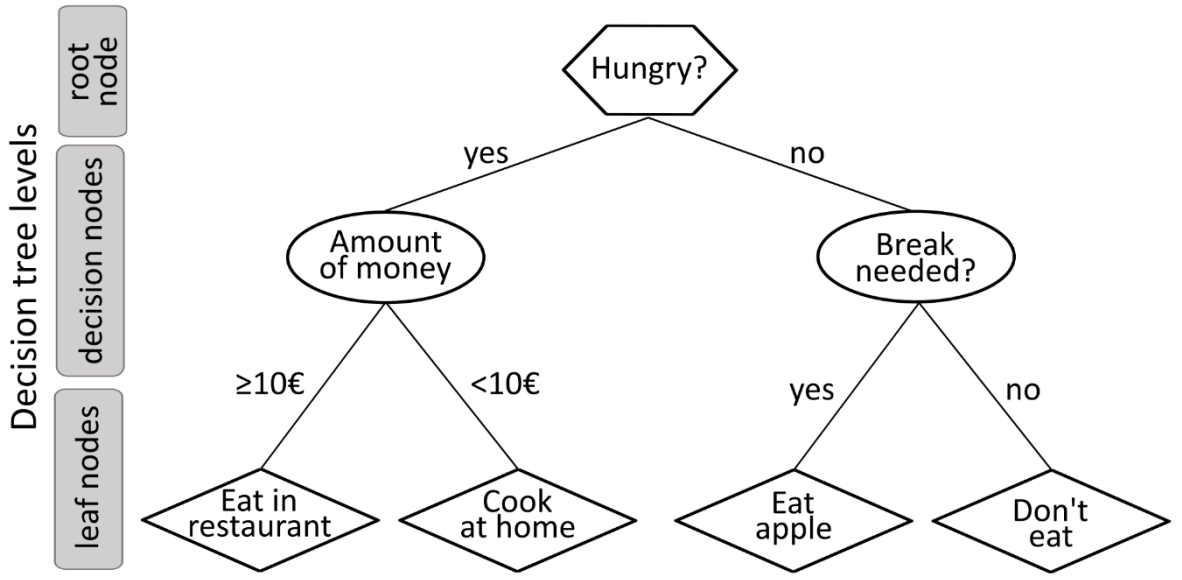


Figure 5: Example of a simple Decision Tree. The root node in this case is the feature "Hungry" which splits the tree based on the answer yes and no, resulting in two different branches with further decision nodes until the last level of the tree with the final leaf nodes. The leaf nodes each hold a class label. Note that while each leaf holds a different class label (multi-class classification) here, several leaves can also hold the same class label. Changing the exemplary Decision Tree to a binary output of "Eat" or "Do not eat", the leaves might change from "Eat in restaurant" to "Eat", "Cook at home" to "Do not eat", and "Eat apple" to "Eat".

The decision for a split is performed in a top-down fashion, where the algorithm chooses the variable that best splits the dataset at each given point (Rokach & Maimon, 2005). Depending on the used algorithm, the underlying metric for choosing the best split may vary; however, the overall goal is to increase the homogeneity of the target variable within the resulting split datasets (Rokach & Maimon, 2005). A single Decision Tree is inherently prone to overfitting, especially if the tree is allowed to branch down fully (Leboeuf et al., 2020). To avoid this, a tree can be pruned – limiting the number of times it is allowed to branch – or several trees can be combined in ensemble techniques (Rokach & Maimon, 2005; Sagi & Rokach, 2018).

Ensemble techniques increase the model performance by building a larger number of trees, minimizing the errors of individual trees (Dietterich, 2000). Bagging and boosting methods are common in building these trees (Breiman, 1996; X. Dong et al., 2020). In bagging, a number of subsets are randomly drawn from the original dataset, and upon each of these subsets, a Decision Tree is trained, whose output is either averaged or majority-voted over all trees to obtain the final ensemble classifier (Dietterich, 2000). The drawn training subsets are independent from one another, so training is performed in parallel (Bauer & Kohavi, 1999). In boosting, the model is built on combining weak classifiers in a chain, where each new classifier attempts to minimize the error of the previous classifier (Freund, 1995; Schapire, 1990). Assembling those weak classifiers results in a more robust prediction. In gradient boosting, the loss function is minimized based on gradient descent (Friedman, 2001). It follows the same principle as general boosting approaches in that trees are iteratively added, and each new tree trains on residual errors of previous trees, thus concentrating and improving on the weak areas of the model performance. One of the most popular ensemble algorithms, often achieving the best results in machine learning competitions on the Kaggle platform, is the XGBoost algorithm (Chen & Guestrin, 2016). XGBoost stands for extreme gradient boosting, which uses a gradient boosting framework optimized for speed and performance (Chen & Guestrin, 2016). While neural networks usually outperform other algorithms in problems involving unstructured data, such as image analysis (Sharada et al., 2023), XGBoost is a popular choice due to its superior performance on tabular data sets, especially small to medium dataset sizes (Grinsztajn et al., 2022; Shwartz-ziv & Armon, 2021).

ML predictions of missense amino acid substitutions

The human population displays substantial genetic variability where the most common genetic difference, a single nucleotide polymorphism (SNP), occurs about every thousand base pairs when comparing two individuals (Auton et al., 2015). While many of those nucleotide exchanges may not result in a difference on the protein sequence level, a considerable subset does result in a single-site amino acid exchange (also frequently referred to as variant or mutation) (Auton et al., 2015; Thusberg & Vihinen, 2009). Mutations can have pronounced effects, and even single missense mutations have been identified as disease-causing for a variety of disorders (Botstein & Risch, 2003) such as Alzheimer's disease (Goate, 2006), amyotrophic lateral sclerosis (Rosen et al., 1993) and PFIC (Dröge et al., 2017). Mutations might impact functionally important sites, change protein dynamics or cellular localization,

affect the protein's structural properties, disturb inter- or intramolecular residue networks, prevent or change post-translational modifications (PTMs), or impact protein translation on the mRNA level by altered mRNA stability or splicing (Thusberg & Vihinen, 2009; Z. Zhang et al., 2012). On the other hand, many mutations might not alter the protein function at all or even give it an evolutionary benefit (Tóth-Petróczy & Tawfik, 2014). Evaluating mutations *in vitro* is time- and cost-intensive, and accordingly, predicting the effect of a mutation on the protein function is a field of intense research, and a range of predictors are available (Mooney et al., 2010). Broadly, predictors can be categorized into sequence- or structure-based or considering information from both areas. Sequence-based predictors estimate evolutionary conservation based on multiple sequence alignments, following the reasoning that benign substitutions are less evolutionary penalized (Miller & Kumar, 2001). Structure-based predictors take protein structural effects of mutations into account, either based on available protein structures or local or global structure predictions (Ittisoponpisan et al., 2019). Combined approaches with sequence conservation as well as structural impact considerations have been found to further improve predictions (Folkman et al., 2013) and are employed in widely used tools such as PolyPhen-2 (Adzhubei et al., 2010). Most prediction tools are designed to predict substitutions for any given protein, which gives the developer the advantage of a bigger available dataset for developing the tool and a larger potential user group. However, it does not guarantee good performance of the predictor on every protein as one potential pitfall can be a skewed training dataset towards a certain protein class, resulting in weaker performance on other protein classes. Several studies have benchmarked predictors using different proteins with established missense substitutions, resulting in vastly varying performances (Choudhury et al., 2022; Livesey & Marsh, 2023; Riera et al., 2016). Accordingly, protein-specific predictors have also been established to increase performance for specific proteins of interest (Crockett et al., 2012; Niroula & Vihinen, 2015; Riera et al., 2016). For the case of MDR3, no protein-specific predictor was available despite its importance in liver health and bile homeostasis. A previous study claimed MutPred to be a well-performing general protein predictor on MDR3 variants (Khabou et al., 2017); however, its performance was tested only on a small set of variants and thus might not be representative. As missense variants in MDR3 have been associated with a range of liver diseases, accurate predictions for this specific protein are of high interest. Hence, I established a dataset containing MDR3-missense variants with pathogenic and benign effects and trained

an XGBoost ML model specifically for MDR3 (Publication I, Chapter 4). The general structure of the code and the integration of some features was established by Pegah Golchin (Heinrich Heine University Düsseldorf, Germany; currently at TU Darmstadt, Germany). We used a combination of input features from established general protein predictors, both sequence- and structure-based, and features to explicitly include secondary structure effects such as PTM changes and solvent accessibility for the ML model. The approach led to improved prediction results, outperforming general protein prediction tools.

2.1.2 Molecular dynamics simulations

The computational method of molecular dynamics (MD) simulations, pioneered by work from Alder and Wainwright in the 1950s (Alder & Wainwright, 1959), allows studying the motions of a biomolecular system based on solving Newton's laws of motion. Biomolecular structures like proteins can be embedded in a specific environment to mimic the cellular context, and the movement of each atom during specified time steps is calculated based on physical properties and interatomic interactions. Depending on the investigated research question, a quantum mechanical description of the system with an explicit representation of electrons can be necessary. However, due to their complexity, such calculations are computationally expensive and currently out of range for larger systems such as proteins (Bottaro & Lindorff-Larsen, 2018). In turn, molecular mechanics is mainly used to describe protein systems, where each atom is described as a point connected with springs to represent the bonds to other atoms (Braun et al., 2019). Hybrid models of quantum mechanics and molecular mechanics MD can be used when explicit electronic description of a part of the system is required (Horn, 2003). For even larger systems and/or if the problem allows a reduced degree of complexity of the system, atoms can be grouped, e.g., atoms of an amino acid residue will be represented by a pseudo-atom (so-called coarse-grained MD) (Levitt & Warshel, 1975). Within the here presented thesis, molecular mechanics MD (in the following referred to as MD) was used to investigate conformational changes and missense variant impact within proteins. The protein of interest is described on an all-atom level, and the forces acting on each atom of the system are calculated based on the bonded (bond, angle, and torsion terms) and non-bonded (electrostatic and van der Waals terms) interactions. After a given amount of time steps, the newly calculated atomic positions and velocities are stored in a so-called "snapshot" or

"frame" that, taken together over the entire simulated time, form the trajectory of the system representing the 3D dynamical movement of the analyzed system (Hollingsworth & Dror, 2018). Accurately describing the atomic interactions is of great importance to computing interatomic forces. Accordingly, considerable research has been and continues to be carried out on generating functions (referred to as "force fields") that describe the atomic behavior well, matching simulated properties to experiments from physics and chemistry (Koes & Vries, 2017; Love et al., 2023). The general force field form to compute the potential energy of a system (Cornell et al., 1995) consists of bonded and non-bonded terms (Equation 2):

$$E_{total} = \sum_{bonds} K_r(r - r_{eq})^2 + \sum_{angles} K_\theta(\theta - \theta_{eq})^2 + \sum_{dihedrals} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] + \sum_{i<j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\epsilon R_{ij}}\right]$$

Equation 2

The first three terms describe the bonded interactions with bond stretching and bond compression, bond angle deviations, and torsion angle deviations expressed in the terms, respectively. Non-bonded interactions are described in the last term of the equation, combining a Lennard-Jones (12,6) potential and a Coulomb potential to describe van der Waals and electrostatic interactions (Cornell et al., 1995).

Within the open-source biomolecular MD program AMBER (Assisted Model Building with Energy Refinement (Case et al., 2021, 2023)), several sets of force fields are integrated and can be used for simulations of varying molecules, including proteins (e.g., force field ff19SB (Tian et al., 2020)), DNA (e.g., force field OL21 (Zgarbová et al., 2021)), carbohydrates (e.g., force field GLYCAM_06j (Kirschner et al., 2008)) and lipids (e.g., force field lipid21 (Dickson et al., 2022)). Additionally, AMBER provides, amongst others, a set of programs for the preparation and execution of molecular simulations and as such has become widely popular within biomolecular research (Case et al., 2021; Salomon-Ferrer et al., 2013). Furthermore, a wide range of compatible force fields exist, designed for specific cases such as phosphorylated amino acids (Stoppelman et al., 2021), fluorescent dye-linked proteins (Schepers & Gohlke,

2020), or to simulate gold-nanoparticles linked with bioactive molecules (Pohjolainen et al., 2016), bridging thus several research fields and broadening the potential applications. Within unbiased MDs, the system will explore its energetically accessible free energy landscape over the simulation time, creating a dynamic sampling of energetic states (Karplus & McCammon, 2002; Orellana, 2019). In a perfect ergodic trajectory, the system will visit all available states, resulting in a full view of conformational space in the case of simulated proteins (Abrams & Bussi, 2013; Pietrucci, 2017). Frequently, however, such an ergodic state is not reached as it requires long simulation times, resulting in the accessing of available energetic minima more frequently, while higher energetic states will be visited less frequently (Abrams & Bussi, 2013). Accordingly, a range of computational methods have been derived to accelerate and enhance the sampling of conformations blocked by high free energy barriers and higher energy states like transitioning states (Abrams & Bussi, 2013; Y. I. Yang et al., 2019). Despite the potential inaccessibility of certain states, biologically relevant conformational transitioning can be observed within unbiased MD simulations and as such, unbiased MDs have been used frequently to derive answers from a molecular view on protein flexibility, substrate transport, and ligand interactions (Calimet et al., 2013; Halder et al., 2015; Latorraca et al., 2017; Orellana, 2019; Skjaerven et al., 2011).

When studying protein-ligand interactions or within drug design approaches, molecules may be present that cannot be accurately described with the integrated force fields, leading to the development of the general AMBER force field (GAFF), which supplies parameters for all bonded terms and the van der Waals term for most organic molecules (J. Wang et al., 2004). To determine the missing parameters for the electrostatics term (i.e., the atom-centered point charges), charges are fitted to reproduce the molecule's quantum mechanically calculated electrostatic potential (ESP). For contemporary AMBER force fields, a restrained electrostatic potential (RESP) fit (Bayly et al., 1993) has been shown to be superior to an unrestrained fit to the ESP (Cornell et al., 1993) and is therefore considered the standard method to determine atomic partial charges. The basis set used to describe and compute the molecular orbitals determines the quality of the calculated molecular electrostatic potential, and popular basis sets differ in the number of Gaussian functions used to describe the atomic orbitals. Though the general tradeoff between computation cost and level of detail applies – i.e., the inclusion of more Gaussian functions will more accurately depict the true ESP at the cost of increasing computation time – a level of 6-31G(d) is frequently used as the derived ESP charges start to

converge at this level (Dupradeau et al., 2010; Hariharan & Pople, 1972). With the accurate description of molecules of interest for MD approaches, a wide range of protein-ligand interaction studies can be assessed.

Nuclear receptors (NRs) are key regulators of a diverse set of physiological functions. Besides the overall domain structure of a DNA-binding domain (DBD) and an LBD with a hydrophobic binding cavity, NRs show a common active conformational state upon agonist binding with a well-defined helix 12 placement (Aranda & Pascual, 2001; Khan et al., 2022; Saen-Oon et al., 2019; Wurtz et al., 1996) (see Chapter 2.3). Due to the importance of the protein class and their effect range, NRs have been a target for intense research efforts to design molecules for specific regulation (Jiang et al., 2021; Jin et al., 2013; Merk et al., 2019; Saen-Oon et al., 2019). A wide spectrum of NR crystal structures has led to a well-defined but static picture of the active state (Aranda & Pascual, 2001). However, the dynamical conformational change from an inactive to an active state has been difficult to investigate in detail, while it may hold the key for the ligand and effect variability as well as structural flexibility of NRs (D'Arrigo et al., 2022; Folkertsma et al., 2005; Jiang et al., 2021). Accordingly, I employed unbiased MD simulations to investigate the transition from an inactive conformation to an active state, including the impact of a clinically identified variant to uncover its molecular mechanistic effect (Publication II, Chapter 5).

2.2 MDR3 acts as an important transporter in bile homeostasis

The multidrug resistance protein 3 (MDR3) acts as a phosphatidylcholine (PC) floppase at the canalicular membrane, enabling extraction of PC into mixed micelles and thus maintaining healthy levels of phospholipid to bile salts ratios which in turn aid in the solubilization of hydrophobic cholesterol to prevent gallstone formation (Carey & Small, 1978; Elferink et al., 1997; Lammert et al., 2004; Oude Elferink & Paulusma, 2007). Furthermore, the formation of mixed micelles lowers bile salt toxicity towards membranes, thus protecting the biliary tract from detergent effects (Ikeda et al., 2017; Oude Elferink & Paulusma, 2007).

2.2.1 MDR3 transporter structure and function

MDR3 structurally consists of two cytosolic nucleotide binding domains (NBDs), where ATP is bound and hydrolyzed, and two transmembrane domains (TMDs), spanning the membrane

leaflets (Prescher et al., 2019; van der Bliek et al., 1988) (**Figure 6**, A). Despite its high sequence identity of 76% with the P-glycoprotein P-gp (ABCB1 or MDR1), MDR3 does not show the wide substrate range of xenobiotic transport that characterizes P-gp as a major player in multidrug resistance (Finch & Pillans, 2014; L. Mercer & Coop, 2011; Prescher et al., 2021). Additionally, the transport rates for overlapping P-gp and MDR3 drug substrates are much lower in MDR3 (A. J. Smith et al., 2000). Overall, the physiological role of MDR3 is the flopping of phospholipids, especially PC, from the inner to the outer membrane leaflet (Oude Elferink & Paulusma, 2007; Prescher et al., 2019; A. J. Smith et al., 1994; van Helvoort et al., 1996). This translocation might involve a central cavity of MDR3 (Nosol et al., 2021; Olsen et al., 2020) or function via a credit card swipe mechanism along the membrane-facing transmembrane helix (TMH) 1 (Prescher et al., 2021). Canalicular membranes show a membrane asymmetry, with PC being more present in the outer leaflet than in the inner leaflet of the membrane (Eckhardt et al., 1999). Accordingly, PC transport by MDR3 is coupled to ATP hydrolysis, working against a concentration gradient. It is still unclear whether flopped PC is directly exposed by MDR3 for extraction into bile mixed micelles or whether mixed micelles later extract PC lipids from the outer membrane leaflet (Oude Elferink & Paulusma, 2007). The importance of PC transport for proper bile formation, however, is undisputable, and misfunction of MDR3 is associated with a variety of liver diseases (Boyer, 2013).

Highly conserved ABC-specific motif sequences within the NBDs of MDR3 are critical for protein function due to their involvement in ATP binding and hydrolysis (Prescher et al., 2019) (**Figure 6**, B). The Walker A motif with a consensus sequence of GXXGXGKT/S (where X can be any amino acid) is responsible for interacting with the phosphate group of ATP (Schmitt & Tampé, 2002; Walker et al., 1982). Crucial for ATPase activity is the Walker B motif, consisting of a stretch of four hydrophobic residues followed by an aspartate (D), which stabilizes a magnesium (Mg) ion (Rai et al., 2006; Urbatsch et al., 2000). Preceding the Walker B motif is the signature motif C-loop (consensus sequence LSGGQ), uniquely preserved in ABC transporters, and C-terminal of the Walker B motif resides the D-loop (consensus sequence SALD) (Prescher et al., 2019; Schmitt & Tampé, 2002). A conserved histidine is of special importance as it serves as a key networker between ATP, water molecules, the Mg ion, and other amino acids (Zaitseva et al., 2005). The coordination of ATP occurs between the two NBD subunits in a concerted action with Walker A, Walker B, and the conserved histidine of one NBD and the C-loop of the other NBD (Schmitt & Tampé, 2002) (**Figure 6**, B).

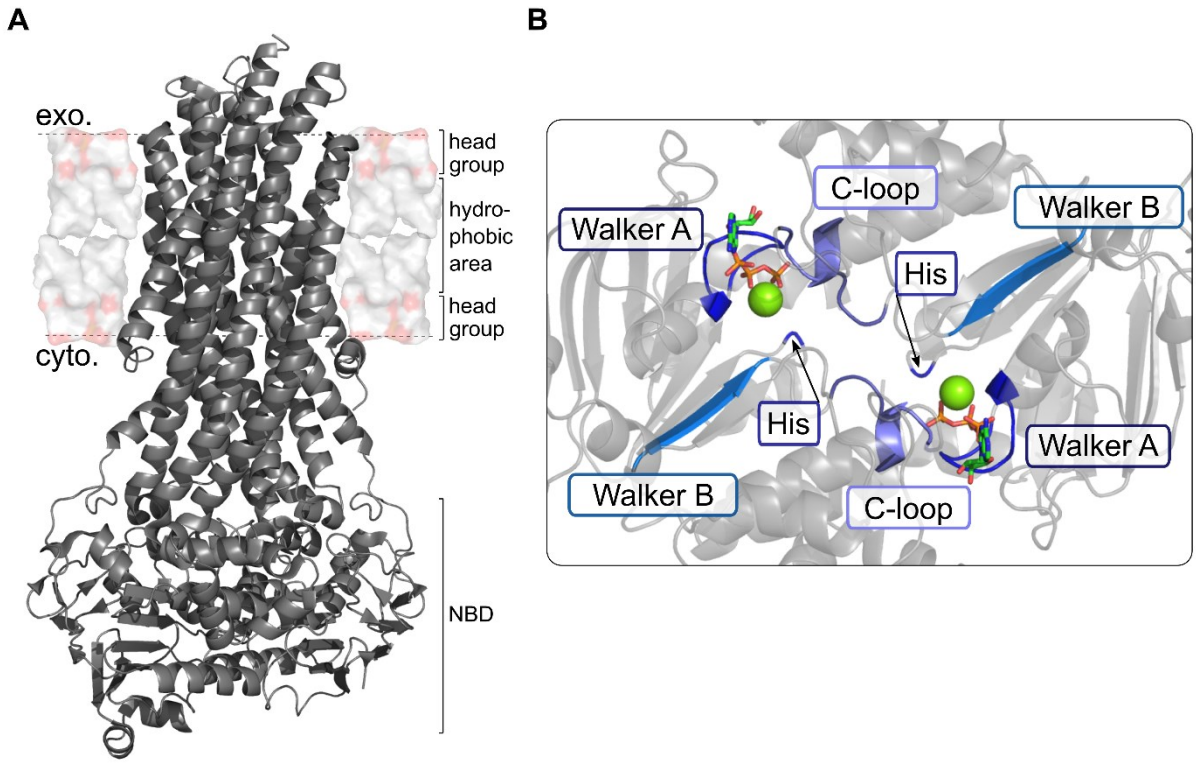


Figure 6: Protein structure of MDR3. [A] Overall MDR3 structure in the ATP-bound outward-facing conformation within the canalicular membrane (PDB ID: 6S7P, Olsen et al., 2020). [B] Rotated view on the NBD as seen from the perspective of the membrane center towards the cytosolically located NBD. Highly conserved and functionally relevant motifs are colored and marked, indicating ATP (depicted as colored licorice) and magnesium ions (depicted as green spheres) coordination between the two NBDs. exo.: extracellular, cyto.: cytosolic, NBD: nucleotide binding domain.

Based on this detailed mechanistic knowledge of key residues and motifs, mutations within these residues will likely impact protein function. Corroborating this theory, several missense mutations have been identified within these key motifs in PFIC patients, and *in vitro* analyses further confirmed that variants were normally processed and targeted to the plasma membrane but exhibited decreased activity (Degiorgio et al., 2013; Delaunay et al., 2017; Dzagania et al., 2012). However, the effect of other identified variants is less easily classified and explained at a molecular level. From the identification of the missense variant on the genetic level to the detailed mechanistic study on the protein level, analyzing a novel variant is both time- and cost-intensive. Accordingly, variant protein predictors are well-established and widely used to aid in the prioritization and analysis of variants (Choudhury et al., 2022). However, there is no established predictor with proven good performance for MDR3 despite its relevance within the liver. Considerable accumulative research has analyzed missense

variants of MDR3 both *in vivo* and *in vitro* (Andress et al., 2014, 2017; Colombo et al., 2011; Davit-Spraul et al., 2010; Degiorgio et al., 2007, 2014; Delaunay et al., 2016, 2017; Dixon et al., 2000; C. Dong et al., 2021; Dröge et al., 2017; L. J. Fang et al., 2012; Floreani et al., 2006, 2008; Frider et al., 2015; Gautherot et al., 2014; Gordo-Gilart et al., 2015, 2016; Gotthardt et al., 2008; C. Hopf et al., 2011; Jacquemin et al., 2001; Keitel et al., 2006, 2016; Khabou et al., 2017; Kluth et al., 2015; Kubitz et al., 2011; Lucena et al., 2003; Olsen et al., 2020; Park et al., 2016; Pauli-Magnus et al., 2004; Poupon et al., 2010, 2013; Rosmorduc et al., 2003; Saleem et al., 2020; Tougeron et al., 2012; Wendum et al., 2012; Ziol et al., 2008). This valuable research provided me with the necessary basis to create a dataset for ML approaches with the aim to further aid researchers and clinicians in the analysis of variants. Following the standardized American College of Medical Genetics and Association for Molecular Pathology (ACMG-AMP) guidelines, the classification by an *in silico* predictor on its own should not be taken as a definitive classification of a variant (Richards et al., 2015). However, it is a valuable help to narrow down and prioritize variants to study *in vitro* (Thusberg & Vihinen, 2009). Since MDR3 plays a vital role in bile homeostasis, its dysfunction is implicated in several diseases (Chapter 2.2.2).

2.2.2 Involvement of MDR3 in disease

MDR3 dysfunction has been linked to ICP, LPAC, DILI, PFIC3, liver fibrosis, liver cirrhosis and hepatobiliary malignancy (Deleuze et al., 1996; C. Dong et al., 2021; Dröge et al., 2017; Gudbjartsson et al., 2015; Lang et al., 2007; Pauli-Magnus et al., 2004; Rosmorduc et al., 2001). Dysfunction leads to decreased PC levels in bile micelles, changing the balance of detergent bile salts to lipids ratio, which can result in free bile salts that are able to attack epithelial tissue (Elferink et al., 1997). In the absence of MDR3, hepatocytes might have to rely fully on their asymmetric membrane composition with high levels of sphingomyelin and cholesterol in the outer canalicular leaflet for protection against detergent effects (Amigo et al., 1999; Oude Elferink & Paulusma, 2007). Furthermore, cholesterol that is not solubilized in the mixed micelles can precipitate and form gallstones (Lammert et al., 2004; Oude Elferink & Paulusma, 2007). In the majority of cases, disease-causing gene variations in the *ABCB4* gene lead to amino acid substitutions, with only a minority leading to protein truncations or other gene alterations (Delaunay et al., 2016). Delaunay et al. further suggested the classification of

variants based on their functional impact in the protein's life cycle (Delaunay et al., 2016), similar to classification schemes for other proteins like phosphatase and tensin homolog (PTEN) (Hasle et al., 2019). As a transmembrane protein, MDR3 translation occurs at the endoplasmatic reticulum (ER), positioning the protein directly into the ER membrane. From here, integral membrane proteins are trafficked to their destined localization. For MDR3, this implies trafficking via the Golgi apparatus to the apical canalicular membrane (Kipp & Arias, 2000). Missense variants can either affect protein maturation, localization, stability, activity, or a combinatorial effect (**Figure 7**). Of note, the chosen categories are protein-dependent, as transmembrane proteins differ from cytosolic proteins in their important steps within their lifecycle. Additionally, categories can be even more fine-tuned, as a minority of genetic variants might impact pretranslational steps such as mRNA stability (Stenson et al., 2003; Thusberg & Vihinen, 2009) and thus preclude the protein maturation step. For the case of MDR3, a classification scheme with protein maturation, localization, activity or stability affected, similar to the proposition of Delaunay et al., seems a sensible choice both in regards to known variant effects and potential drug intervention (Delaunay et al., 2016).

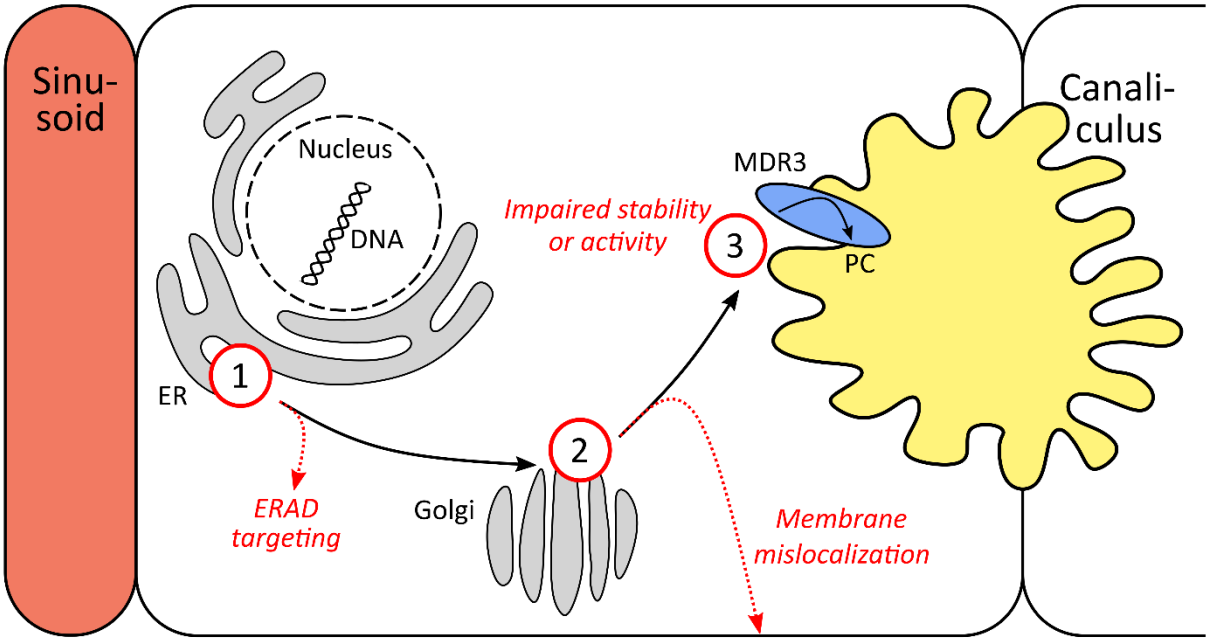


Figure 7: MDR3 protein lifecycle from translation to membrane localization. Genetic variants can impact the lifecycle at several stages, marked with red circled numbers. Missense variants might impact protein maturation, inducing misfolding that will target the protein for degradation via ER-associated degradation (ERAD) to the proteasome (1). Variants might lead to a mislocalization of the protein, preventing the protein from reaching its apical target location (2). Lastly, variants might impair the protein's activity and local stability or influence its turnover time once it is located in the apical membrane (3).

In general, the impact of a missense variant will fall somewhere on the spectrum from no effect on protein function to complete protein failure. In accordance with this, variants associated with ICP, a transient and reversible disease, while being pathogenic variants, usually have a less pronounced effect on protein function (Keitel et al., 2006; Pauli-Magnus et al., 2004) than some PFIC3 associated variants that lead to almost complete loss of function (Delaunay et al., 2016). Another factor in effect strength comes from the genetic status, whether the missense variant is present on one (heterozygous) or on both alleles (homozygous). Homozygous MDR3 variants are often associated with PFIC3 and thus tend to result in more severe phenotypes (Jacquemin et al., 2001; Saleem et al., 2020). Additional complexity within patients arises through the general genetic landscape as well as environmental factors. Compound heterozygosity, where either both alleles or one allele is marked by two or more variants, and their impact might add up to the presentation of the specific phenotype, is known in PFIC cases (Dröge et al., 2017). Furthermore, several risk genes might be impacted and contribute to disease strength, as has been shown for ICP (Keitel et al., 2006) and PFIC3 (Dröge et al., 2017).

In order to analyze this heterogeneous and complex system, I specifically narrowed down the effects of variants into the categories of benign and pathogenic as target categories for an ML approach. While further categories, as well as functional evaluation predictions, were envisioned and would certainly be beneficial, a larger and well-controlled dataset would have been required to enable such predictions. In the case of MDR3, where data from associated liver diseases was pooled, the obtained dataset size only allowed for a binary pathogenicity prediction with high confidence (see Publication I, Chapter 4).

2.3 Nuclear receptor FXR regulates bile homeostasis network

2.3.1 FXR isoform expression within the body

Two FXR genes exist, FXRα and FXRβ (Lee et al., 2006); however, FXRβ is a pseudogene in humans (Otte et al., 2003). As such, within this thesis, I am using FXR as a synonym for FXRα. The FXRα gene encodes for four isoforms formed through alternative splicing (FXRα1, FXRα2, FXRα3, and FXRα4) in human and murine models (Huber et al., 2002; Yanqiao Zhang et al., 2003). Recently, four novel but functionally defective isoforms have been identified in human

hepatocytes (Mustonen et al., 2021), and further research is needed to analyze their physiological and pathological relevance, if any. Considering only the functionally active four isoforms, they differ within their N-terminal activation function domain 1 (AF1) as well as in the presence or absence of a four amino acid long sequence, MYTG, located adjacent to the DBD (Yanqiao Zhang et al., 2003). The short MYTG sequence motif plays a role in FXR target gene activation via differential DNA-binding preferences, conferring the isoforms with sets of different transcriptionally regulated genes (Correia et al., 2015). Overall, all isoforms consist of an AF1 region, followed by the DBD and a flexible hinge region, connecting the N-terminal part of the protein to the ligand binding domain (LBD) with the C-terminal helix 12 (H12), frequently referred to as activation function domain 2 (AF2) (Yanqiao Zhang et al., 2003) (**Figure 8**).

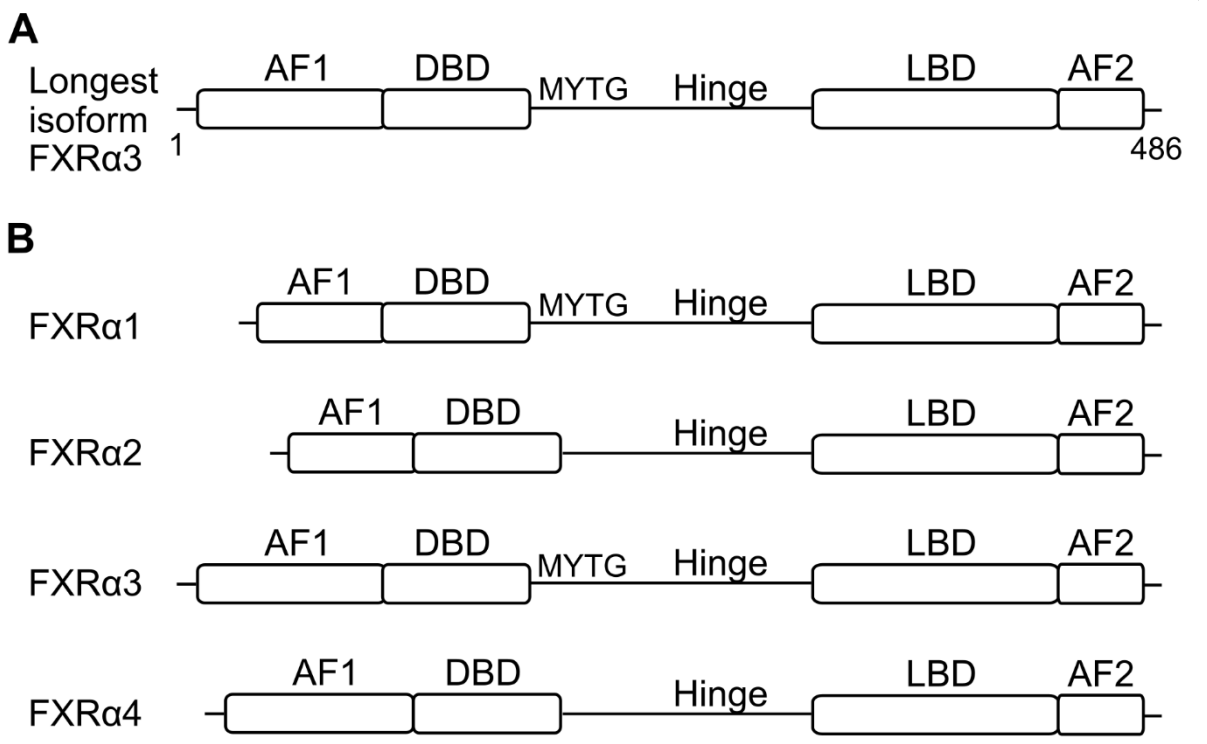


Figure 8: Schematic view on FXR isoforms. [A] The overall organization of FXR consists of an N-terminal activation function 1 (AF1) motif, followed by the DNA-binding domain (DBD) and a hinge region connecting to the C-terminal ligand binding domain (LBD) with the activation function 2 (AF2) motif, exemplarily depicted in the longest isoform FXRα3 with 486 amino acids. [B] The four FXRα differ in their AF1 sequence and in the presence or absence of the short sequence motif MYTG. Figure loosely based on Yanqiao Zhang et al., 2003.

FXR is highly expressed within the liver and small intestines (Vaquero et al., 2013), with lower levels in the kidney and the adrenal gland (Forman et al., 1995). Lower mRNA levels of FXR

have been identified in a variety of tissues and cell types, including in glial and neuronal cells, vascular smooth muscle cells, pancreatic β cells and immune cells (Albrecht et al., 2017; Bishop-Bailey et al., 2004; C. Huang et al., 2016; Renga et al., 2010; Schote et al., 2007). Overall, a combination of both MYTG-positive (α1 or α3) and -negative (α2 or α4) FXR isoforms can be found in FXR-expressing cells (Ramos Pittol et al., 2020), revealing a specific balance of FXR isoform expression. FXRα1 and FXRα2 are predominantly expressed within the liver (Huber et al., 2002; Vaquero et al., 2013), with the metabolism in human and mouse liver cells being mainly driven by the FXRα2 isoform (Ramos Pittol et al., 2020; Vaquero et al., 2013). Within the intestines, FXRα3 and FXRα4 are the predominant isoforms (Huber et al., 2002). Different isoforms showed preferential DNA-binding motifs (Ramos Pittol et al., 2020) and thus differential isoform expression influences FXR downstream targets. Intriguingly, the ongoing investigation of cell type specific FXR effects in a range of cell types highlights the possibility of an even more complex system than currently anticipated. Since the introduction of the concept of the gut-liver axis (Marshall, 1998), the inter-organ connectivity and its interplay has revealed widespread implications in human health and disease states with the bile acid-receptive FXR as a prominent regulator (Blesl & Stadlbauer, 2021; Perino et al., 2021; Tilg et al., 2022) (**Figure 9**). Bile homeostasis regulation (Radun & Trauner, 2021), glucose and lipid metabolism (Ma et al., 2006; Y.-D. Wang, Chen, Moore, et al., 2008), anti-inflammatory effects (Y.-D. Wang, Chen, Wang, et al., 2008) and liver regeneration (W. Huang et al., 2006) are amongst the most prominent FXR functions. Interestingly, with the emergence of the gut-liver-brain axis, FXR modulation might have even further implications (M. Yan et al., 2023). Effects may work indirectly through interactions to the microbiome and bile acid levels (Perino et al., 2021) or through yet-to-be clearly established functions of expressed FXR in neurons and oligodendrocytes within the brain (Albrecht et al., 2017; Deckmyn et al., 2022; C. Huang et al., 2016).

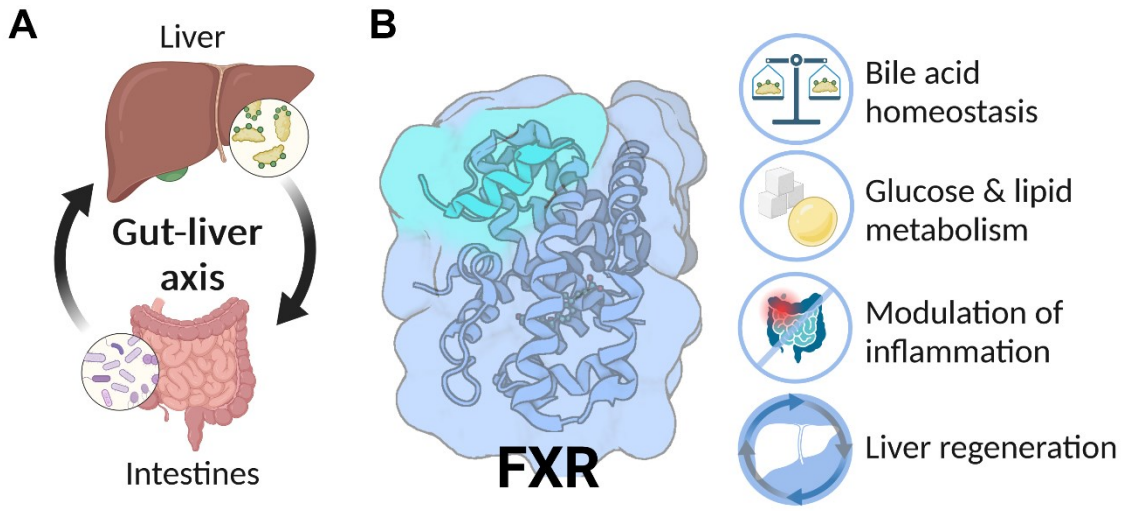


Figure 9: Prominent functions of FXR within the gut-liver axis. [A] An intricate interplay between bile acid pool and microbiome exists within the gut-liver axis. [B] FXR (shown here: LBD structure based on agonist-bound crystal structure (Merk et al., 2019) with a coactivation peptide shown in green) is an important regulator within this inter-organ connectivity. Established FXR functions, maintained by diverse transcriptional regulation, include bile acid homeostasis, glucose and lipid metabolism, modulation of inflammation and liver regeneration. Created with BioRender.

2.3.2 Transcriptional regulation by FXR

A variety of genes related to bile acid, lipoprotein, and glucose metabolism are regulated by FXR (Ma et al., 2006; Sinal et al., 2000). Two zinc finger motifs within the DBD of FXR, consisting of four cysteine residues, each coordinating one zinc ion, form the basis for DNA recognition, a mechanism conserved in the nuclear receptor superfamily (Rastinejad et al., 2000). Canonically, FXR forms a heterodimer with the retinoid X receptor α (RXRα, in the following shortened to RXR) (Forman et al., 1995); however, it can also act as a monomer or homodimer for specific genes such as apolipoprotein A-1 and the glucose transporter GLUT4 (Claudel et al., 2002; Shen et al., 2008). The genomic target sequences are so-called FXR response elements (FXREs) within the promotor region of downstream target genes, with an inverted repeat sequence of AGGTCA bases separated by a 1-base pair spacer (IR-1) being the highest affinity binding site for the FXR/RXR dimer (Laffitte et al., 2000). Additionally, the FXR/RXR dimer can bind to other DNA sequences, such as direct repeat sequences with one to five nucleotide spacers (DR-1 to DR-5) (Laffitte et al., 2000). Furthermore, FXRα2 and FXRα4 can bind to everted repeat sequences with a nucleotide spacer of two base pairs (ER-2), and binding was shown to be an important regulator in mouse and human liver cells besides the canonical IR-1 binding site (Ramos Pittol et al., 2020). Regulation by FXR is target specific and

thus can either induce (e.g., GLUT4 (Shen et al., 2008)) or repress (e.g., APOA1 (Claudel et al., 2002)) the target gene.

Adding to the complexity of the FXR-regulated network are tissue-specific effects. Within the intestines, the apical sodium-dependent bile acid transporter (ASBT, also called ileal bile acid transporter IBAT) is responsible for ileal reabsorption of bile acids, and its expression is downregulated upon FXR activation (Neimark et al., 2004). Upon absorption of bile acids into the enterocytes, the small cytosolic protein ileal bile acid-binding protein (IBABP) facilitates intracellular trafficking of bile acids (Alrefai & Gill, 2007; Trauner & Boyer, 2003). IBABP expression is increased on FXR activation (Coppola et al., 1998; Grober et al., 1999; Nakahara et al., 2005), ensuring functional sensing of the bile acid levels. Furthermore, activated FXR induces the expression of human fibroblast growth factor 19 (FGF19) (Song et al., 2009) or the corresponding ortholog gene FGF15 in mouse models (Inagaki et al., 2005). Within mouse models, intestinal FXR activation leads to FGF15 expression, export and subsequent suppression of liver-specific cholesterol 7α-hydroxylase (CYP7A1) via FGF receptor 4 (FGFR4) binding and a c-Jun N-terminal kinase (JNK)-dependent pathway (Holt et al., 2003; Inagaki et al., 2005; Kim et al., 2007; Xie et al., 1999). CYP7A1 is the rate-limiting enzyme within the bile acid synthesis (Russell, 2003), and accordingly, the regulation of this critical enzyme impacts overall bile homeostasis. Interestingly, liver FXR stimulation did not repress CYP7A1 within the liver in a knockout mice model with tissue-specific intestinal FXR deficiency (Kim et al., 2007), verifying gut-liver signaling and transportation of intestinally secreted FGF15 to the liver. Within humans, there are contradictory indications about whether the same clear tissue specificity takes place. While FGF19 mRNA was not detectable in human liver samples (Nishimura et al., 1999), mRNA and protein FGF19 could be detected at low levels in primary human hepatocytes and positively responded to FXR agonist treatment (Song et al., 2009). Thus, it seems likely that within humans, both intestinal and liver FXR activation leads to FGF19 upregulation and secretion and consequently to suppression of CYP7A1 to decrease bile acid synthesis.

Within the liver, the bile salt export pump (BSEP) is a prominent and well-studied example of FXR regulation, where expression of BSEP is driven by binding of the heterodimer FXR/RXR to an IR-1 motif in the BSEP promotor (Ananthanarayanan et al., 2001; Gerloff et al., 2002; Ijssennagger et al., 2016; Plass et al., 2002). BSEP, as the main bile salt efflux transporter located at the canalicular membrane of hepatocytes, is a critical factor for proper bile flow

(Strautnieks et al., 1998). Further, active FXR transactivates the orphan nuclear receptor small heterodimer partner (SHP), which can negatively regulate other nuclear receptors such as the liver receptor homolog-1 (LRH-1) (Goodwin et al., 2000; Lu et al., 2000). In turn, LRH-1 is essential for CYP7A1 gene expression (Nitta et al., 1999). As such, downregulation of CYP7A1 via bile acid-mediated FXR activation occurs via FGF19 signaling and via the SHP and LRH-1 axis, thus limiting novel bile acid synthesis. On the other hand, SHP represses the expression of the hepatic basolateral located bile salt importer, also called sodium taurocholate cotransporting polypeptide (NTCP) (Denson et al., 2001), consequently limiting the uptake of bile acids into the hepatocyte and preventing toxic effects. To further the same end, efflux of bile acids into the bloodstream is upregulated upon FXR activation through increased protein expression of the heterodimer transporter organic solute transporter α (OSTα) and OSTβ as evidenced in human hepatoma cell lines (Landrier et al., 2006) and in sandwich-cultured human hepatocytes (Guo et al., 2018; Y. Zhang et al., 2017). Additionally, FXR transactivates the expression of MDR3, thus upregulating the secretion of phospholipids into bile (L. Huang et al., 2003; Ijssennagger et al., 2016). PTMs have been shown to further influence and regulate FXR (reviewed in Appelman, van der Veen, and van Mil 2021). Besides bile homeostasis regulation, FXR is linked to the regulation of hepatic inflammation and inflammation-driven development of hepatocellular carcinoma (HCC). FXR represses the nuclear factor-κB (NF-κB) signaling pathway in human hepatoblastoma and in primary mouse cells (Y.-D. Wang, Chen, Wang, et al., 2008), explaining the increased inflammation found in FXR knockout mice (Kim et al., 2007; F. Yang et al., 2007). However, in a reciprocal fashion, the inflammatory response, in turn, downregulates FXR via NF-κB activation (Wagner et al., 2008; Y.-D. Wang, Chen, Wang, et al., 2008). Exemplarily, this connection further highlights the versatility and importance of FXR. It further provides another challenge for specific drug intervention as changes in FXR activity can have implications in other important pathways. On the other hand, it could open up novel therapeutic options in diseases associated with intestinal inflammation such as inflammatory bowel disease (Gadaleta et al., 2011).

In summary, research into the FXR network has revealed a complex system with FXR as a central regulator, in which cellular responses are dependent on the tissue, the isoform expression, the ligands, and the nuclear interaction partners. Within human hepatocytes, bile acid binding to FXR induces upregulation of BSEP, MDR3, and OSTα/β (Ananthanarayanan et al., 2001; L. Huang et al., 2003; Landrier et al., 2006). Simultaneously, a negative feedback loop

is triggered that acts via upregulation of SHP to lower CYP7A1 and NTCP levels (Goodwin et al., 2000) (**Figure 10**).

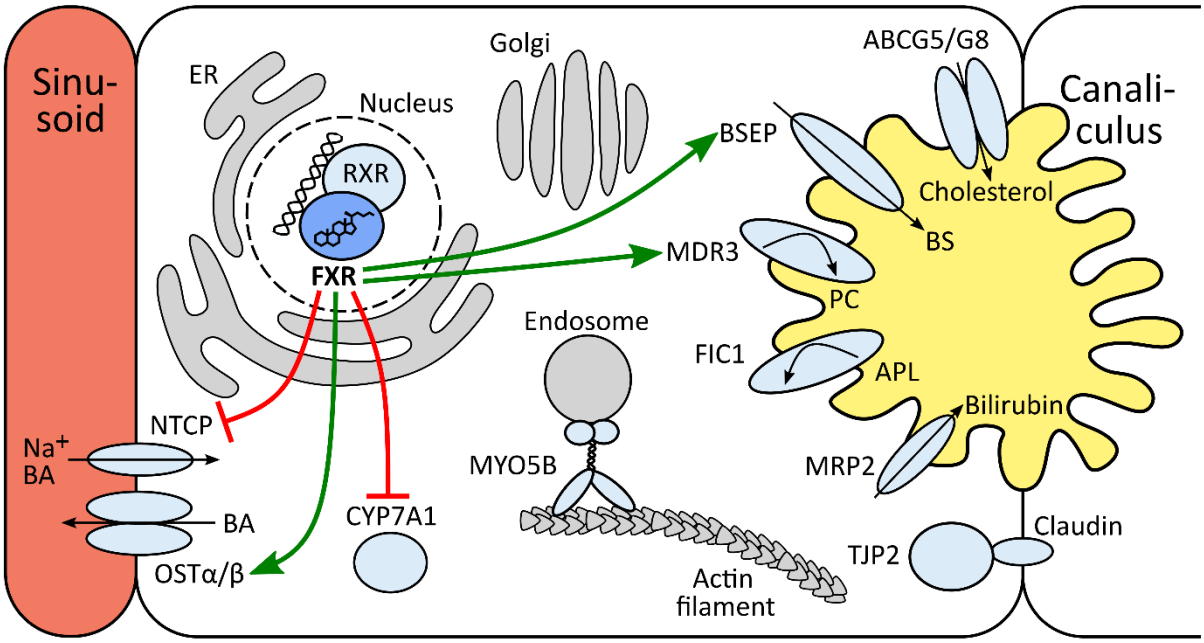


Figure 10: FXR-regulated network within hepatocytes. FXR (marked in dark blue with BA-ligand bound) acts as a central regulator, upregulating key proteins for increasing bile acid efflux and bile formation towards the canaliculus via BSEP (Ananthanarayanan et al., 2001) and MDR3 expression (L. Huang et al., 2003; Ijssennagger et al., 2016) and efflux of bile acids towards the blood stream via the expression of OSTα/β (Landrier et al., 2006). Through SHP expression, FXR additionally downregulates the uptake of bile acids from sinusoids via NTCP and represses the building of novel bile acids via CYP7A1 downregulation (Goodwin et al., 2000). Accordingly, FXR senses the current bile acid levels and acts to prevent bile acids reaching toxic levels within the cell.

Based on these regulatory links, the expression of downstream gene targets of FXR correlate with the transactivation activity of FXR and can be measured accordingly. Luciferase-based transactivation assays have been well established for FXR and other nuclear receptors (Cui et al., 2002; Elbrecht et al., 1999), and allow measuring the activity of FXR on different gene targets. In short, investigated cell lines are usually transfected with plasmids encoding for FXR and its canonical binding partner RXR to ensure overexpression. Additionally, a plasmid encoding for luciferase is used under the promoter control of a known FXR gene target, e.g., BSEP or SHP promoter. Using this system, the activity of FXR can be investigated based on its binding to the promoter sequence, which induces the expression of luciferase. Based on normalization to a control luciferase signal, the effect of different ligands for FXR or amino acid substitutions in FXR can be analyzed. We employed this assay to study a FXR variant in

liver-relevant isoforms and decipher its effect on two well-known gene targets, BSEP and SHP (see Publication II, Chapter 5).

In the following subchapters, I will give a short overview of common FXR ligands, both agonists and antagonists, before discussing the structure of the LBD in detail, as it is crucial for protein activity.

2.3.3 Diversity of FXR ligands

Due to its importance within metabolism, FXR is a promising target for drug intervention in metabolic disorders (Claudel et al., 2002; S. Fang et al., 2015) and liver diseases (Adorini et al., 2012; Merk et al., 2019). However, the complexity of the system makes detailed studies necessary to minimize side effects. For example, elevated cholesterol levels have been described in clinical trials of the FXR agonist obeticholic acid (OCA) (Neuschwander-Tetri et al., 2015) due to overactivation of FXR and its subsequent effect of blocking bile acid synthesis, which uses cholesterol as a primary building block. Accordingly, current research often focuses on finding partial agonists or on identifying selective bile acid receptor modulators (SBARMs) designed to activate or repress certain FXR functions (Massafra et al., 2018; Merk et al., 2019). Many FXR ligands are based on the steroidal backbone, building on the endogenous bile acid ligand's structure. In decreasing potency, these endogenous bile acids are chenodeoxycholic acid (CDCA), deoxycholic acid (DCA), lithocholic acid (LCA), and cholic acid (CA) (Makishima et al., 1999; Parks et al., 1999; H. Wang et al., 1999), with the secondary bile acids DCA and LCA being generated from the primary ones CDCA and CA, respectively (Fiorucci et al., 2020; Jiang et al., 2021). Derived FXR ligands, as well as the endogenous ones, often have poor aqueous solubility and bioavailability and show promiscuity towards the G protein-coupled bile acid receptor (GPBAR1, also called Takeda G protein-coupled receptor 5 (TGR5)) (Kawamata et al., 2003; Massafra et al., 2018). Targeting both receptors is not necessarily an unwanted off-target effect. The dual ligand for FXR and TGR5, BAR502, showed positive results in a nonalcoholic steatohepatitis (NASH) mouse model, interestingly without triggering pruritus (Carino et al., 2017; Cipriani et al., 2015). However, since the activation of TGR5 has been linked to pruritus (or, in layman's terms, itch) in mice (Lieu et al., 2014) and pruritus was a frequent side effect of OCA treatment in a primary biliary cholangitis (PBC) clinical trial (Markham & Keam, 2016), dual agonistic ligands are not the answer for every hepatic disorder.

Nonsteroidal FXR agonists became of interest to increase selectivity (Jiang et al., 2021). The molecule GW4064 was developed as a potent and specific FXR agonist (Maloney et al., 2000), although studies also indicate histamine receptors as additional targets for GW4064 (N. Singh et al., 2014). Nonetheless, GW4064 is frequently used as an investigational tool and as a lead structure for developing agonists that overcome its predecessor's limitations (Jiang et al., 2021). While nonsteroidal FXR agonists do not show TGR5 induction, they do need to avoid the pitfall that complete FXR activation leads to elevated cholesterol levels. Partial agonists are therefore of increasing interest for fine-tuning FXR functions. DM175, a nonsteroidal molecule, induced a conformational change in FXR different to endogenous CDCA binding and exhibited a partial agonistic and partial antagonistic profile (Merk et al., 2019). TERN-101 is another potential partial FXR agonist (Genin et al., 2015), currently in clinical trials for NASH (Y. Wang et al., 2021). On the other hand, FXR antagonists are useful for elucidating physiological functions, shedding light on molecular mechanisms, and balancing the activity state of FXR. Guggulsterone is a natural compound that has been identified initially as a FXR antagonist (Urizar et al., 2002), while later studies identified it as a likely SBARM, as it further enhanced BSEP expression in presence of other FXR agonists (Cui et al., 2003). Due to its high affinity towards other NRs (Burris et al., 2005), its usefulness as selective FXR ligand is limited. Often, compounds need to be reclassified due to novel insights and so far, no antagonist has been found to block all FXR targets. For the nonsteroidal compound ivermectin, despite being initially identified as a partial agonist (Jin et al., 2013), it has been referred to as FXR antagonist in the literature as an ivermectin-bound crystal structure showed preferred corepressor binding and a dynamic helix 12, indicative of the inactive state (Jiang et al., 2021; Jin et al., 2013). Potentially, ivermectin acts in a tissue-specific fashion, with high activity in the intestines while displaying lower effects in the liver (Jin et al., 2015), thus highlighting its potential use as a SBARM. However, antagonistic effects of ivermectin on other NRs, namely LXR and PXR, have also been identified (Hsu et al., 2016) and have to be taken into account. These examples already indicate the troubles of FXR ligand research, as meticulous efforts must be undertaken to study the ligand effect on different subsets of FXR targets, within different tissues, evaluate protein interaction partner binding and analyze off-target binding to other NR or other proteins. Nonetheless, FXR ligand research is an ongoing topic due to FXRs widespread functions and associated promises in disease amelioration.

In Publication II (Chapter 5) we were, amongst others, interested in the transitioning of FXR from the inactive to the active state. Accordingly, I used CDCA as strongest endogenous FXR agonist in MD simulations to drive the system towards the active state. In *in vitro* experiments, we (cellular assays performed by Dr. Jan Stindt (Heinrich Heine University Düsseldorf, Germany) and Dr. Alex Bastianelli (Otto von Guericke University Magdeburg, Germany), recombinant protein purification and assay performed by me) used OCA as a well-established and potent agonist to analyze protein activity and variant effects on ligand binding. While there are many ligands available, ligands were chosen here based on their closeness to the *in vivo* situation and potency to maximize the signal and drive the protein to activity. OCA has an increased potency of roughly 100-fold over CDCA, while structurally it remains a close analog of CDCA with only an additional ethyl group at C_6 (Pellicciari et al., 2002). In line with using the promoter sequences of established FXR targets BSEP and SHP (see Chapter 2.3.2 and Publication II, Chapter 5), OCA has been shown to upregulate both BSEP and SHP expression (Y. Zhang et al., 2017).

Besides broadening the spectrum of research tools and potential treatment options, research into FXR ligands has increasingly also provided information on the structural basis and molecular mechanisms of FXR activation. Due to the structural similarity within the NR superfamily (R. Kumar & Thompson, 1999; Weikum et al., 2018), certain mechanisms can potentially be inferred from and transferred to other NRs.

2.3.4 Protein structure and conformational states of FXR

Almost all proteins of the superfamily of NRs share the overall architecture of the N-terminal domain containing the AF1 region, DBD, followed by a hinge region and the LBD with a C-terminal AF2 domain (except the receptors SHP and DAX) (Weikum et al., 2018). The N-terminal AF1 domain shows low structural order, and as such, efforts to determine its structure have been unsuccessful so far. Its flexibility likely enables different transit interaction surfaces, providing the possibility for interacting with a broader spectrum of binding partners (Simons et al., 2014). As isoforms differ in the N-terminal domain (Ramos Pittol et al., 2020) and several functionally important PTMs have been identified (Anbalagan et al., 2012; Appelman et al., 2021), this region has a certain influence which downstream gene targets are controlled by the NR. However, DNA binding – and thus target gene

determination – is mainly controlled by the highly conserved DBD (Devarakonda et al., 2003). Based on the few structural data available on (almost) full-length NRs, namely peroxisome proliferator-activated receptor γ (PPARγ) and hepatocyte nuclear factor 4α (HNF4α), domain-domain interactions between DBD and LBD occur and critically affect the activity (Chandra et al., 2008, 2013; Simons et al., 2014). Transference towards FXR or other NRs, however, is difficult as it seems likely that domain-domain interactions change depending on the binding partner. Accordingly, the heterodimer PPAR/RXR displayed different domain-domain interaction patterns than the HNF4α homodimer (Chandra et al., 2008, 2013; Simons et al., 2014). In solution, structural analyses emphasized the importance of the hinge region for the integrity of the DNA-bound structure and further pointed to the fact that different DNA-binding elements lead to different conformations within NRs (Rochel et al., 2011). These revelations further highlight the overall flexibility of NRs and, for the canonical heterodimer FXR/RXR (Forman et al., 1995) compared to FXR as a monomer (Shen et al., 2008), could explain differential DNA-binding preferences with potentially different structural conformations. While the dimerization mechanism and interface between the FXR-DBD and RXR-DBD are unresolved to date (Jiang et al., 2021), crystal structure determination on the FXR/RXR LBD complex revealed stabilizing effects of the RXR LBD on the active conformation of the FXR LBD (Zheng et al., 2018). Similar to other NR heterodimers (Gampe et al., 2000; Svensson et al., 2003), FXR/RXR LBD dimerization relies especially on interactions between the helix 10 of both receptors (Zheng et al., 2018).

The LBD is critical for the overall protein activity based on ligand binding and interactions with coregulator proteins, either coactivator proteins like the nuclear receptor coactivator 2 (NCoA2) or nuclear receptor corepressor proteins (NCoR) (Jiang et al., 2021; Zheng et al., 2018). Accordingly, intense research focused on elucidating structural features and molecular mechanisms in the LBD of NRs. Overall, the LBDs of NRs show high structural similarity, with twelve α-helices folded in a three-layered arrangement. Of special interest for protein activity is the short C-terminal α-helix, the helix 12 (H12). In the active state, H12, together with parts of helix 3 and helix 4, forms part of the activation function 2 (AF2) surface, a binding surface for nuclear coactivation proteins to enhance transcriptional initiation (Aranda & Pascual, 2001; Mi et al., 2003). Accordingly, deletion of H12 within HepG2 cells abolished FXR transactivation activity, i.e., its ability to bind to its DNA response elements (Ananthanarayanan et al., 2001). Coactivators bind to the hydrophobic AF2 surface groove

with an α-helix containing a signature LXXLL motif (where X can be any amino acid) (Heery et al., 1997). Corepressors interact via a larger (L/I)XX(I/V)I or LXXX(I/L)XXX(I/L) motif, thus blocking sterically H12 positioning (Nagy et al., 1999). A mousetrap mechanism was initially proposed to explain the underlying molecular mechanism, in which an unliganded and inactive LBD with an extruding H12 (pointing away from the LBD) would transition to the active state upon ligand binding with an LBD-bound H12. This theory was proposed based on crystal structures of apo and agonist-bound NR RXR (Renaud et al., 1995). An alternative model, termed dynamic stabilization, argues for a highly flexible H12 in the apo state, which shifts towards a stable active conformation upon ligand binding (Kallenberger et al., 2003). In contrast to the mousetrap model with its two specific stable states, the dynamic stabilization model is characterized by a highly mobile and unstructured H12 (Weikum et al., 2018). A flexible and likely unstructured H12 in the apo state is supported by studies using fluorescence spectroscopy (Kallenberger et al., 2003) and nuclear magnetic resonance in PPARγ and RXR (Hughes et al., 2012; X. Yan et al., 2004). Additionally, studies observed a H12 positioning towards the LBD within the apo state in several NRs (thyroid hormone receptor (Figueira et al., 2011), estrogen receptor (Dai et al., 2009), FXR (Merk et al., 2019)). Taken together, it seems likely that in an unliganded state, H12 of the FXR LBD moves flexibly and visits the active conformation with some regularity but does not remain stably in this conformation in the absence of a ligand or a coactivating protein to stabilize the state. This is corroborated by the identification of transient interactions of H12 to the FXR LBD core in the absence of agonists in an NMR study (Merk et al., 2019). Additionally, the presence of crystal structures of apo FXR LBDs associated with a nuclear coactivation peptide indicates its ability to interact with coactivators even in the absence of ligands (Gaieb et al., 2018; Merk et al., 2019). However, this recruitment of coactivation protein was not observed in NMR studies and might represent a crystallization artifact (Merk et al., 2019). Based on the comprehensive study by Merk et al., apo FXR can bind corepressor, and subsequent agonist binding induces conformational changes leading to weakened interactions to the corepressor peptide, shifting the balance to preferred coactivator peptide binding (Merk et al., 2019). Binding of the coactivator peptide has a greater influence on the stability of the active conformation than the ligand binding itself has; while ligand binding increases the propensity of the protein to associate with the coactivator, it can still partly bind the corepressor. Antagonist binding, however, stabilizes interactions with the corepressor so that even in the presence of coactivators, the protein will

stay bound to the corepressor. Partial agonists infer their function due to conformational changes in which the LBD has partly affinity to the corepressor and partly to the coactivator (Merk et al., 2019).

MD simulations have been increasingly used to analyze the dynamics of NR LBDs and elucidate the influence of ligands, coactivators, or corepressors binding as well as NR heterodimerization (Chrisman et al., 2018; Díaz-Holguín et al., 2023; Heidari et al., 2019; Kumari et al., 2021, 2023; Saen-Oon et al., 2019). In a comprehensive study by Chrisman et al. on the PPARγ LBD, MD and NMR data confirmed its structural flexibility, indicating a range of possible conformations available to the protein (Chrisman et al., 2018). The AF2 surface, including the H12, switches rapidly between several conformations in the µs to ms time range in the apo state. Agonist or inverse agonist binding, however, limits the available conformations with only rare switching (Chrisman et al., 2018). MD simulation studies on the heterodimer FXR/RXR and the FXR monomer further indicated a destabilization of the H12 in antagonist-bound states compared to agonist-bound states, as well as changes in the interaction interface between FXR and RXR (Díaz-Holguín et al., 2023). Overall, this further strengthens the picture of the LBD of NR as a flexible module in the apo state, moving relatively freely between conformations. Ligand binding and coregulatory binding limit this flexibility and push the system towards specific conformations. Within a study employing MD simulations and NMR techniques for the PPARγ protein, the authors observed a conformational change from inactive to an almost-perfect placement of the H12, potentially representing the active state, in a system with an inverse agonist and corepressor peptide present (Chrisman et al., 2018). However, revealing the dynamic pathway from inactive to active conformation in MD simulations has – prior to Publication II (see Chapter 5) – not been shown for FXR. Furthermore, the influence of variants on the FXR function has not been studied in depth using MD simulations so far.

2.3.5 Dysfunction of FXR

FXR dysfunction can severely affect the intricate network of bile regulation. Accordingly, several FXR variants have been identified in intrahepatic cholestasis of pregnancy (ICP) (van Mil et al., 2007). Although ICP is usually transient, affected patients have an increased risk of developing other liver-associated diseases (Ropponen et al., 2006). Additionally, FXR has been

linked with various cancers (Girisa et al., 2021; Kainuma et al., 2018; You et al., 2019) with 172 mutations listed in the cBioPortal database for cancer genomics (Cerami et al., 2012; Gao et al., 2013), out of which 131 are missense mutations, 33 are truncations, and 8 mutations affect splicing. Overexpression of FXR was identified in breast, lung, and pancreatic cancer and was associated with increased proliferation (Girisa et al., 2021; You et al., 2017, 2019) and increased epithelial-mesenchymal transition in hepatocellular carcinoma (HCC) (Kainuma et al., 2018). Within PFIC5, identified FXR mutations were leading to a premature stop codon and truncation of the protein (p.Arg176*) or to in-frame insertion on one chromosome (p.Tyr139_Asn140insLys) and a partly deletion (first two exons of FXR) on the other chromosome, affecting FXR function to a high degree (Gomez-Ospina et al., 2016). PFIC5 clinically presents with liver dysfunction at an early age with severe cholestasis, accumulation of bile acids in hepatocytes resulting in elevated aminotransferases levels, and low bile salt export pump (BSEP) expression (Gomez-Ospina et al., 2016). In the HiChol consortium, a rare homozygous variant in FXR was identified in a patient presenting with a clinical phenotype in line with PFIC5 (Pfister et al., 2022). However, the molecular pathomechanism was unknown for the variant.

Chapter 3 Scope of the Thesis

Within my work for the HiChol consortium, I focused on the investigation of variant impact in the proteins MDR3 and FXR. Despite its importance within the liver and frequent association of variants with liver diseases, there is no well-established protein predictor for MDR3 (see Chapter 2.2). A proposed predictor, MutPred, was tested mainly on pathogenic variants and lacked testing over a higher number of variants (Khabou et al., 2017). Based on the extensive research over the years on MDR3 variants within the field of liver research and advances in ML on small datasets, the possibility to establish a protein-specific dataset to enable machine learning-based classification of variants arose. While this approach does not reach the level of depth as single variant studies can provide, it has the advantage of being applicable to future novel identified variants outside the direct project time scope. The protein-specific predictor should satisfy strict criteria. First, it needs to outperform general protein predictors such as the previously proposed general predictor MutPred. Second, it is desirable that the tool can classify any variant possible within the protein. Additionally, this implies that the predictor should be sensitive to any potential pathogenic variant and not limited to a specific liver disease. Accordingly, the training dataset will need to be assembled from MDR3-affected liver diseases without the limitation to PFIC3. Third, it needs to be easy to use to enable wide usage and easy interpretation. The project is described in detail in Chapter 4.

The identification of a homozygous variant in FXR identified in a PFIC type 5 presenting patient (Pfister et al., 2022) demanded an in-depth analysis to understand the molecular mechanism. Accordingly, a collaborative strategy was established to employ cellular and protein assays, analyze patient tissue samples, and perform MD simulations in order to investigate the variant effect and unravel its molecular pathomechanism. Based on the clinical presentation, the focus was put on liver-relevant isoforms with associated downstream targets and known ligands (see Chapter 2.3.2 and Chapter 2.3.3). Further, analyzing the variant in the inactive and the active state using MD simulations may provide a deeper understanding of misfunction (see Chapter 2.3.4). The project is described in detail in Chapter 5.

Chapter 4 Publication I

Vasor: Accurate prediction of variant effects for amino acid substitutions in multidrug resistance protein 3

A. Behrendt, P. Golchin, F. König, D. Mulnaes, A. Stalke, C. Dröge, V. Keitel, H. Gohlke.

Original publication. Contribution: 40%

Study conception and design: A. Behrendt, P. Golchin, A. Stalke, C. Dröge, V. Keitel, H. Gohlke; data collection: A. Behrendt, P. Golchin; analysis, interpretation, and visualization of results: A. Behrendt, F. König, D. Mulnaes; draft manuscript preparation: A. Behrendt, H. Gohlke. All authors contributed to scientific discussions, reviewed the results and approved the final version of the manuscript.

(I adapted parts of the following text and figures from the respective publication.)

4.1 Background

The prediction of an amino acid missense substitution within a protein has received much attention in the last decades due to the rapidly increasing identification of genetic variations based on large sequencing efforts (F. S. Collins & Fink, 1995; Gudbjartsson et al., 2015; Oh et al., 2020; T. Singh et al., 2022; Trubetskoy et al., 2022). Since not every substitution can be analyzed by time- and cost-consuming *in vitro* assays, *in silico* tools provide important information and can narrow down substitutions for further subsequent analysis (Thusberg & Vihinen, 2009). General protein predictors, designed and trained to predict effects for any given protein, often show varying performance when tested on individual proteins (Riera et al., 2016; Choudhury et al., 2022; Livesey & Marsh, 2023). Furthermore, predictors do not guarantee coverage of every possible substitution (Riera et al., 2016). Of note, while there is a tendency for protein-specific predictors to rank higher than general predictors, they do not outperform general predictors in every case (Riera et al., 2016), and as such, careful evaluation for every protein is needed. Transporting phosphatidylcholine from the inner canalicular leaflet to the outer, the MDR3 protein performs an essential function within bile homeostasis

(Boyer, 2013; A. J. Smith et al., 1994; van Helvoort et al., 1996). Dysfunction, thus, is linked to a range of liver diseases such as PFIC, cholelithiasis, cholestasis, cirrhosis, DILI, LPAC, ICP, and HCC (Boyer, 2013; Deleuze et al., 1996; C. Dong et al., 2020; Dröge et al., 2017; Gotthardt et al., 2008; Gudbjartsson et al., 2015; Lang et al., 2007; Pauli-Magnus et al., 2004; Rosmorduc et al., 2001; Stättermayer et al., 2020). Given a genetic cause, the majority of cases (an estimation of 70%) are caused by amino acid substitutions (in the following referred to as 'variant') (Delaunay et al., 2016). However, a reliable (general or protein-specific) predictor with specific evaluation on MDR3 prediction performance is missing, although it would provide a valuable tool for clinicians and researchers. The general predictor MutPred was proposed as a reliable predictor for MDR3 variants based on a group of 21 variants (Khabou et al., 2017; B. Li et al., 2009), but the small size of tested variants as well as a bias towards pathogenic variants within this group were not addressed and may hamper generalization. Accordingly, we set out to establish an ML model to classify variants into the categories benign or pathogenic while comparing our model to the updated version of MutPred, MutPred2 (Pejaver et al., 2020), as well as other integrated general protein predictors.

4.2 Results

Creation of an MDR3-specific dataset

To create a basis for an ML model, I first constructed a dataset specifically for MDR3 variants. Obtaining variants based on literature search allowed the exclusion of variants with no clear disease association (i.e., no *in vitro* verification and no information on clinical indications for disease association), creating a manually curated dataset (**Figure 11**, A). Due to the scarceness of well-studied benign variants, I additionally resorted to known variants from the Genome Aggregation Database (gnomAD), a database based on large-scale genome sequencing projects where pediatric disease patients and their close relatives have been excluded (Karczewski et al., 2020). Despite the possibility of a few disease-associated variants being included in the gnomAD dataset, the benefit of increasing a dataset with highly likely benign variants currently outweighs the risk, and as such, inclusion is a common strategy in ML approaches (Ioannidis et al., 2016; Jagadeesh et al., 2016; Livesey & Marsh, 2023; Wu et al., 2021). While, in principle, a filter step screening out low allele frequency variants would lower the risk of disease-associated variants within gnomAD, it drastically reduces the number of

obtainable benign variants. Similarly, others have refrained from using such an allele frequency filter with parallel reasoning (Livesey & Marsh, 2023). In order to further exclude possible false negative variants from the obtained set of gnomAD variants, I further filtered the variants using the VarSome platform (Kopanos et al., 2019), a tool following the ACMG-AMP guidelines (Richards et al., 2015) to classify variants, and variants with likely pathogenic score were excluded (**Figure 11**, A). While such methods for balancing benefits and risks, quality and quantity of datasets, are currently often unavoidable, advances in multiplexing assays may provide help in the future (Esposito et al., 2019; Starita et al., 2017; Weile & Roth, 2018). The final high-quality dataset contained 85 pathogenic and 279 benign variants. Mapping the variant locations on to the structure of MDR3 revealed a good distribution over the entirety of the protein, with no distinct clustering of benign or pathogenic variants (**Figure 11**, B). While such clustering can occur in certain areas, for example, for pathogenic variants within ligand-binding pockets in cancer-related proteins to form aberrant constitutively active proteins (Niu et al., 2016), it could also introduce unwanted hidden bias for an ML model.

Establishing predictive features

Next, we* used established general protein predictors to predict the variants and established further informative features, namely post-translational modifications (PTM) site impact, variant location within α-helical or β-sheet secondary structure, and residue solvent accessibility (**Figure 11**, A).

* Integrating predictors and other informative features was performed by P. Golchin and A. Behrendt.

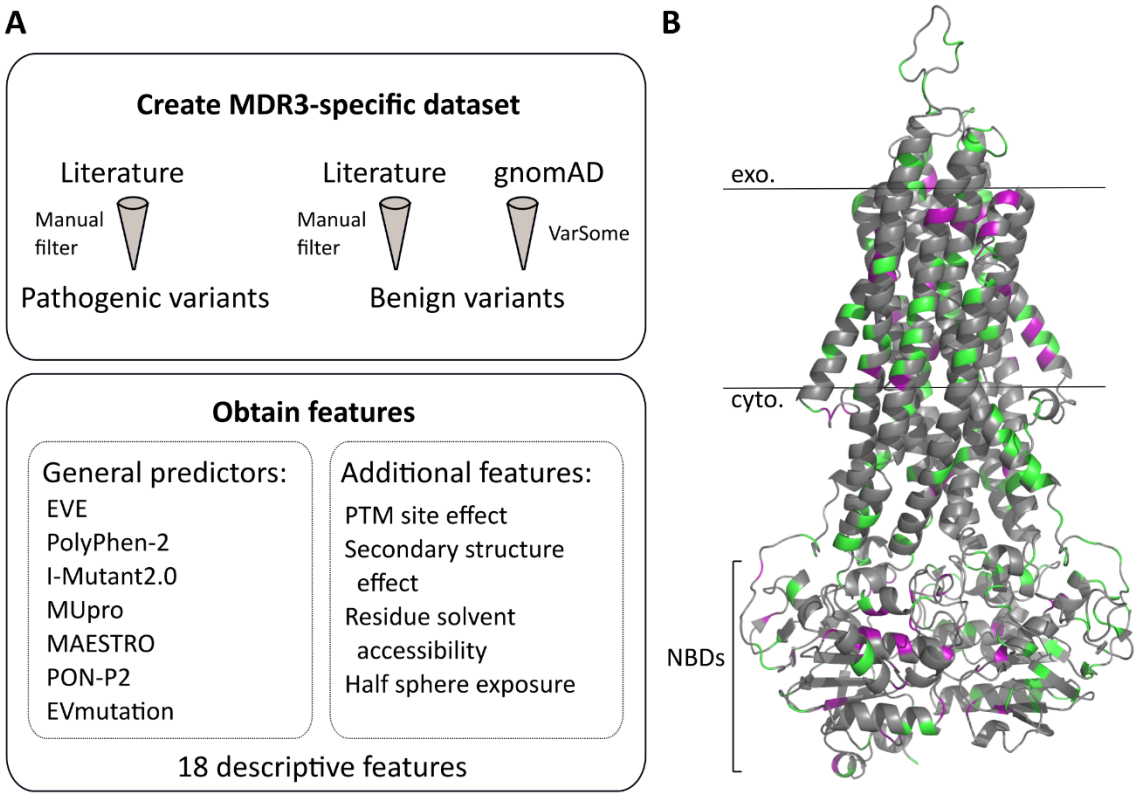


Figure 11: Generation of an MDR3-specific dataset. [A] Overview of the creation of the dataset with variants from the literature and the database gnomAD and the establishment of features. [B] Mapping of included variants within the dataset onto the protein structure revealed a good distribution of benign (green) and pathogenic (purple) variants.

The general protein predictor EVE, a multiple sequence alignment-based classifier trained using an unsupervised ML approach, was integrated as a feature. The naïve Bayes classifier PolyPhen-2 (Adzhubei et al., 2010), frequently used for clinical variant interpretation (Gunning et al., 2021), predicts variant impact based on sequence and structural considerations (Adzhubei et al., 2010). I-Mutant2.0 and MUpro both employ support vector machine approaches to predict stability changes of proteins (Capriotti et al., 2005; Cheng et al., 2006). The tool MAESTRO uses a combination of ML models to derive predictions of stability changes upon point mutations, including a confidence score (Laimer et al., 2015). Using evolutionary conservation information, biochemical considerations, and (functional) annotations, PON-P2 classifies variants based on a random forest classifier (Niroula et al., 2015). EVmutation specifically includes residue interdependencies, showing improvements over using only evolutionary conservation features, and derives predictions using an unsupervised statistical model (T. A. Hopf et al., 2017). A specific feature for PTM sites was derived from literature knowledge and predicted PTM spots from PhosphoMotif (Amanchy et al., 2007),

PhosphoSitePlus (Hornbeck et al., 2015), NetPhos (Blom et al., 1999) and the Eukaryotic Linear Motif database (M. Kumar et al., 2019). Using the database of secondary structure assignments DSSP (Joosten et al., 2011; Kabsch & Sander, 1983), the secondary structure for the MDR3 protein (Protein Data Bank identification number 6S7P (Olsen et al., 2020)) was extracted and further used for a rudimentary feature of secondary structure impact and calculation of relative solvent accessibility (RSA). RSA was calculated using DSSP-based residue exposure divided by the maximal residue solvent accessibility (Tien et al., 2013). Half-sphere exposure (HSE), a measure derived to surmount RSA limitations in measuring residue solvent exposure, was implemented using the biopython HSExposure module (Hamelryck, 2005). In preparation for ML, the obtained dataset with the features was cleaned from non-numerical values.

Establishing a well-balanced test set

Creating a sensible test set is not always straightforward. Considerations range from size to class distribution within the test set, and often, the answers depend on the individual research question and on the available dataset (Dobbin & Simon, 2011). Borrowing from the Pareto principle, people often use an arbitrary split of 80/20 for dividing a dataset into training and test set (Joseph, 2022). Due to the relatively small overall dataset, the test set was designed to contain 40 variants with equal class distribution (20 benign variants and 20 pathogenic variants). In order to avoid biases within the test set towards the overrepresentation of specific amino acids – and the potential exclusion of other amino acids – I established the test set by performing a root-mean-square deviation (RMSD)-based minimization of the amino acid distribution within the test set against the overall dataset. As a first step, 10 variants were randomly chosen for the test set to calculate an initial amino acid distribution for comparison with the distribution of amino acids within the entire dataset. Following, randomly chosen variants were only transferred into the test set if the RMSD decreased or only marginally increased (as otherwise, the limited size of the dataset could have resulted in failures to generate a test set). Using such an approach, I ensured that the test set included a good distribution of variants (**Figure 12**), and the resulting test set was withheld from machine learning until the final validation of the model.

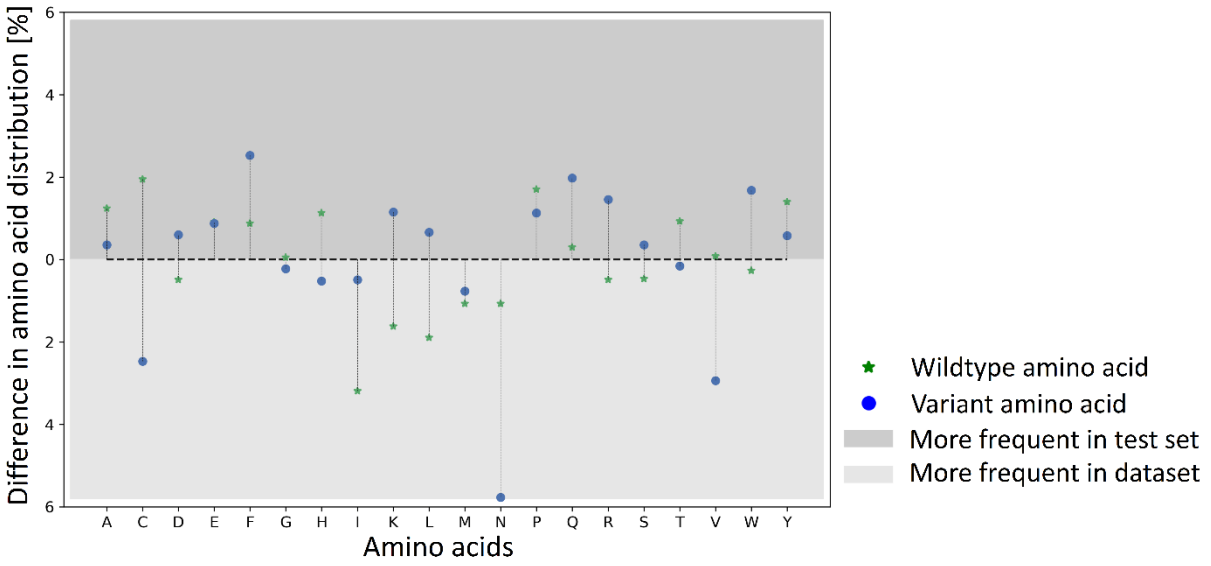


Figure 12: Comparison of the amino acid distributions. Plotted differences between amino acid distributions, both for the wildtype (green star) and variant amino acid (blue dot), with points above the horizontal dotted line indicating a higher representation in the test set and below the line a higher representation in the overall dataset. Due to the limited number of variants, it was not possible to minimize the distribution differences to zero. However, the obtained test set displayed an overall distribution of amino acids similar to that of the general dataset.

Training and evaluation of the ML tool

Since the overall dataset displayed a clear class imbalance (85 pathogenic and 279 benign variants) and such imbalances can influence predictor performance (Wei & Dunbrack, 2013), I employed an established technique to generate synthetic new data points within the N-dimensional data set space with the synthetic minority oversampling technique (SMOTE) (Chawla et al., 2002). Next, the training dataset was used to train an XGBoost model (with a default gradient boosting tree, maximum tree depth set to 3, and a learning rate of 0.02) (Chen & Guestrin, 2016). To evaluate the performance, repeated k-fold cross-validation was used with a split of 3 and the number of repeats set to 5. Performance on the respective internal fold used for evaluation within the cross-validation was visualized using receiver operating characteristics (ROC) curves and compared to the final evaluation on the test set to detect potential overfitting (see Chapter 2.1.1). Further, I calculated the feature importance using two approaches, the XGBoost internal tree-based feature importance and permutation-based feature importance, to reduce the number of features. The four shared least-informative features were removed with marginal impact on model performance. Additionally, such a feature evaluation provides insights into the usefulness of specific features on the overall

prediction outcome and indicated EVE as the most important feature. In the specific case with a relatively small dataset and feature space, a reduction of features is not computationally necessary; however, it is a common practice in the field and aims towards the highest efficiency (Jia et al., 2022). Performance with the reduced number of features was again evaluated with repeated k-fold cross-validation and assessed against the predictions of the final model, termed Vasor (Variant assessor of MDR3), on the withheld test set (**Figure 13**, A). Calculation of the confusion matrix with True Negative (TN), False Positive (FP), False Negative (FN), and True Positive (TP) predictions revealed only four mis-classified variants for the test set (**Figure 13**, B).

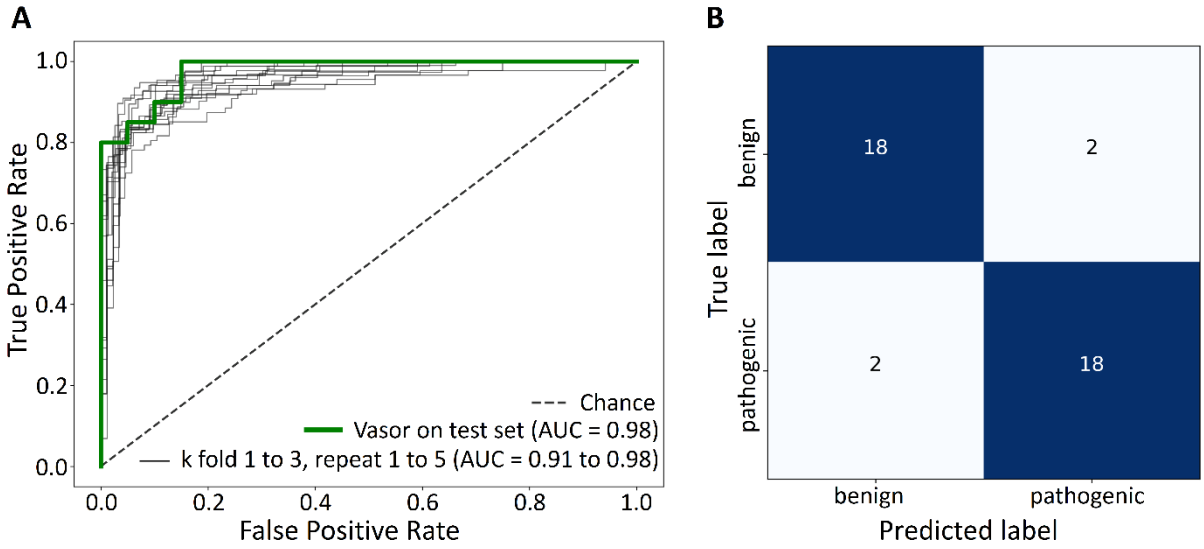


Figure 13: Performance of Vasor. [A] The performance estimations within the repeated k-fold cross-validation (thin black lines) show similar ROC curves and area under the curve (AUC) values as the evaluation on the final test set (thick green line), indicating a well-fit model without over- or underfitting. [B] Confusion matrix of Vasor performance on the test set.

Next, I compared the performance of Vasor against other integrated general protein predictors and against MutPred2 as a previously suggested high-performing predictor on MDR3 (Khabou et al., 2017) (**Figure 14**). In line with other studies identifying a combination of general predictors (meta-predictors) to outperform their individual contributors (Broom et al., 2017; Gunning et al., 2021), Vasor achieved the highest ROC curve and highest AUC value (**Figure 14**, A). Looking at the coverage of the predictors, EVE and PON-P2 did not derive predictions for the full dataset (**Figure 14**, B). To obtain a fair comparison of predictors, the ROC curves and precision-recall-curves indicative of performance (**Figure 14**, A and C) were normalized to the coverage of the dataset. Vasor outperformed the closest competitor, EVE,

based on performance scores and additionally on protein coverage (**Table 1**). While EVE achieved the lowest number of FP predictions on the overall dataset, it only covered 85.7% of the dataset and failed to recognize 19 pathogenic variants (FN). Vasor, with 100% coverage of the dataset, achieved low numbers of 14 FN and 12 FP predictions, indicating a good balance. Of note, MutPred2 achieved an admirable low number of only 6 FN predicted variants, but at the expense of a high number of 93 benign variants falsely classified as pathogenic (FP). Accordingly, the superiority of Vasor resulted in the highest values in the weighted measures of F1-score (0.85) and Matthew's correlation coefficient (MCC) (0.80) (**Table 1**).

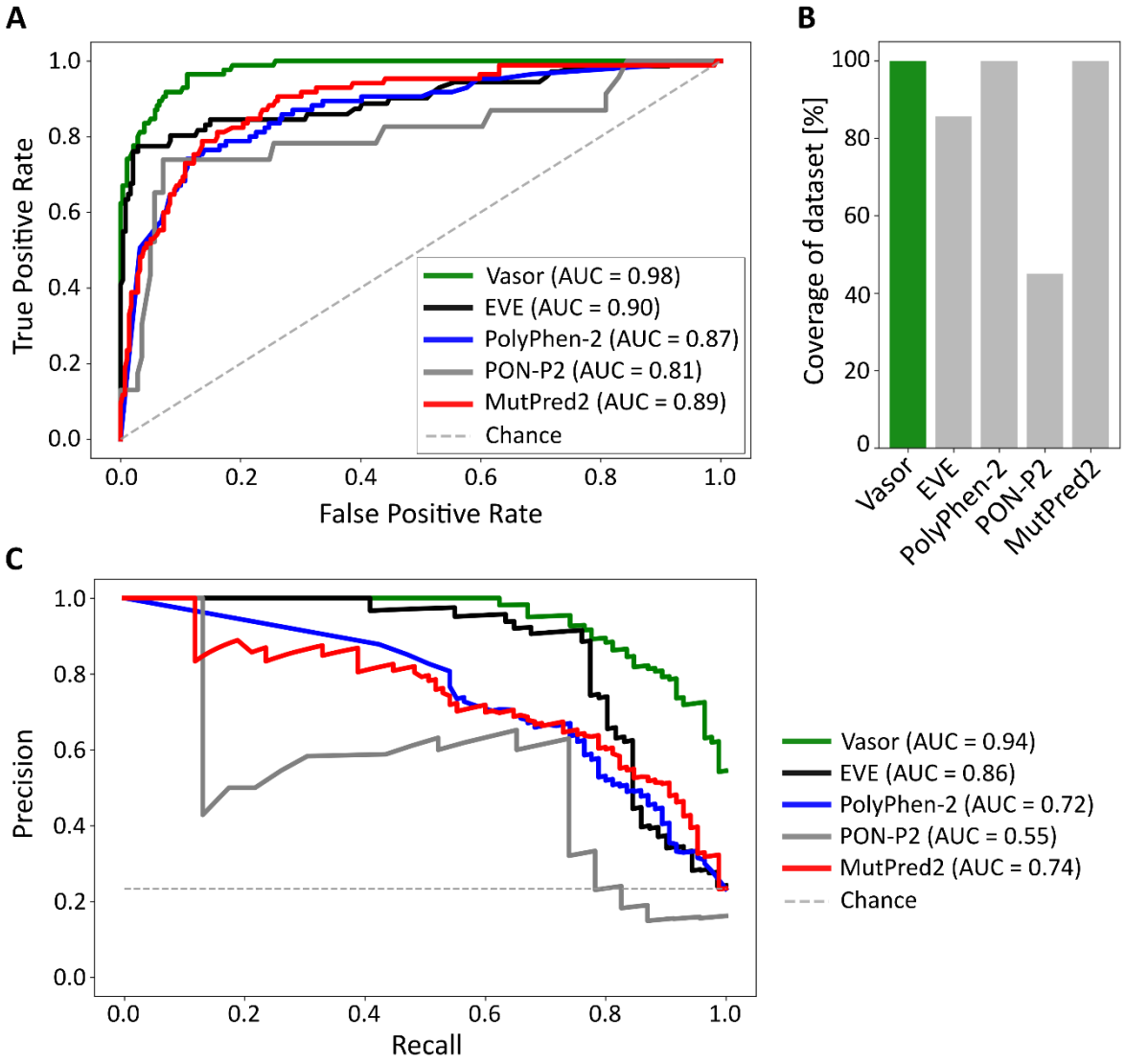


Figure 14: Performance comparison of Vasor against other predictors. [A] ROC curve comparison of Vasor and the integrated general protein predictors EVE, PolyPhen-2, PON-P2, as well as the external MutPred2 predictor. [B] Coverage of the MDR3-specific dataset for the respective prediction tools. [C] Precision-recall curves for the respective predictors on the MDR3-specific dataset. Values for both the ROC curves and precision-recall curves were normalized to the covered set of variants for each predictor, respectively.

Table 1: Detailed performance measures of predictors on the entire dataset.

	Vasor	EVE	PolyPhen-2	PON-P2	MutPred2
Recall	0.84	0.73	0.84	0.74	0.93
Specificity	0.96	0.98	0.74	0.89	0.67
Precision	0.86	0.91	0.49	0.52	0.46
NPV	0.95	0.93	0.94	0.95	0.97
Accuracy	0.93	0.92	0.76	0.87	0.73
F1-Score	0.85	0.81	0.62	0.61	0.61
MCC	0.80	0.77	0.50	0.54	0.51
TP	71	52	71	17	79
FN	14	19	14	6	6
TN	267	236	206	125	186
FP	12	5	73	16	93
Coverage [%]	100	85.7	100	45.1	100

Abbreviations: NPV, negative predictive value; MCC, Matthew's correlation coefficient; TP, true positive; FN, false negative; TN, true negative; FP, false positive.

Of note, such a full description of performance measures is recommended for an accurate judgment of binary predictors (Vihinen, 2012). To further investigate Vasor performance, I assessed how certain Vasor was in its predictions. Accordingly, I assessed its output, the probability of pathogenicity, with values below 0.5 leading to a classification as benign and values above 0.5 leading to a classification as pathogenic. Good predictors show a distinctive clustering towards very low and very high probabilities of pathogenicity (Ioannidis et al., 2016; Pejaver et al., 2017, 2020). Visualizing the probability of pathogenicity for every variant within the dataset as well as the SMOTE-generated points for the minority pathogenic class, Vasor showed high peaks towards low probability and high probability values, with few variants in the range between 0.3 to 0.7 probability of pathogenicity (**Figure 15**). The distribution further

indicated Vasor as a well-performing predictor, classifying the majority of cases with a high certainty.

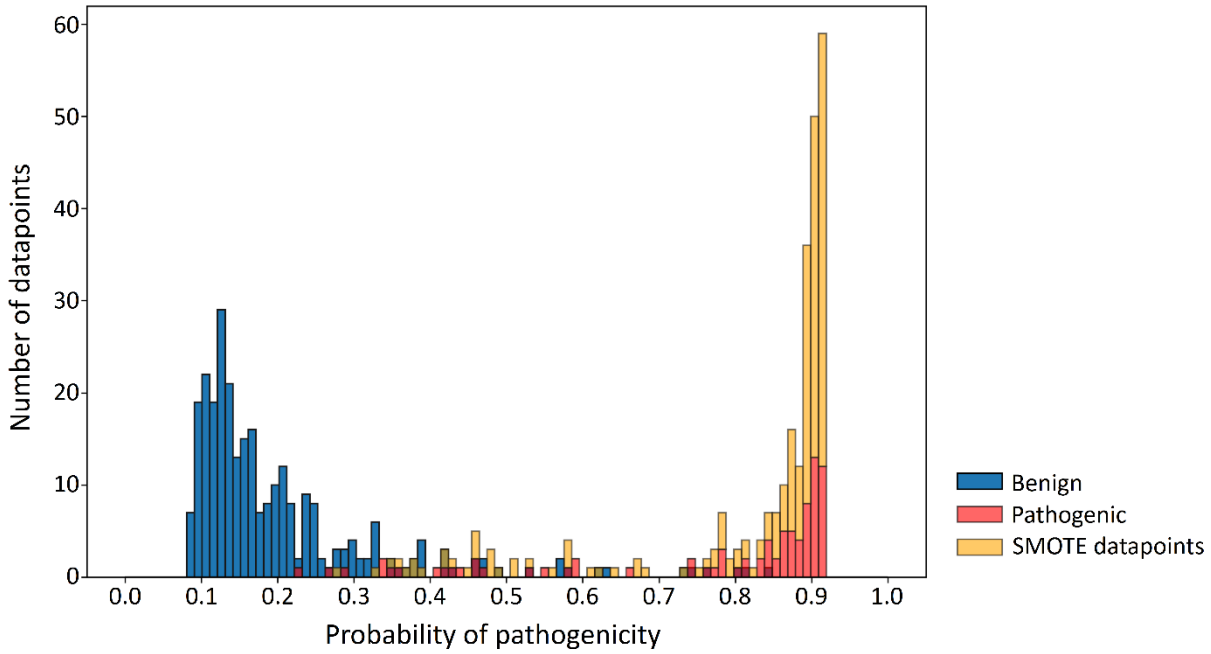


Figure 15: Distribution of probability of pathogenicity values of Vasor. The generated output of Vasor, the probability of pathogenicity, for each benign (blue), pathogenic (red), or SMOTE-generated datapoint for the minority class (orange) showed a good class separation with peaks towards low and high probabilities.

Generating predictions for every substitution and providing easy access to Vasor

Having established a high-performing predictor, I next predicted every possible amino acid substitution within MDR3. This precomputed prediction map was used as the basis for retrieving predictions from the Vasor webserver for rapid assessment of variant impact (accessible at https://cpclab.uni-duesseldorf.de/mdr3_predictor/). We[†] integrated a structure visualization feature specific to the variant entered and offer downloadable enlarged images of the variant and wildtype. Additionally, Vasor can be downloaded and locally installed, allowing users to access the source code. Similarly, visualization of the variant from the webserver can be enhanced by the user based on a downloadable PyMOL script. In general, these steps were taken to enable researchers and clinicians from different fields to use the tool, as ML-based tools often remain cumbersome to handle for non-experts in the field.

[†] Webserver establishment and structural visualization was executed by F. König in accordance with A. Behrendt.

Mapping the average probability of pathogenicity value over every possible substitution for each position back onto the protein structure (**Figure 16**) revealed an additional view of areas of high susceptibility to harmful substitutions.

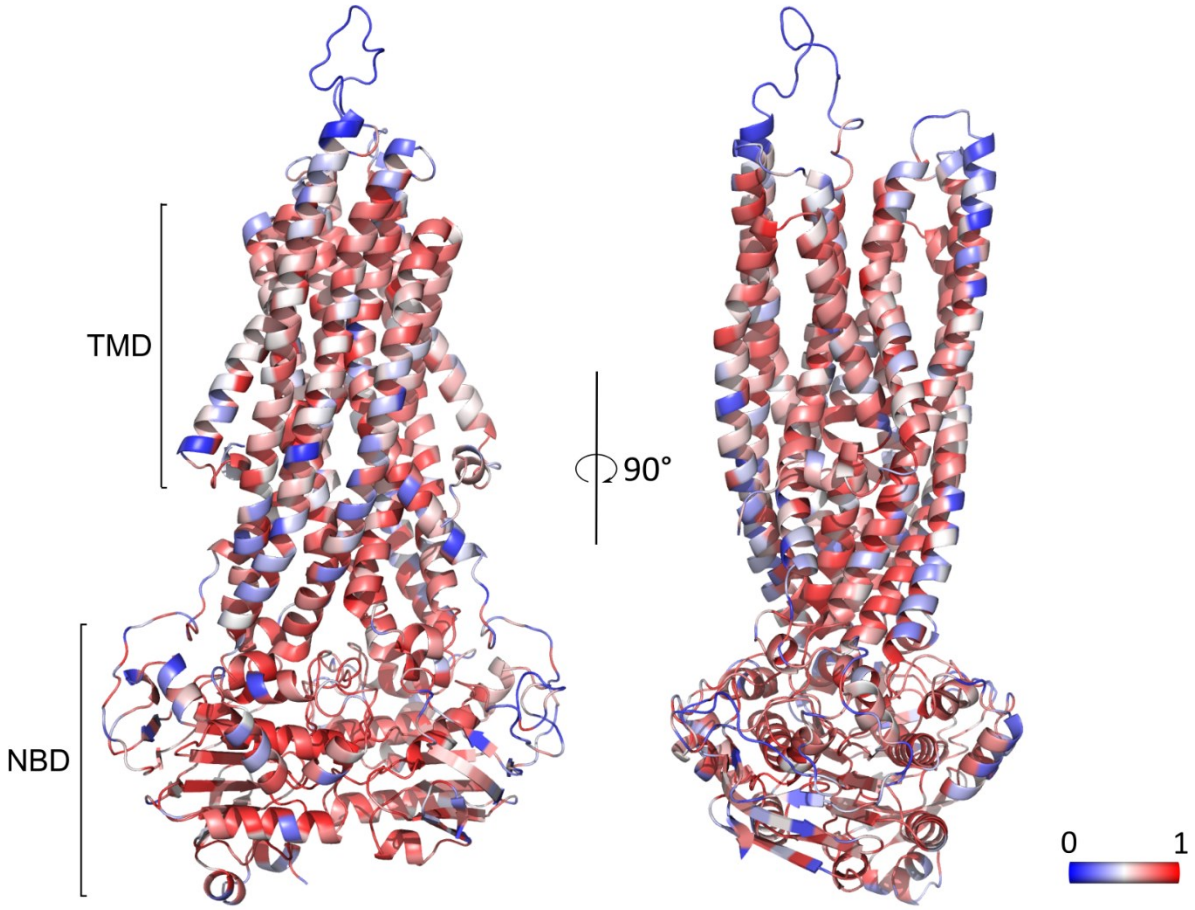


Figure 16: Average probability of pathogenicity per position mapped onto the protein structure of MDR3. Vasor-derived prediction values were averaged over all possible substitutions at each position and color-coded with values closer to 0 (blue), indicating the average probability corresponds to benign predictions, while highly susceptible positions where the average probability corresponds to pathogenic classifications are closer to 1 (red). TMD: transmembrane domain, NBD: nucleotide binding domain.

In line with knowledge about functional motifs, such as Walker A and Walker B (Schmitt & Tampé, 2002), buried residues within the NBDs of MDR3 showed in tendency a high average probability of pathogenicity value, indicating that most substitutions at those position were predicted as pathogenic and, thus, to result in functional impairment. Buried residues within the helices of the TM were predicted as more susceptible to pathogenic substitutions than more exposed residues of the protein, in line with previous studies on RSA and evolutionary conservation (Franzosa & Xia, 2009; Ramsey et al., 2011). Variants within the helices forming

the TM domain of MDR3 might lead to disruption of helical structure, providing a functional explanation for the overall pattern. Specific variants might, however, diverge from the trend due to averaging over possible substitutions and as such, a detailed view on every substitution is necessary.

4.3 Conclusion and significance

Focusing on a single protein for an ML predictor can be beneficial to predictor performance, providing increased accuracy for the protein of interest (Riera et al., 2016). While the trend to increasingly larger datasets to provide predictions for every known protein is undisputably valuable, the publication presented here highlights the additional benefit of further creating specific protein predictors. Key points that are addressed and provided within this publication:

i. Generation of an MDR3-specific dataset
   The largest dataset specifically for the MDR3 protein to date was derived using a combination of literature-based knowledge and filtered variants from the gnomAD database.

ii. Development of a highly reliable MDR3-specific predictor
   A unique combination of general protein predictors and additional features resulted in increased predictor performance, outperforming single included protein predictors and the external general predictor MutPred2.

iii. Providing access to prediction results and structural visualization
   A webserver was implemented to allow easy access and rapid assessment of variants. Due to the precomputation of all possible substitutions, waiting time for the user is minimized. Visualization of the variant site within the protein structure is provided to further engage users. Additionally, source code, precomputed substitution map, and standalone version of Vasor can be downloaded for more experienced users within the ML field.

The successful collaboration of experts from different fields was vital for this project to shape a well-rounded prediction tool. With the developed ML-based tool Vasor, I provide a specified predictor for single-site amino acid substitutions in MDR3. Based on the importance of MDR3, research on a range of diseases, including PFIC3, can benefit from such a highly reliable predictor.

Chapter 5 Publication II

Impaired transitioning of the FXR ligand binding domain to an active state underlies a PFIC5 phenotype

A. Behrendt, J. Stindt, E.-D. Pfister, K. Grau, S. Brands, C. Dröge, A. Stalke, M. Bonus, M. Sgodda, T. Cantz, A. Bastianelli, U. Baumann, V. Keitel, H. Gohlke.

Manuscript in peer-review. Contribution: 40%

Accessible as preprint at bioRxiv, DOI: 10.1101/2024.02.08.579530

Study conception and design: A. Behrendt, J. Stindt, M. Bonus, U. Baumann, V. Keitel, H. Gohlke; data collection: A. Behrendt, J. Stindt, E.-D. Pfister, K. Grau, S. Brands, C. Dröge, A. Stalke, M. Sgodda, T. Cantz, A. Bastianelli; analysis and interpretation of results: A. Behrendt, J. Stindt, C. Dröge, V. Keitel, H. Gohlke; draft manuscript preparation: A. Behrendt, J. Stindt, V. Keitel, H. Gohlke. All authors contributed to scientific discussions, reviewed the results and approved the final version of the manuscript.

(I adapted parts of the following text and figures from the respective manuscript.)

5.1 Background

Nuclear receptors (NRs) mediate a wide range of functions, orchestrating different downstream target gene expression based on the ligand, isoform, and tissue-specific effects (Kim et al., 2007; Massafra et al., 2018; Merk et al., 2019; Ramos Pittol et al., 2020). Subtle ligand changes have been found to change the ligands impact from agonistic to partial agonistic or antagonistic effects, indicating a highly sensitive and flexible ligand binding domain (LBD) (Merk et al., 2019). The activation function 2 (AF2) surface is of high importance for protein function as it mediates binding to coactivator or corepressor proteins, depending on the positioning of the helix 12 (H12) as a crucial part of the AF2 surface (Aranda & Pascual, 2001; Mi et al., 2003). Coactivators interact with the AF2 surface using a conserved LXXLL motif (Heery et al., 1997), while partial agonists and antagonists have been found to disturb the proper placement of H12, leading to favored corepressor binding with a larger hydrophobic

motif that additionally blocks the positioning of H12 required for an active state (Merk et al., 2019; Xu et al., 2002). A plethora of crystallization studies have revealed a highly similar LBD structure for NRs with a conserved H12 positioning for the active conformation (Chrisman et al., 2018; Kroker & Bruning, 2015; Wurtz et al., 1996; Xu et al., 2002; Zheng et al., 2018). Revealing the structure of H12 in the inactive state, however, has proven more difficult. Studies suggest that H12 is highly flexible in the apo state and does not form connections to the core of the LBD (Kallenberger et al., 2003; Renaud et al., 1995; Weikum et al., 2018), while others indicate that H12 can be bound to the LBD even within the apo state (Merk et al., 2019) (see Chapter 2.3.4). Crystal structure determination of antagonist-bound states failed to resolve H12, further indicating high flexibility within inactive states (Jiang et al., 2021; Jin et al., 2013). Overall, NR LBDs likely can access a range of different conformations, with one well-defined active state, and both ligand and coactivator or corepressor binding influence the likelihood of certain states. The NR farnesoid X receptor (FXR) is involved in glucose and lipid metabolism (Jiao et al., 2015; Ma et al., 2006; Sinal et al., 2000), immune response (Fiorucci et al., 2018, 2022) and bile production (Goodwin et al., 2000), based on a range of transcriptionally regulated genes as well as tissue-specific differences (reviewed in Han, 2018; Jiang et al., 2021; Massafra et al., 2018). The bile acid-responsive FXR protein is a key regulator in hepatocytes and maintains bile homeostasis by transcriptional control of the BSEP promotor (Ananthanarayanan et al., 2001; Ijssennagger et al., 2016) as well as the SHP promotor (Goodwin et al., 2000; Lu et al., 2000). Its widespread functions have made FXR a target for pharmaceutical interventions (Jiang et al., 2021; Massafra et al., 2018). To maximize desired targeting while avoiding side effects, detailed molecular mechanistic studies and an in-depth understanding of the dynamical movement of FXR are needed to enable future targeted approaches. Genetic variations within FXR may lead to a predisposition for ICP (van Mil et al., 2007) or inflammatory bowel diseases (Attinkara et al., 2012). Further, variants have been linked to PFIC subtype 5 (Gomez-Ospina et al., 2016; Mehta et al., 2022; Pfister et al., 2022). A novel homozygous missense variant has been identified in a patient and has been classified as PFIC5 (Pfister et al., 2022). Within this publication, we[‡] studied the effect of the variant on FXR activity using *in vitro* and *in silico* studies. Additionally, by assessing FXR-

[‡] Cellular assays and patient tissue were analyzed by J. Stindt, A. Bastianelli, C. Dröge and V. Keitel; *in silico* studies and *in vitro* ligand binding were performed by A. Behrendt and H. Gohlke.

regulated gene expression in patient tissue, we confirmed FXR dysfunction *in vivo*. Using unbiased MD simulations, I uncovered the conformational change from the inactive to the active state of the wildtype (WT) FXR LBD and deciphered the variants' effect on both inactive and active states, enabling a detailed mechanistic interpretation of the variant effect.

5.2 Results

The variant FXR T296I is located within the LBD

The identified variant, a mutation from a threonine at position 296 (reference sequence UniProt entry Q96RI1-1) to an isoleucine (in short T296I), lies within the helix 3 in the LBD (**Figure 17**). Based on its localization (**Figure 17**, A), we hypothesized an influence of the variant on forming the active state. Accordingly, I prepared four systems for MD simulations to study the variant influence compared to the WT protein: "active WT", "active T296I", "inactive WT" and "inactive T296I" (**Figure 17**, B). All systems further contained the most potent *in vivo* endogenous FXR agonist CDCA (H. Wang et al., 1999), as well as a short peptide of the nuclear receptor coactivator 2 (NCoA2) to drive the systems towards the active conformation. Protein activity measurements in cellular assays were based on the transcriptional activity of FXR in HEK293 cells using reporter-based luciferase assays.

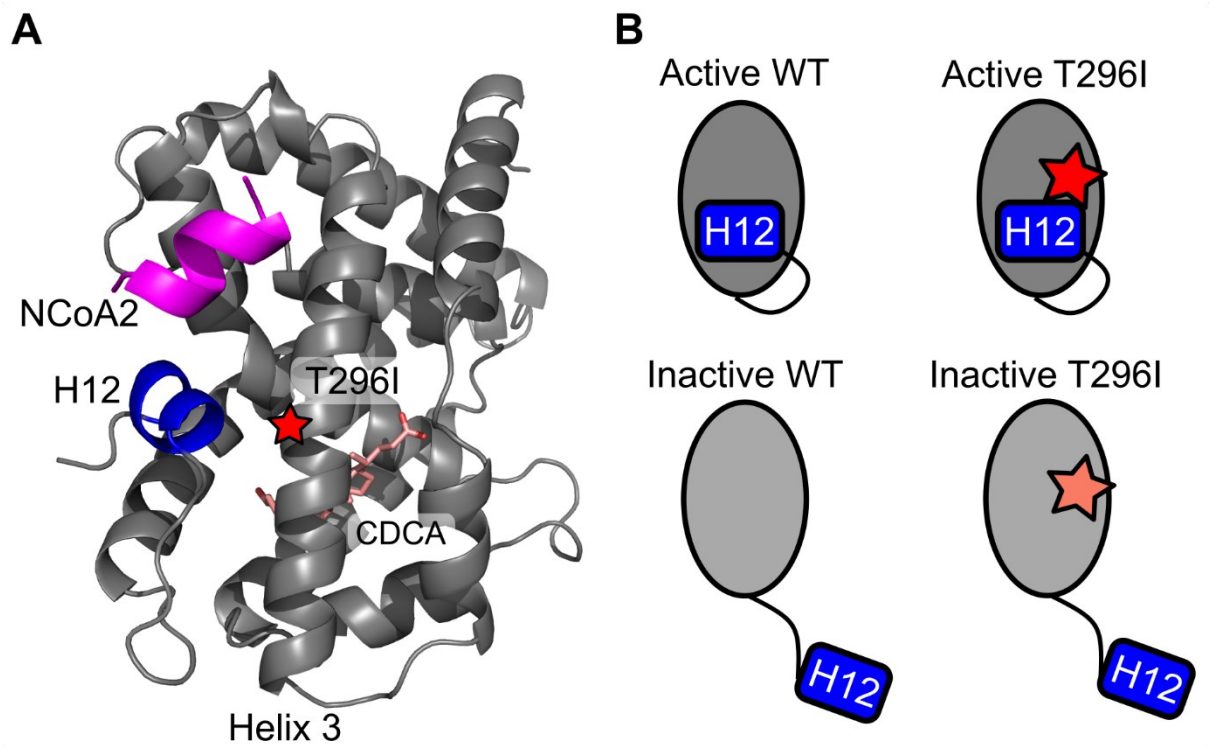


Figure 17: Schematic overview of the variant localization and MD simulation setup of the FXR LBD. [A] T296I variant localization (red star) within the active state of the FXR LBD, based on the crystal structure of agonist-bound FXR (PDB ID 6HL1) (Merk et al., 2019). Helix 12 (H12, shown in blue) is in close proximity to the variant site in the active conformation. The endogenous ligand Chenodeoxycholic acid (CDCA, shown as licorice in pink) is bound within the LBD core. A short peptide containing the LXXLL interaction motif, belonging to the nuclear receptor coactivator 2 (NCoA2, purple) binds to the surface formed by H12, helix 3, and helix 4. [B] Setup of the four systems for MD simulations to analyze the variant impact within the active and the inactive conformation.

FXR T296I decreases transcriptional activity in cellular assays

HEK293 cells were co-transfected with both liver-expressed FXR isoforms, FXR1α and FXRα2, as well as RXRα. Of note, FXRα2 is the main metabolic regulator in hepatocytes (Ramos Pittol et al., 2020; Vaquero et al., 2013). Cells were subjected to immunostaining and Western blotting to exclude any effect of the variant T296I on protein localization and overall expression levels. Both FXR WT and T296I showed the expected nuclear localization with similar protein levels (**Figure 18**, A and B). A luciferase-based assay was performed to study the protein activity of FXR WT and T296I. Cells were transfected with FXR and RXR constructs and a vector containing the luciferase gene under the control of either the BSEP- or SHP-promotor sequence. Both BSEP and SHP are well-established transcriptionally regulated FXR targets (Ananthanarayanan et al., 2001; Goodwin et al., 2000; Lu et al., 2000; Plass et al., 2002). Cells were stimulated with the FXR agonist obeticholic acid (OCA) (Pellicciari et al.,

2002) and RXR-agonist 9-cis-retinoic acid (Heyman et al., 1992) to provide optimal conditions for protein activity. Values were normalized to FXRα1 WT and RXRα or FXRα2 WT and RXRα signals (**Figure 18**, C and D), as these conditions are expected to lead to the highest protein activity. FXRα1 WT or FXRα2 WT transfection alone resulted in a significant decrease of protein activity since the functional readout is based on binding to the BSEP and SHP promotor, containing the IR-1 canonical motif for FXR/RXR heterodimers (Forman et al., 1995). FXR T296I transfection consistently resulted in a significant decrease in transcriptional activity compared to the WT in both isoforms and on BSEP and SHP promotor targets. Specifically, co-transfection of FXRα1/2 T296I with RXRα showed significantly reduced luciferase activity in BSEP- (**Figure 18**, C) and SHP-promotor regulated luciferase readouts (**Figure 18**, D). Overall, the data indicated decreased functional activity of the FXR T296I protein while subcellular localization and protein expression were unaffected.

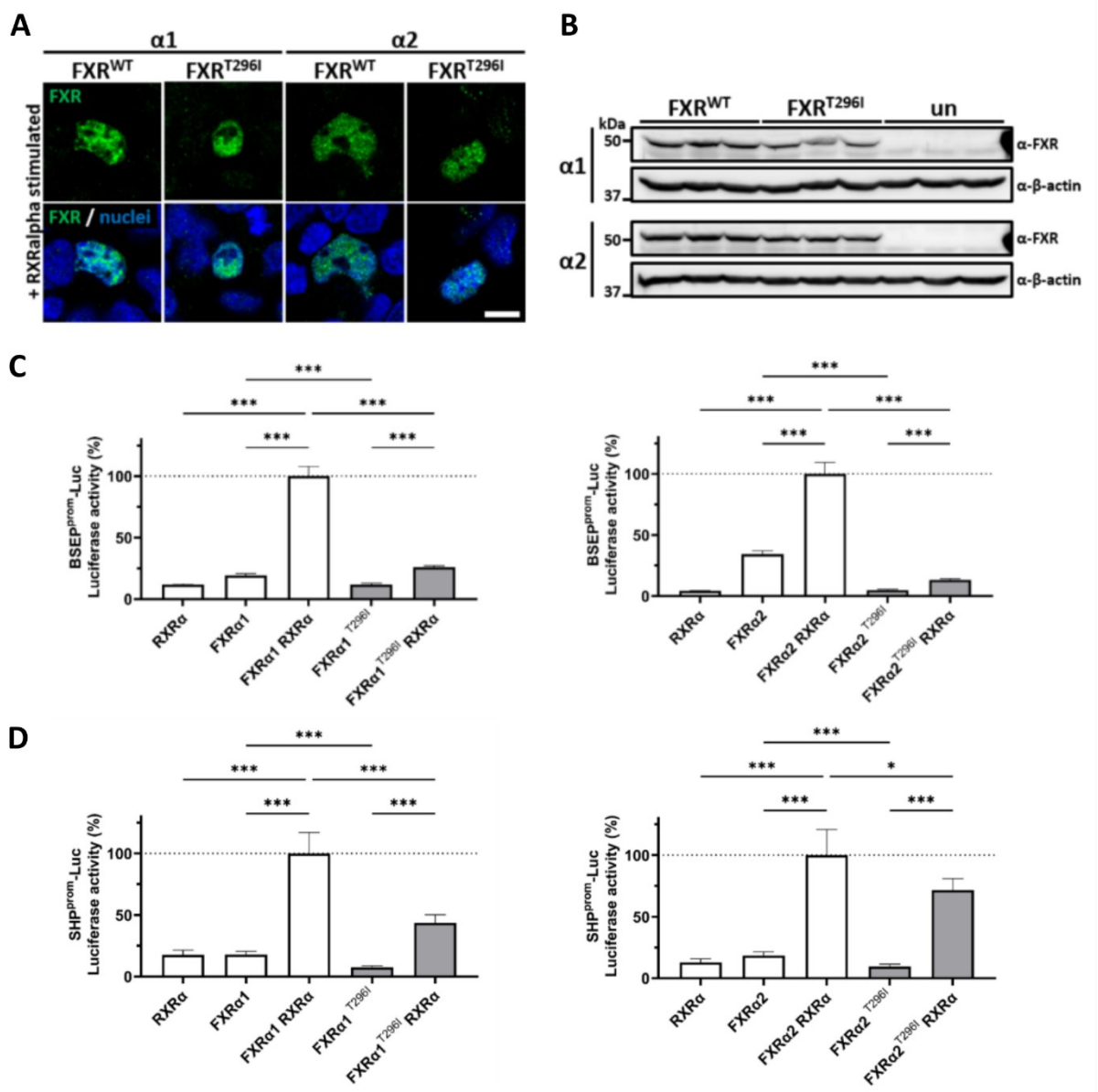
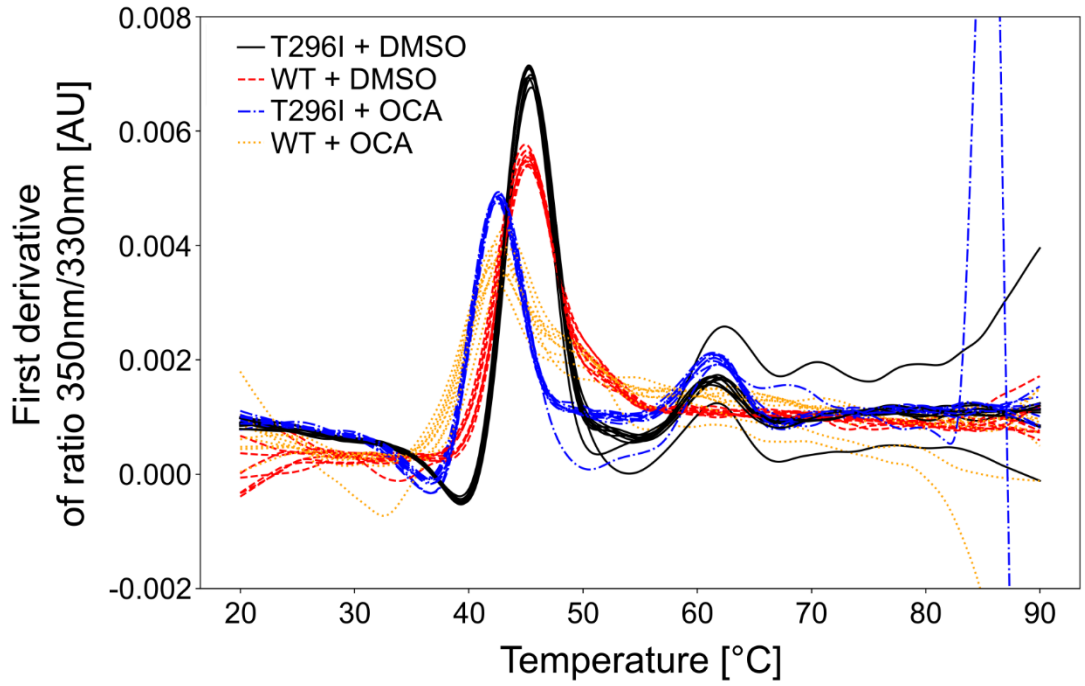


Figure 18: FXR WT and T296I localization, protein levels, and transcriptional activity in HEK293 cell assays. [A] HEK293 cells were transiently transfected with either FXRα1-WT, FXRα1-T296I, FXRα2-WT, or FXRα2-T296I in combination with RXRα. Staining was performed with an anti-FXR antibody (H-130, Santa Cruz Biotechnology, shown in green) and with the nuclear counterstain 4',6-diamidino-2-phenylindole (DAPI, shown in blue), revealing nuclear localization for both WT and T296I protein. [B] Western blot of transfected HEK293 cells indicated similar overall protein levels of variant and WT protein. [C] Transcriptional activity of FXR constructs and in combination with RXRα (or RXRα only as control) measured using a Luciferase assay readout, with the luciferase gene under BSEP-promoter control. [D] Transcriptional activity of FXR constructs and in combination with RXRα (or RXRα only as control) measured using a Luciferase assay readout, with the luciferase gene under SHP-promoter control. Significance testing was performed using a two-tailed Student's t-test.

Furthermore, to exclude the possibility that the variant T296I impacts ligand binding, which in turn could affect protein activity, we analyzed recombinant FXR WT and T296I protein in the presence and absence of the ligand. Of note, within the simulated time of MDs, no unbinding

events of the ligands were observed, neither in the WT nor in the variant protein, indicating stable ligand binding once positioned in its pocket independently of helix 12 placement. To study the ligand binding and its impact on protein stability, FXR WT and FXR T296I recombinant proteins (with a 6xHis- and small ubiquitin-related modifier (SUMO)-tag for easier purification and increased solubility (Butt et al., 2005; Malakhov et al., 2004)) were expressed within *E. coli* Rosetta cells and separated from other bacterial proteins using a two-step procedure. First, the His-tagged FXR protein was subjected to a HisTrap column and, in a second step, further purified using a size exclusion chromatography column. Purified and concentrated protein was aliquoted and stored at -80°C until further usage in melting temperature experiments. NanoDSF, a differential scanning fluorescence method, was employed in which a protein solution is gradually heated while measuring the autofluorescence of intrinsic tryptophan residues within the protein as a measure of structural unfolding (J. Wen et al., 2020). Both FXR WT and FXR T296I exhibited a similar melting temperature in the absence of OCA, indicating that the variant does not impact the overall structure of the protein fold (**Figure 19**, A and B). In the presence of the ligand, both WT and T296I showed a significant shift towards decreased melting temperature while showing no significant difference between each other, indicating ligand binding to both wildtype and variant protein (**Figure 19**, B).

A



B

		Mean [°C]	Standard deviation [°C]
n.s.	FXR WT + DMSO	45.17	0.09
*	FXR T296I + DMSO	45.36	0.12
***	FXR WT + OCA	42.84	0.41
n.s.	FXR T296I + OCA	42.58	0.12

Figure 19: Melting temperature of FXR WT and T296I protein. [A] NanoDSF-measured melting temperature of FXR WT or FXR T296I protein (25µM) with either DMSO (2.5%) only or OCA (250µM) dissolved in DMSO (2.5%) present. [B] Derived mean values and standard deviations over three experiments with three replicates each. Significance testing was performed using Welch's t-test.

Supporting our data that ligand binding is likely undisturbed by the variant, steered MD studies on the FXR LBD with the agonist GW4064 have indicated an egress pathway facing helix 1-helix 2 loop and helix 5-helix 6 loop as energetically most favorable (W. Li et al., 2012), thus facing away from the H12 and the variant site. Furthermore, computational studies on ligand binding and unbinding in related NRs such as retinoic-acid related-orphan-receptor-C gamma (RORγ) identified the so called "backdoor" pathway, facing away from the AF2 surface (Saen-Oon et al., 2019). Overall, the variant T296I, facing towards H12 and in close proximity to the AF2 surface, did not disturb the binding of OCA or general protein properties.

The variant FXR T296I lowers the probability of H12 placement correlated to the active state

Next, I employed MD simulations to investigate the molecular mechanism underlying the *in vitro* identified decreased functional activity of FXR T296I. Using the four different systems "active WT", "active T296I", "inactive WT" and "inactive T296I" (**Figure 17**, B), 15 replica per system with a simulation time of 1 µs per replica were prepared and analyzed. Within the crystal structure of agonist-bound FXR LBD (Merk et al., 2019), residue 296 likely interacts with a threonine directly preceding H12, T466 (**Figure 20**, A). The derived distance between residue 296 and residue 466 was used as a reference value indicating a likely active conformation and compared to measured distances over the simulation time. Comparing the active WT with the active T296I system indicated increased distances. Active WT systems showed a distance distribution with a large peak around the reference distance cutoff, indicating an active conformation, and a smaller peak with slightly higher distances (**Figure 20**, B, first panel). Active T296I, however, revealed a shift of the distance distribution to one broadened peak towards higher distances (**Figure 20**, B, second panel). Analyzing the frequency of reaching the reference cutoff (converted into percentages as measured over the simulation time for each replica) revealed a significant decrease of the active T296I system (mean value of 0.40%) in reaching the reference cutoff compared to the active WT, which showed close contact to T466 below the reference value for about one-fourth of the entire simulation time (mean value of 26.95%) (**Figure 20**, C). Accordingly, even in the active WT, the active conformation is not always perfectly preserved, which is attributable to the dynamic movement of proteins. Due to the high degree of flexibility for H12 in the inactive systems, measured distances show a broad fluctuation (**Figure 20**, B, third and fourth panel). Interestingly, the inactive WT system reached distances below the reference value in several replicas (mean value of 1.79%), resulting in a small peak around the reference distance (**Figure 20**, B and C), indicating that inactive WT might transition into an active conformation. However, this was not observed for the inactive T296I system (**Figure 20**, B and C), where distances below the reference cutoff were only reached briefly in one replica (mean value of 0.03%).

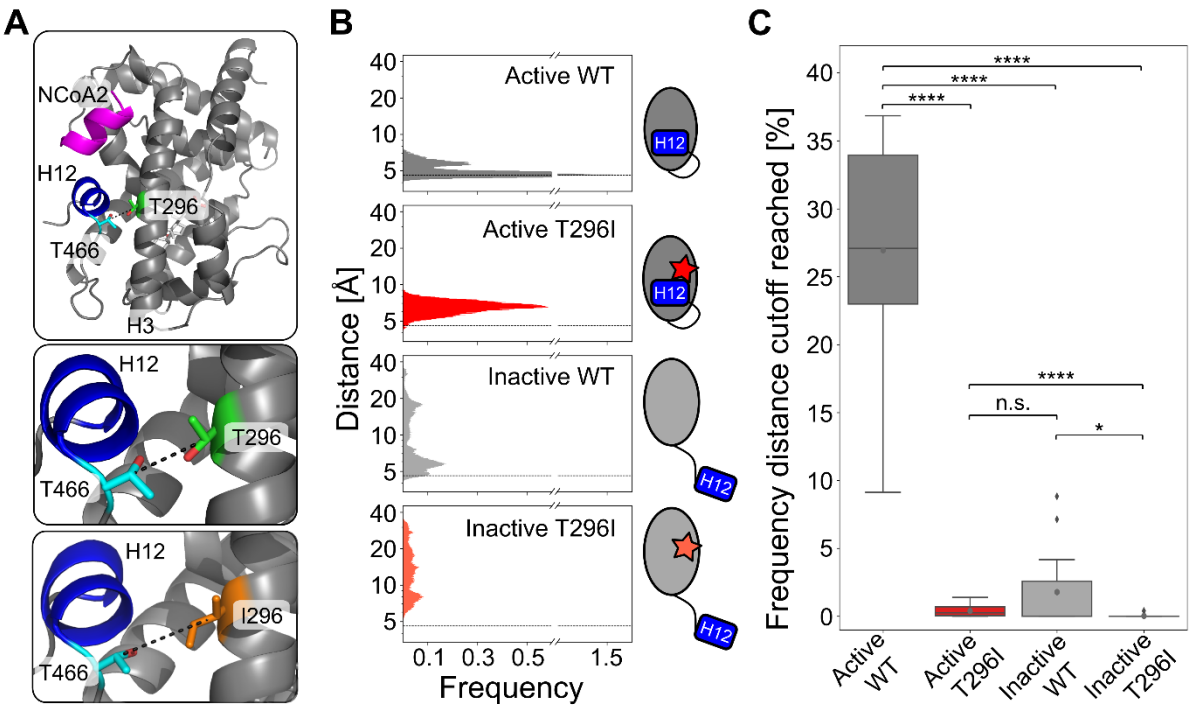


Figure 20: T296I variant leads to increased distance to residue T466. [A] Overview of measured distance within the FXR LBD with marked residue T296 and T466 (upper panel). The distance was measured between the C_β atoms of T296 (middle panel, with WT residue T296 shown in green) or I296 (lower panel, with variant residue I296 shown in orange) and T466. The mean distance over the simulation time is increased in the active T296I system (6.6 Å) compared to the active WT system (5.0 Å) and the reference distance as measured in the agonist-bound crystal structure (4.6 Å). [B] Histogram of measured distance distribution for each system setup. The reference distance cutoff is indicated as a dashed grey line. [C] Frequency of each system reaching the reference distance, calculated per replica and pooled per system. Boxes depict the quartiles of the data with the median (straight black lines) and mean (grey dots) indicated; the whiskers indicate the minimum and the maximal values, outlier points are depicted as rhombus. Differences in the mean values were statistically evaluated using a two-sided Mann-Whitney U test (N = 15, n.s.: not significant; *: $p \leq 0.05$, **: $p \leq 0.01$, ***: $p \leq 0.001$, ****: $p \leq 0.0001$).

Visualization of the distance measurement over the simulated time for each replica further provided an overview of which replica might transition from an inactive to active conformation (**Figure 21**). While 6 out of 15 replicas for the inactive WT reached the reference distance value, only 1 replica of the inactive T296I system transiently reached below the cutoff (replica 6). In summary, the data revealed an increased distance between the variant site and T466 as an interacting residue next to H12, indicating that the active state is destabilized in the variant protein.

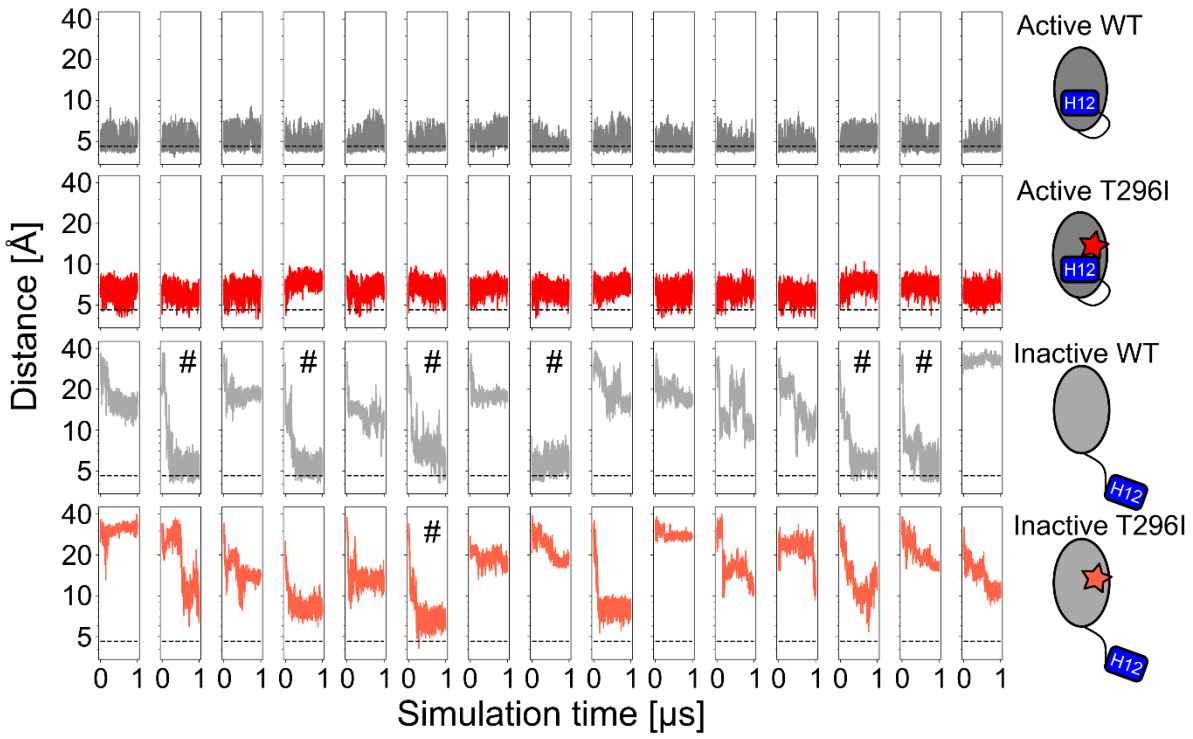


Figure 21: Distance measurement between T466 and residue 296 within MD replicas. The distance measured between C_β atoms of T466 and residue 296 over the simulated time for each replica and system. Histograms and calculated frequencies of **Figure 20** were calculated based on the data. The reference distance value is marked as a dashed grey line. Inactive system replicas that reach the reference value are marked (#).

The variant FXR T296I leads to decreased conformational change into the active conformation

Several MD studies using the FXR LBD have confirmed the importance of the H12 positioning (Kumari et al., 2021) and investigated changes associated with novel drug candidates (Díaz-Holguín et al., 2023; Kumari et al., 2023). However, the transitioning from the inactive to the active conformation has so far not been shown in MD studies. Based on the indication from the previous distance analysis that the inactive WT system may transition into an active conformation, I visually inspected MD trajectories with a special focus on H12 placement in line with the active conformation (**Figure 22**). Of note, several replicas of the inactive WT showed similar transitioning and accordingly, one replica (replica 2) was chosen at random (**Figure 22**, A). Replica 6 of the inactive T296I was chosen for visualization as it showed conformational transitioning closest to the active state although not fully reaching perfect H12 placement (**Figure 21** and **Figure 22**, B).

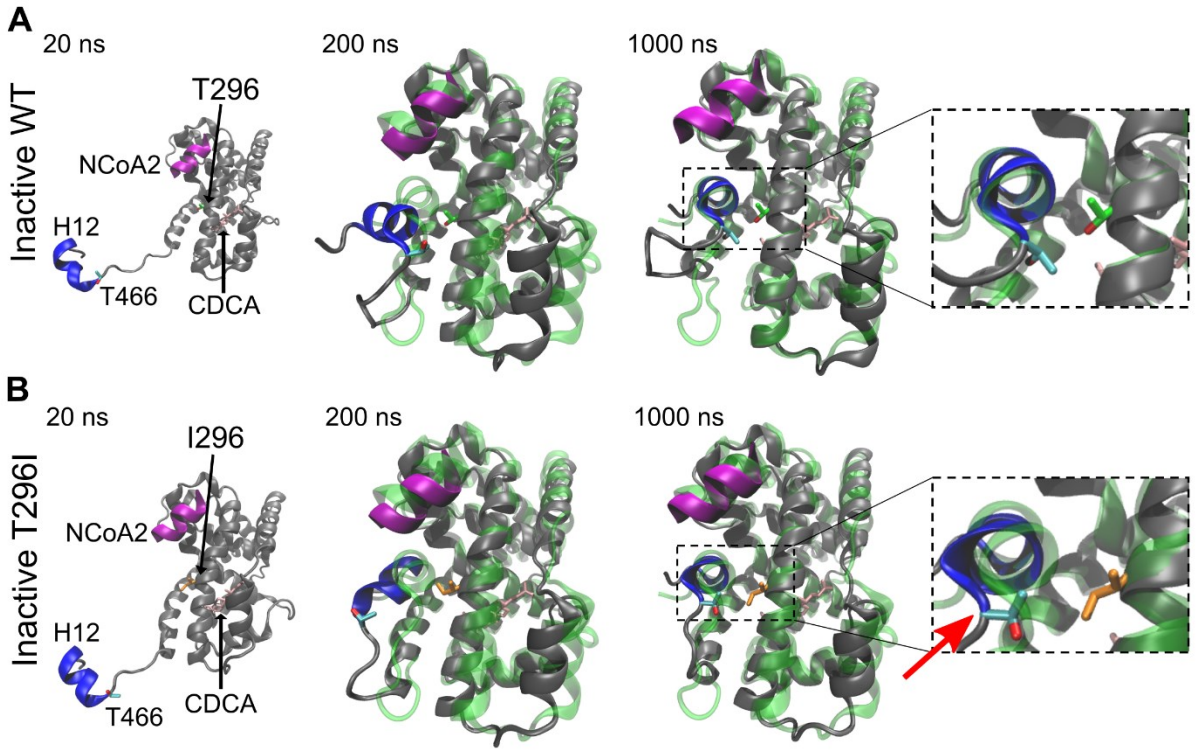


Figure 22: Conformational change of H12 over exemplary MD trajectories of inactive WT or T296I systems. [A] Inactive WT (replica 2) transitioned from the initial inactive state into a conformation with H12 closely aligning to the active reference state (based on the crystal structure of agonist-bound FXR LBD (Merk et al., 2019), green translucent structure). [B] Inactive T296I (replica 6) showed a conformational change into a close to the active state structure but with an imperfect H12 placement (marked with a red arrow). The side chain of residue T466 (light blue licorice), T296 (green licorice) or I296 (orange licorice), as well as H12 (blue cartoon), NCoA2 peptide (purple cartoon), and CDCA ligand (pink licorice) are highlighted (oxygen atoms within side chains are consistently colored red).

To further investigate and quantify the observed conformational change, I employed an RMSD-based measurement to analyze atomic coordinate distances between H12 residues over the MD simulation time compared to the initial reference crystal structure of agonist-bound FXR. In detail, I first fixed the conformations of the trajectory to the most stable core, calculated over all four MD systems, to avoid arbitrary distortion of the RMSD values by, e.g., rotational movement. Next, I calculated the all-atom RMSD of the H12 residues and the preceding T466 against the active reference structure and visualized the derived distribution for the active WT and active T296I systems (**Figure 23**, A). In line with the results of the distance analysis, RMSD distribution is significantly shifted to higher RMSD values based on fitted skewed Gaussian functions on the active T296I histogram compared to the active WT histogram (**Figure 23**, A). Further, the RMSD distribution of the active WT system was used to derive a reference RMSD value, indicating the mean RMSD value for H12 fluctuations that can

be expected in an uninhibited active state. The calculated value of 1.9 Å was used as a reference in the histogram distribution of all four MD states (**Figure 23**, B). Calculation of average time spent reaching the reference value over the simulation time per replica (**Figure 23**, C) revealed a significant decrease in the frequency of occupying the active state in all three systems compared to the active WT state. RMSD value distribution of the inactive WT system (**Figure 23**, B, third panel) revealed a peak close to the reference value, indicating again that transitioning into an active conformation can occur, confirming the indication from the basic distance analysis (**Figure 20** and **Figure 21**) and in line with visual analysis (**Figure 22**).

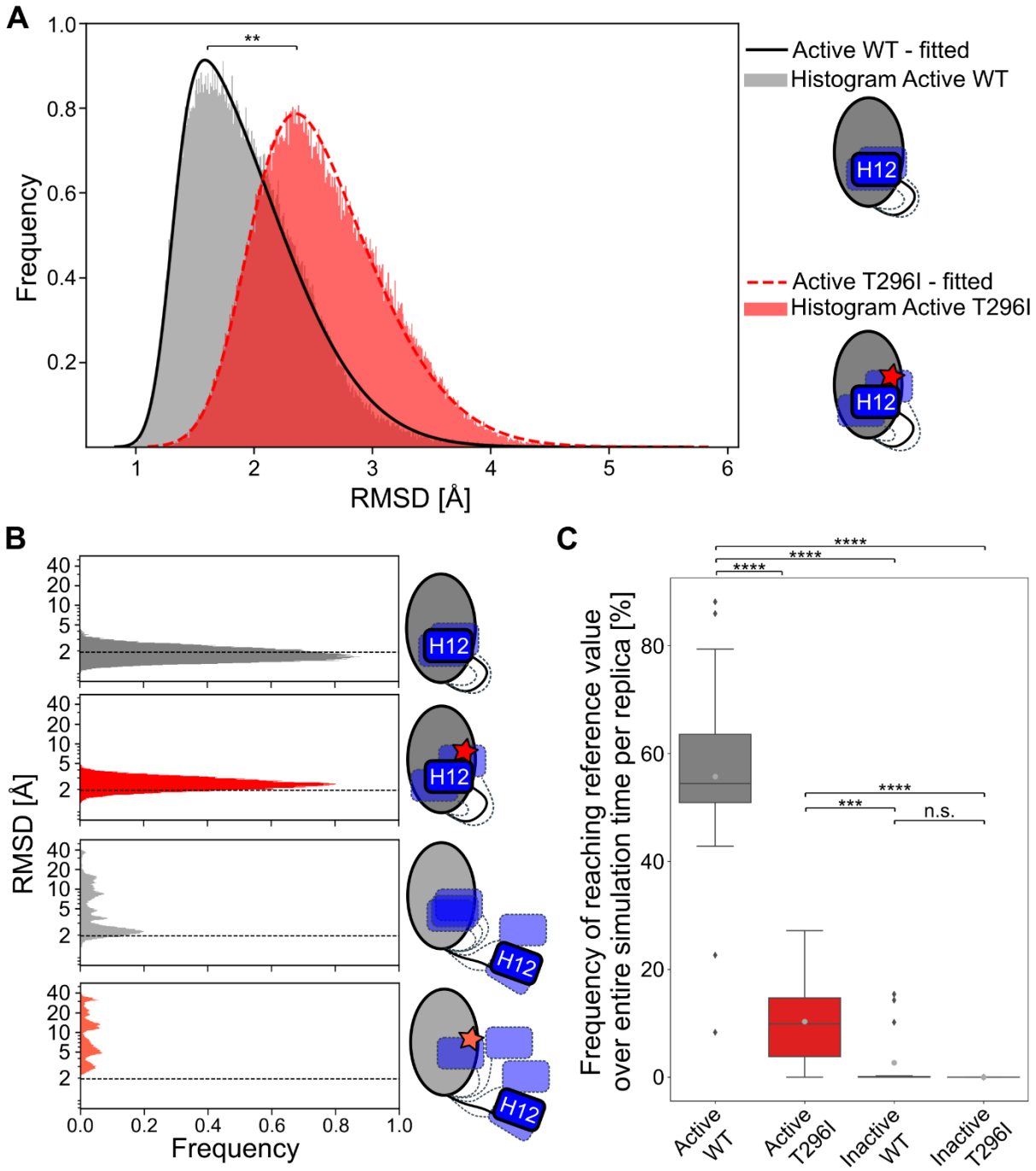


Figure 23: Movement of H12 in MD systems based on RMSD measurement. [A] RMSD value distribution of H12 and preceding T466 over all replicas, compared to the initial crystal structure as the active reference state. Skewed Gaussian functions were fitted to the distributions of active WT and active T296I systems, revealing a significant shift towards higher RMSD values in the active T296I system (two-sided Students t-test). The derived mean of the active WT system (1.9 Å) was further used as a reference value for expected RMSD fluctuations for H12. [B] RMSD value distribution for all four MD systems, with the reference value derived from [A] marked as a dashed grey line. [C] Frequency of each system reaching the reference value, calculated per replica and pooled per system. Boxes depict the quartiles of the data with the median (straight black lines) and mean (grey dots) indicated; the whiskers indicate the minimum and the maximal values, outlier points are depicted as rhombus. Differences in the mean values were statistically evaluated using a two-sided Mann-Whitney U test (N = 15, n.s.: not significant; *: $p \leq 0.05$, **: $p \leq 0.01$, ***: $p \leq 0.001$, ****: $p \leq 0.0001$).

Accordingly, inactive WT reached below the reference value in several replica, translating into frequencies of occupying states below the reference RMSD value as high as ~15% (**Figure 23**, C). The data indicates that once the inactive WT system transitioned, the system stably stays within the active state, in line with the indications from the distance analysis. However, since only 4 out of 15 replicas reach below the reference value for the inactive WT while the inactive T296I system reached in one replica with a calculated frequency of 0.01%, differences between the inactive systems are not significant.

Overall, within the active systems of MD simulations, the variant T296I showed structural deviation from the stable active conformation, indicating a destabilization of the active state. Further, analyzing the transitioning from inactive to active conformation, the variant likely impedes effective conformational change, decreasing the frequency of FXR within the active state and accordingly its protein activity. This observation correlates with the decreased protein transcriptional activity identified *in vitro*. Of note, the remaining activity indicated by both *in vitro* and *in silico* data might explain the clinical manifestation. Despite high disease severity with the necessity for organ liver transplantation at the age of 8 months due to terminal liver disease (Pfister et al., 2022), this homozygous variant is not *per se* incompatible with life. The reduced protein function was further verified within a patient's tissue sample by analyzing the expression of downstream targets BSEP and SHP, revealing a significant decrease in protein expression and, thus, decreased FXR T296I transcriptional activity. My work within this project substantially contributed to understanding the functional impact of the variant on a molecular level.

5.3 Conclusion and significance

Within this highly interdisciplinary project, we combined patient sample data, cellular assays, and *in silico* analysis to unravel the molecular mechanism and functional impact of a missense variant in the NR FXR. Key points within this project include:

i. Functional impairment of FXR variant protein in *in vitro* and *in vivo* assays
The variant T296I reduced the transcriptional activity significantly, while FXR protein levels, localization, and ligand binding were not affected. Functional impairment was further validated *in vivo* in patient tissue based on reduced expression of transcriptionally regulated target genes BSEP and SHP.

ii. Decreased transitioning of FXR variant protein into the active state in unbiased MD
The variant T296I critically impacted the positioning of H12, showing impairments when comparing the active systems. Further, and potentially more impactful, T296I reduced the frequency of transitioning from the inactive to the active conformation. Together, the data explains the functional impairment of FXR T296I on a molecular level.

iii. Uncovering transitioning of FXR WT from inactive to active state in unbiased MD
For the FXR WT, protein functionality is dependent on conformational changes from inactive to active states. To our knowledge, this is the first study to reveal the pathway of this transitioning for the FXR LBD using unbiased MD simulations.

Beyond understanding the effect of a missense variant in detail, the work may provide a basis for future revelations. Extrahepatic FXR expression is widespread with diverse tissue-specific functions and accordingly, dysregulation and disease involvement of FXR in cholestatic diseases, non-alcoholic fatty liver disease (NAFLD), inflammation, and various cancers have made FXR a pharmacological target (reviewed in Han, 2018). Safely targeting and modulating FXR function requires a detailed understanding of protein dynamics, wherein the inclusion of inactive to active transitioning may provide valuable information for future rational drug design.

Chapter 6 Summary and Perspective

During the work performed for this thesis, I have achieved and successfully used skills from the computational fields of machine learning (ML) and molecular dynamics (MD) simulations (see **Figure 24**). In short, in collaboration with Pegah Golchin and Filip König (both Heinrich Heine University Düsseldorf, Germany), I built an MDR3-specific dataset of variants that are either disease-associated or benign to train a ML algorithm for classifying single-site mutations into benign or pathogenic (see Chapter 4, Publication I). The generated tool, called Vasor (Variant assessor of MDR3), enables users to rapidly assess the impact of a novel variant and thus prioritize variants for further experimental evaluation. In order to facilitate access to users, especially novice ones in the field of bioinformatics, Vasor was made available as a webserver (https://cpclab.uni-duesseldorf.de/mdr3_predictor/). To further engage users, a structural overview of the MDR3 protein was additionally provided with an automatic highlighting of the entered variant as well as automated image generation of wildtype and variant protein. The python-coded Vasor program can also be downloaded and locally installed. The program has been tested against current state-of-the-art mutation predictor, MutPred2, and outperformed it as well as other predictors, which were included as features for the machine learning approach.

The established approach for a protein-specific predictor has proven beneficial and accordingly will be further used for the protein BSEP (see Chapter 2.3.2), which is a bile salt transporter located at the canalicular membrane. Similar to the MDR3 protein, there is no protein-specific prediction tool available yet despite BSEP's disease involvement. Using information from extensive studies on missense variants (see e.g., Dröge et al., 2017; Sohail et al., 2021) may provide a good dataset for machine learning to enable classification of novel variants. We envision this project in the continuation of the HiChol consortium, which achieved continued funding from the BMBF based on its success. Further, the establishment of protein-specific prediction tools contributes to the active field of applying machine learning for research problems. While a high number of available tools may seem daunting at first glance, they offer the possibility to identify of best-suited tools for specific problems. Accordingly, besides well-performing general protein predictors, protein-specific tools fill important niches and provide a great asset to researchers and clinicians.

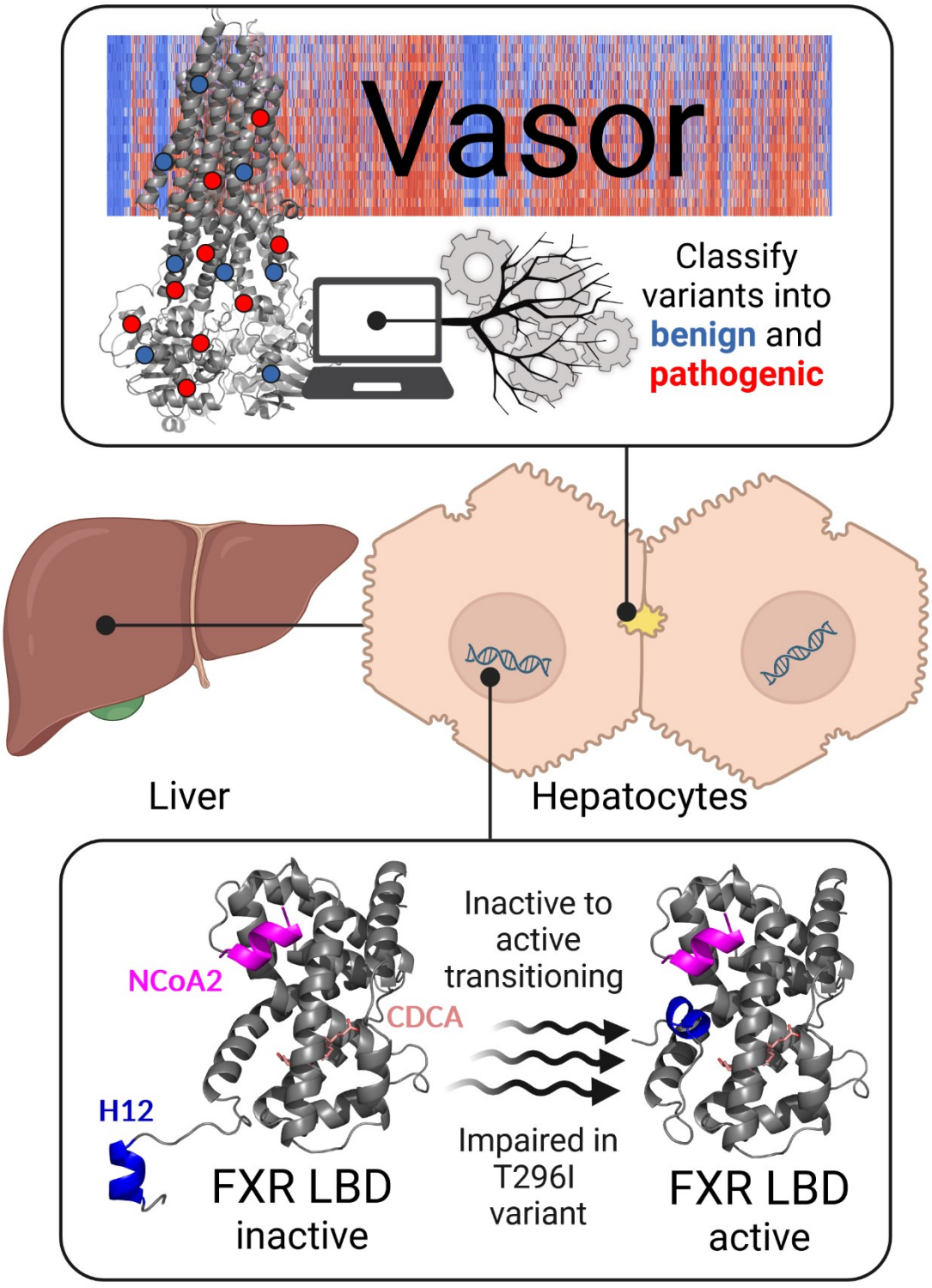


Figure 24: Overview of the presented work. The ML-based predictor Vasor classifies variants of the MDR3 protein, located within the canalicular membrane in hepatocytes (upper panel). Transitioning from the inactive to active positioning was uncovered using MD simulations for the FXR LBD (lower panel). A PFIC5-associated variant impaired this transitioning, in line with *in vitro* assays. Created with BioRender.

For the FXR protein, MD simulations were employed to analyze the effect of the variant T296I within the protein (see Chapter 5, Publication II). Combining the work with *in vitro* and *in vivo* studies performed by Dr. Jan Stindt (Heinrich Heine University Düsseldorf, Germany), Dr. Malte Sgodda, Prof. Dr. Tobias Cantz (Medizinische Hochschule Hannover, Germany), Dr. Alex Bastianelli, Dr. Carola Dröge and Prof. Dr. Verena Keitel-Anselmino (Otto von Guericke University Magdeburg, Germany), the work provides an in-depth analysis of mutational impact on the protein function. Of note, I provided a detailed mechanistic understanding of variant impact within the activation dynamics of the FXR protein, and I reveal the transitioning from inactive to active state for FXR, a conformational change not yet described for the FXR LBD in MD simulations.

In a novel project within the continued HiChol consortium, we aim to investigate residue specific importance in the LBD of FXR using an Alanine Mutation Scanning approach. A combination of *in silico* and *in vitro* data will be used to create an extensive dataset for a machine learning approach to predict variant impact. Due to the structural similarity of NR LBDs and the high research interest in the area, the provided data (both from finished and novel projects) can provide valuable information for research on other NRs. Additionally, the establishment of the inactive and the active system including its transitioning may be used to study and design novel FXR ligands, without the previous limitation of analyzing effects only on the active state. Given the complexity of the FXR network (see Chapter 2.3.2 and Chapter 2.3.3), increased knowledge of residue importance and including explicitly both conformational states may provide a next step in understanding and regulating FXR functions.

Furthermore, I provided expertise on six clinically identified variants of interest within the ATP7B protein and supplied structure-based estimations of variant influence (Stalke et al., 2023). Similarly, I enriched the assessment of a clinically relevant MDR3 variant using the previously described derived MDR3-prediction tool as well as a structural assessment of the variant (Dröge et al., 2023). Overall, the collaborative effort and the unique combination of different fields of expertise to understand variant impact within important liver proteins has proven to be prosperous. It led to a substantial increase of knowledge within the field and hopefully will contribute to further research efforts in tackling and mediating variant impact to improve patient care and outcomes in the future.

Chapter 7 Acknowledgment

First, I wish to thank my supervisor Prof. Dr. Holger Gohlke. Entrusting me with highly interesting projects, that challenged me in getting familiar with new technologies and techniques, has helped me grow as a person and scientist. Furthermore, I am grateful for his encouragement and optimism to pursue my own ideas within research projects and for the trust and freedom that this path entailed.

I also wish to thank my second supervisor Prof. Dr. med. Verena Keitel-Anselmino. Her drive and expertise have always motivated me. Additionally, her perspective as a highly accomplished clinician has provided me with valuable insights and further formed my view on scientific problems as multifaceted issues. Bringing such different research areas together can be a challenge, and I admired her project steering and truly experienced the projects as positive examples of interdisciplinary research.

A big, heartfelt thank you to all my external collaboration partners, especially Dr. Carola Dröge, Dr. Jan Stindt, Dr. Amelie Stalke, and Dr. Alex Bastianelli, for their efforts, feedback, and expertise. It was a pleasure to work with you.

Next, I'd like to thank the entire CPC and CBC lab for the support and positive atmosphere. My special thanks goes out to Dr. Michele Bonus, who has patiently answered my questions, provided help and expertise, and discussing potential ideas with him has been a pleasure. My additional thanks to Dr. Michele Bonus and Yu Lin Ho for proofreading and valuable feedback. My deepest gratitude goes out to Pegah Golchin, who made starting out in a new area a lot more accessible and fun. It was a pleasure to work with her and learn from her. Thanks to Filip König for the pleasantly easy collaboration, valuable input, and extra drive to make Vasor even better. Thank you to Dr. Daniel Mulnaes for his expert view on machine learning and valuable insights. A big thank you to all my lab colleagues, especially Dilara Nemli, Laura Munoz Gloder, Yu Lin Ho, Kathrin Grau, Dr. Stefanie Brands and Mauricio Munoz, for providing expertise and diversion whenever needed and for rekindling my joy in wet lab experiments. A heartfelt thanks also to the student support from Adeline Schiwe, Jule Meister, Julia Thomas, and Katja Schötteler. Their motivation, especially within the lab, was inspiring.

A special thank you goes to Dr. Alexandra Hamacher, who was both an inspiration and motivator in tough situations. My gratitude and thanks to my colleagues of AMA and the entire research team of Prof. Dr. Kassack. My special gratitude goes of course to Laura Pradel.

My deepest thanks to all my friends, who might not always be aware of it but who are truly my backbone. I will keep it very brief and just say a huge thank you to everyone! A special thanks to my family for their support and their love.

Thank you to the person without none of this would have been possible. You gave me more than I thought possible despite tough times, and I will be forever grateful. Also, thank you to the other person without this would not have happened. To growth, on a personal, emotional, and scientific level!

Scientific career

12.2019 – present Ph. D. student in the research group of Prof. Dr. Holger Gohlke, Institute for Pharmaceutical and Medicinal Chemistry, Heinrich Heine University
Düsseldorf, Germany

Key tasks:
- Analyzing missense variants in proteins involved in cholestatic disorders using computational tools.
- Course assistant "Drug Analysis" (Pharmacy studies, 7th semester)
- Project management and research design, communication of scientific findings to scientific peers.

10.2018 – 10.2019 Research Associate in the group Epigenetics and COPD, BioMed X
Heidelberg, Germany

Key tasks:
- Establishing novel protocol for lung tissue immunofluorescence imaging.
- Processing of human lung tissue for visualization of protein expression, cryopreservation and cell type-specific isolation.
- Project management and research design, communication of scientific findings to industrial mentors and broad range of international scientific peers with different scientific backgrounds.

04.2018 – 09.2018 Research Associate in the group Tau-mediated neurodegeneration in Alzheimer's disease, BioMed X
Heidelberg, Germany

Key tasks:
- Elucidating caspase-cleaved tau fragment neurotoxicity *in vivo* by AAV-mediated transductions with mice hippocampal injections.
- Studying localization, uptake, cleavage and secretion of asparagine endopeptidase-cleaved tau *in vitro*.
- Project management and research design, communication of scientific findings to industrial mentors and broad range of international scientific peers with different scientific backgrounds.

10.2017 – 03.2018 Student intern in the group Tau-mediated neurodegeneration in Alzheimer's disease, BioMed X
Heidelberg, Germany

10.2013 – 10.2017 Master of Science in Molecular Biotechnology
Ruprecht-Karls-University
Heidelberg, Germany

Master thesis: Aggregation and seeding properties of tau fragments.

10.2010 – 09.2013 Bachelor of Science in Molecular Biotechnology
Ruprecht-Karls-University
Heidelberg, Germany

Scientific achievements

Publications

Amelie Stalke, **Annika Behrendt**, Finja Hennig, Holger Gohlke, Nicole Buhl, Thea Reinkens, Ulrich Baumann, Brigitte Schlegelberger, Thomas Illig, Eva-Doreen Pfister, Britta Skawran. Functional characterization of novel or yet uncharacterized ATP7B missense variants detected in patients with clinical Wilson's disease. *Clinical Genetics*, (2023). https://doi.org/10.1111/cge.14352

Lydia Reinhardt, Fabrizio Musacchio, Maria Bichmann, **Annika Behrendt**, Ebru Ercan-Herbst, Juliane Stein, Isabelle Becher, Per Haberkant, Julia Mader, David C. Schöndorf, Melanie Schmitt, Jürgen Korffmann, Peter Reinhardt, Christian Pohl, Mikhail Savitski, Corinna Klein, Laura Gasparini, Martin Fuhrmann, Dagmar E. Ehrnhoefer. Dual truncation of tau by caspase-2 accelerates its CHIP-mediated degradation. *Neurobiology of Disease*, Volume 182, (2023). https://doi.org/10.1016/j.nbd.2023.106126.

Carola Dröge, Tobias Götze, **Annika Behrendt**, Holger Gohlke, Verena Keitel. Diagnostic workup of suspected hereditary cholestasis in adults: a case report. *Exploration of Digestive Diseases*, (2023). https://doi.org/10.37349/edd.2023.00016

Annika Behrendt, Pegah Golchin, Filip König, Daniel Mulnaes, Amelie Stalke, Carola Dröge, Verena Keitel, Holger Gohlke. Vasor: Accurate prediction of variant effects for amino acid substitutions in multidrug resistance protein 3. *Hepatology Communications*, 6(11), (2022). https://doi.org/10.1002/hep4.2088

Maria Llamazares-Prada, Elisa Espinet, Vedrana Mijošek, Uwe Schwartz, Pavlo Lutsik, Raluca Tamas, Mandy Richter, **Annika Behrendt**, Stephanie T. Pohl, Naja P. Benz, Thomas Muley, Arne Warth, Claus Peter Heußel, Hauke Winter, Jonathan J. M. Landry, Felix J.F. Herth, Tinne C.J. Mertens, Harry Karmouty-Quintana, Ina Koch, Vladimir Benes, Jan O. Korbel, Sebastian M. Waszak, Andreas Trumpp, David M. Wyatt, Heiko F. Stahl, Christoph Plass, and Renata Z. Jurkowska. Versatile workflow for cell type–resolved transcriptional and epigenetic profiles from cryopreserved human lung. *JCI Insight*, Volume 6, (2021). https://doi.org/10.1172/jci.insight.140443

Ebru Ercan-Herbst, Jens Ehrig, David C. Schöndorf, **Annika Behrendt**, Bernd Klaus, Borja Gomez Ramos, Nuria Prat Oriol, Christian Weber, Dagmar E. Ehrnhoefer. A post-translational modification signature defines changes in soluble tau correlating with oligomerization in early stage Alzheimer's disease brain. *Acta Neuropathologica Communications* **7**, 192 (2019). https://doi.org/10.1186/s40478-019-0823-2

Annika Behrendt*, Maria Bichmann*, Ebru Ercan-Herbst, Per Haberkant, David C. Schöndorf, Michael Wolf, Salma A. Fahim, Enrico Murolo, Dagmar E. Ehrnhoefer. Asparagine endopeptidase cleaves tau at N167 after uptake into microglia. *Neurobiology of Disease*, Volume 130, (2019)**.** https://doi.org/10.1016/j.nbd.2019.104518 (*: shared first author)

Ebru Ercan, Sameh Eid, Christian Weber, Alexandra Kowalski, Maria Bichmann, **Annika Behrendt**, Frank Matthes, Sybille Krauss, Peter Reinhardt, Simone Fulle & Dagmar E. Ehrnhoefer. A validated antibody panel for the characterization of tau post-translational modifications. *Molecular Neurodegeneration*, Volume 12, Article 87, (2017). https://doi.org/10.1186/s13024-017-0229-1

Holger Dinkel, Kim Van Roey, Sushama Michael, Manjeet Kumar, Bora Uyar, Brigitte Altenberg, Vladislava Milchevskaya, Melanie Schneider, Helen Kühn, **Annika Behrendt**, Sophie Luise Dahl, Victoria Damerell, Sandra Diebel, Sara Kalman, Steffen Klein, Arne C. Knudsen, Christina Mäder, Sabina Merrill, Angelina Staudt, Vera Thiel, Lukas Welti, Norman E. Davey, Francesca Diella, Toby J. Gibson. ELM 2016—data update and new functionality of the eukaryotic linear motif resource. *Nucleic Acids Research*, Volume 44, Issue D1, (2016). https://doi.org/10.1093/nar/gkv1291

Conferences: Oral and poster presentations

John von Neumann Institute for Computing (NIC) Symposium, September 29th – 30th 2022
Jülich, Germany
Behrendt et al., Vasor: accurately predicting single amino acid substitution impact within the MDR3 protein. (Poster presentation)

International Bile Acid Meeting, July 8th – 9th 2022
Amsterdam, Netherlands
Behrendt et al., Vasor: accurately predicting single amino acid substitution impact within the MDR3 protein. (Poster presentation)

German Conference on Cheminformatics, May 8th – 10th 2022
Garmisch-Partenkirchen, Germany
Behrendt et al., Vasor: accurately predicting single amino acid substitution impact within the MDR3 protein. (Poster presentation)

Neuronus (IBRO Neuroscience Forum), April 20th – 22nd 2018
Krakow, Poland
Behrendt et al., Asparagine endopeptidase cleaves tau at a novel cleavage site in vivo. (Oral presentation)

Grants

Computing time grant on JUWELS/JURECA, Jülich Supercomputing Centre (Project ID: FIC1) May 2022 - April 2023

Chapter 9 Reprinted publications

Publication I *Page 80 to 101*

Reprinted from

Vasor: Accurate prediction of variant effects for amino acid substitutions in multidrug resistance protein 3

A. Behrendt, P. Golchin, F. König, D. Mulnaes, A. Stalke, C. Dröge, V. Keitel, H. Gohlke.

Hepatology Communications (2022) 6, 11.

Received: 15 May 2022 | Accepted: 16 August 2022

DOI: 10.1002/hep4.2088

ORIGINAL ARTICLE

Vasor: Accurate prediction of variant effects for amino acid substitutions in multidrug resistance protein 3

Annika Behrendt[1] | **Pegah Golchin**[2] | **Filip König**[1] | **Daniel Mulnaes**[1] | **Amelie Stalke**[3,4] | **Carola Dröge**[5,6] | **Verena Keitel**[5,6] | **Holger Gohlke**[1,7]

[1]Institute for Pharmaceutical and Medicinal Chemistry, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

[2]Department of Electrical Engineering and Information Technology, Technische Universität Darmstadt, Darmstadt, Germany

[3]Department of Human Genetics, Hannover Medical School, Hannover, Germany

[4]Division of Kidney, Department of Pediatric Gastroenterology and Hepatology, Liver, and Metabolic Diseases, Hannover Medical School, Hannover, Germany

[5]Department for Gastroenterology, Hepatology, and Infectious Diseases, Medical Faculty, Otto von Guericke University, Magdeburg, Germany

[6]Department for Gastroenterology, Hepatology, and Infectious Diseases, University Hospital, Medical Faculty, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

[7]John-von-Neumann-Institute for Computing, Jülich Supercomputing Center, Institute of Biological Information Processing (IBI-7: Structural Biochemistry), and Institute of Bio- and Geosciences (IBG-4: Bioinformatics), Forschungszentrum Jülich GmbH, Jülich, Germany

Correspondence
Holger Gohlke, Institute for Pharmaceutical and Medicinal Chemistry, Heinrich-Heine-Universität Düsseldorf, Universitätsstr. 1, 40225 Düsseldorf, Germany.
Email: gohlke@uni-duesseldorf.de

Funding information
Bundesministerium für Bildung und Forschung, Grant/Award Number: 01GM1904A and 01GM1904B

Abstract

The phosphatidylcholine floppase multidrug resistance protein 3 (MDR3) is an essential hepatobiliary transport protein. MDR3 dysfunction is associated with various liver diseases, ranging from severe progressive familial intrahepatic cholestasis to transient forms of intrahepatic cholestasis of pregnancy and familial gallstone disease. Single amino acid substitutions are often found as causative of dysfunction, but identifying the substitution effect in *in vitro* studies is time and cost intensive. We developed variant assessor of MDR3 (Vasor), a machine learning-based model to classify novel MDR3 missense variants into the categories benign or pathogenic. Vasor was trained on the largest data set to date that is specific for benign and pathogenic variants of MDR3 and uses general predictors, namely Evolutionary Models of Variant Effects (EVE), EVmutation, PolyPhen-2, I-Mutant2.0, MUpro, MAESTRO, and PON-P2 along with other variant properties, such as half-sphere exposure and posttranslational modification site, as input. Vasor consistently outperformed the integrated general predictors and the external prediction tool MutPred2, leading to the current best prediction performance for MDR3 single-site missense variants (on an external test set: F1-score, 0.90; Matthew's correlation coefficient, 0.80). Furthermore, Vasor predictions cover the entire sequence space of MDR3. Vasor is accessible as a webserver at https://cpclab.uni-duesseldorf.de/mdr3_predictor/ for users to rapidly obtain prediction results and a visualization of the substitution site within the MDR3 structure. The MDR3-specific prediction tool Vasor can provide reliable predictions of single-site amino acid substitutions, giving users a fast way to initially assess whether a variant is benign or pathogenic.

INTRODUCTION

Bile formation is a carefully regulated system, from bile acid synthesis to secretion of bile acids across the canalicular membrane. Adenosine triphosphate (ATP)-binding cassette (ABC) transporters present on the canalicular membrane of hepatocytes are responsible for the transport of primary bile components, namely, bile acids through the bile salt export pump (BSEP, *ABCB11*), cholesterol through the ABC subfamily G members 5 and 8 (ABCG5/ABCG8), and phospholipids through multidrug resistance protein 3 (MDR3, *ABCB4*). MDR3 acts as a floppase, translocating substrates, such as phosphatidylcholine, from the inner to the outer membrane leaflet[1,2] and exposing the substrate for extraction into primary bile.[3] Recent studies have suggested different transport pathways that follow either an alternating two-site access model through the protein's inner cavity[4] or a credit-card swipe mechanism along transmembrane helix 7 (TM H7).[5] MDR3 dysfunction has been linked to various liver-associated diseases, including intrahepatic cholestasis of pregnancy, low phospholipid-associated cholelithiasis, drug-induced liver injury, progressive familial intrahepatic cholestasis type 3, liver fibrosis/cirrhosis, and hepatobiliary malignancy.[6–12]

It is estimated that at least 70% of disease-causing *ABCB4* variants are amino acid substitutions, whereas variants leading to premature stop codons and protein truncations are in the minority.[13] However, while the advancement of sequencing allows rapid testing of patients, it remains challenging for clinicians and researchers to assess the potential impact of novel missense variants.

Evaluation of newly found MDR3 amino acid substitutions by *in vitro* cellular assays remains time consuming. Machine-learning-based prediction tools instead offer rapid analysis and have led in recent years to many predictors.[14,15] Nonetheless, general predictors do not consistently perform well on all proteins, necessitating the development of protein-specific prediction tools. To date, there is no MDR3-specific predictor available for classifying amino acid substitutions despite the vital role of MDR3 in bile homeostasis. An initial evaluation of general predictor performances on MDR3 variants suggested MutPred as a well-performing tool[16,17]; however, generalization is difficult due to only 21 tested variants with established cellular effects. Additionally, the tested variants presented a clear bias toward pathogenic effects.

Here, we created an MDR3-specific variant data set and trained a machine-learning algorithm using established general prediction tools, namely Evolutionary Models of Variant Effects (EVE), EVmutation, PolyPhen-2, I-Mutant2.0, MUpro, MAESTRO, and PON-P2,[18–24] as well as half-sphere exposure and posttranslational modification (PTM) site influence as features to obtain an MDR3-specific prediction tool for help in classifying variants as benign or pathogenic (see Figure 1 for a graphical overview). Our predictor, variant assessment of MDR3 (Vasor), performed better than each integrated general predictor. Additionally, Vasor outperformed MutPred2,[25] a general predictor we chose for comparison based on the suggested high performance of its predecessor MutPred on MDR3.[16] We provide easy access to Vasor through a webserver where users can enter a missense variant of interest and obtain a prediction if it is benign or pathogenic together with an estimate of the prediction probability. Additionally, the mutation site is displayed on the structure of MDR3, giving the user a comprehensive view of the local site and the overall position of the assessed variant.

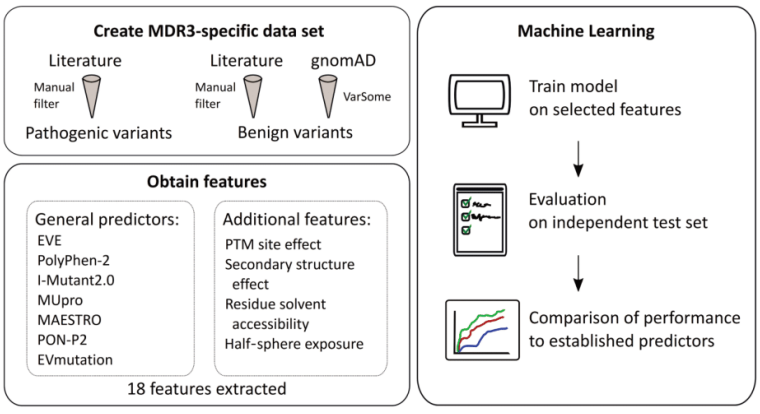


FIGURE 1 Graphical overview of data set generation and machine-learning approach. For details, see text. EVE, Evolutionary Models of Variant Effects; gnomAD, Genome Aggregation Database; MDR3, multidrug resistance protein 3; PTM, posttranslational modification.

MATERIALS AND METHODS

MDR3 missense variants

MDR3 variants were obtained from a literature search for variants causative of MDR3 dysfunction or known variants with no effect in any MDR3-associated disease (see Table S1). We excluded variants with unclear information on disease association (i.e., no *in vitro* verification analysis and no information on clinical indications for disease association) to eliminate false positives (FPs) or false negatives (FNs). As studied benign variants for MDR3 are rare,[13,16] further missense variants were obtained from Genome Aggregation Database (gnomAD) v2.1.1[26] to increase the number of benign variants. During the generation of the gnomAD database, individuals with severe pediatric diseases are removed; however, it is possible that pathogenic variants exist in the gnomAD data set. Accordingly, we employed a selection step to exclude FN cases of MDR3 variants. Using the platform VarSome,[27] variants were preclassified following the guidelines of The American College of Medical Genetics and Association for Molecular Pathology (ACMG-AMP)[28] rules, and variants with a likely pathogenic or pathogenic effect were removed, whereas variants with uncertain significance, likely benign, or benign classification by VarSome were integrated into the data set. These steps were included to create a high-quality data set to keep the number of misclassified variants low but at the same time retain a sufficiently high number of variants. The final list of variants contained 85 pathogenic and 279 benign variants. Every variant was mapped to the longest MDR3 isoform, corresponding to Uniprot[29] entry P21439-1.

Data set and features

The list of MDR3 variants was subjected to general predictors for missense mutations (EVE, PolyPhen-2, I-Mutant2.0, MUpro, MAESTRO, PON-P2, and EVmutation), and additional features (half-sphere exposure, secondary structure disruption, PTM site, and relative solvent accessibility) were computed, creating an MDR3-specific feature set.

EVE is a recently developed, unsupervised, computational method that trained Bayesian variational autoencoders on multiple sequence alignments to classify variant effects based on a computed evolutionary index followed by a fitted global–local mixture of Gaussian mixture models.[18] PolyPhen-2 employs a naive Bayes classifier for predicting variant effects using sequence-based features and structure-based features.[19] I-Mutant2.0 predicts protein stability changes by using a support vector machine-based tool trained on either sequence or structural information.[20] MUpro predicts stability changes on single-site mutations by using sequence and structural information with a support vector machine.[21] Both I-Mutant2.0 and MUpro predict the direction of stability change and the energy difference. MAESTRO employs a combination of machine-learning approaches to predict the energy difference introduced by missense mutations based on consensus, along with predicting a confidence score.[22] PON-P2 applies selected features from evolutionary conservation and biochemical properties of amino acids to develop a random forest classifier that classifies mutations as benign or pathogenic or those with unknown significance.[23] EVmutation explicitly considers interdependencies between residues or nucleotide bases in their unsupervised statistical method to include epistasis.[24]

EVE and EVmutation predictions for the MDR3 protein were accessed using the precomputed data set available from the method creators (https://evemodel.org/, https://marks.hms.harvard.edu/evmutation/human_proteins.html). I-Mutant2.0, Mupro, and MAESTRO predictions were generated using their standalone downloadable versions. PolyPhen-2 predictions were accessed using the batch query of the webserver (http://genetics.bwh.harvard.edu/pph2/bgi.shtml) with the default values. PON-P2 predictions were generated using the sequence submission feature for variants of the webserver (http://structure.bmc.lu.se/PON-P2/).

Additional features were added to explicitly integrate effects on PTM sites, variant location in α-helical or β-sheet secondary structure, and effects on residue solvent accessibility. Known PTM sites from the literature were supplemented by potential PTM sites predicted by PhosphoMotif,[30] PhosphoSitePlus,[31] NetPhos,[32] and the Eukaryotic Linear Motif (ELM) database.[33] The secondary structure was extracted from the MDR3 structure (Protein Data Bank identification [PDB ID]: 6S7P), using the database of secondary structure assignments DSSP.[34,35] Relative solvent accessibility was computed based on residue exposure calculated with DSSP divided by the maximal residue solvent accessibility.[36] Half-sphere exposure was introduced before[37] to measure residue solvent exposure and surpass limitations of relative solvent accessibility. It was implemented using values from the Biopython HSExposure module calculated according to the half-sphere corresponding to the direction of the sidechain of the residue as measured from the Cα atom.

Machine learning

The obtained data set was cleaned from non-numerical values. In the case of binary features, such as classification features of general predictors, −1 was set if no prediction was available to distinguish from benign (value 0) or pathogenic (value 1) predictions. Additionally, relative solvent accessibility and

half-sphere exposure were set to −1 if no prediction value was obtained in order to distinguish from prediction values of 0. Other numerical features were replaced by 0 if no prediction for the respective feature was available. The correlation between features within the data set was assessed by the Spearman *R* correlation coefficient.

A test set was generated by selecting 20 benign and 20 pathogenic variants from the overall data set. To avoid a bias toward specific amino acids, we minimized the root-mean-square deviation (RMSD)-based difference between the amino acid distribution of the variants within the test set compared to the overall data set (Figure S1). After randomly drawing 10 variants into the test set, the RMSD-based difference between the amino acid distribution of the general data set and current test set was computed; further variants were only transferred into the test set if they met one of the following conditions: (a) the RMSD between reference sequence and substituted amino acid distributions decreased by addition of the new variant, (b) the RMSD between reference sequence amino acid distributions decreased while the RMSD between substituted amino acid distributions did not increase more than 0.1, or (c) the RMSD between substituted amino acid distributions decreased while the RMSD between reference sequence amino acid distributions did not increase more than 0.1. Due to the limited size of the data set, it might not otherwise be possible to draw a variant for the test set. The test set was withheld from the machine-learning training step and used for final validation.

To handle the imbalance between the pathogenic (85 variants) and benign (279 variants) class, we used the synthetic minority oversampling technique (SMOTE).[38] This method generates new synthetic data points by using existing minority data points within the *N*-dimensional data set space, drawing lines to the five nearest minority class neighbors, and randomly selecting synthetic data points along these lines to balance out the classes.

On the training data set, the XGBoost algorithm[39] (as implemented in the Python library) was trained using the default gradient-boosted tree (gbtree); the maximum depth of a tree (max_depth) was 3, subsample 0.6, and step size (learning_rate) 0.02. The training was evaluated using repeated *k*-fold cross-validation, with $k = 3$ and the value of repeats (n_repeats) = 5. Using this procedure, the training data set was randomly split into three equally sized folds, where each fold is used as an internal test data set with the remaining two folds as training data sets. The performance results were measured and visualized in receiver operating characteristic (ROC) curves for comparison to the final test set. These steps were repeated 5 times.

To reduce features and estimate feature importance, we analyzed the tree-based feature importance and the permutation importance, leading to the removal of the four least informative features shared in both feature-importance measures: relative solvent accessibility, I-Mutant2.0 stability sign, I-Mutant2.0 deltaG value, and secondary structure disruption. Tree-based feature importance was computed using the XGBoost algorithm built-in feature and the "gain" (average gain across all splits where a feature is used). Permutation-based feature importance was computed by random shuffling each feature consecutively, followed by a performance test; this denoted performance alterations following feature permutation. The performance of the model without feature selection is shown in Figure S2.

The trained model, termed Vasor, predicts a probability ranging from 0 to 1 for a given variant to belong to the pathogenic class. Predictions above (below) 0.5 are classified as pathogenic (benign).

Comparison to established predictors

To assess the general performance of Vasor, we compared it to the general predictors EVE, PolyPhen-2, PON-P2, and MutPred2. MutPred2 predictions were used to compare our prediction tool to an external general predictor as MutPred2 was not used as an input feature for Vasor. The standalone version of MutPred2 was used to classify each variant within the entire data set, and a threshold of 0.5 was used to classify pathogenicity.[25] The performance of Vasor and the other predictors was evaluated on the entire data set and the test set. This ensured increased fairness for the performance comparison as Vasor may have an advantage over other predictors based on its training on the training data set. ROC and precision-recall curves were adjusted to the availability of variants each predictor was able to classify over the entire data set (i.e., if general predictors did not classify a variant into the category benign or pathogenic, the respective variant could not be assessed and curves were shown only on assessable variants). To account for this, the coverage of each predictor of the MDR3 data set was computed.

Performance evaluation

The performance of Vasor and the other prediction tools was evaluated using recommended measures for binary classifiers,[40] including additionally the F1-score as well as visualization in ROC and precision-recall curves. The measures are based on the values of correctly classified variants, indicated by true positives (TPs) for correctly predicted pathogenic variants and true negatives (TNs) for correctly predicted benign variants as well as incorrectly classified variants indicated by FPs for variants predicted as pathogenic

albeit benign and false negatives FNs for variants predicted as benign albeit pathogenic. The analyzed measures of recall, specificity, precision, negative predictive value (NPV), accuracy, F1-score, and Matthew's correlation coefficient (MCC) were calculated as

$$\text{Recall} = \frac{TP}{TP + FN} \quad (1)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (2)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (3)$$

$$\text{NPV} = \frac{TN}{TN + FN} \quad (4)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

$$F1 - \text{score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2\,TP}{2\,TP + FP + FN} \quad (6)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (7)$$

Webserver tool

Vasor can be accessed online at https://cpclab.uni-duesseldorf.de/mdr3_predictor/. Users can enter a single-site amino acid missense MDR3 variant; the tool will only recognize MDR3 variants corresponding to the largest protein isoform UniProt ID: P21439-1. The entry needs to be in the format of the standard International Union of Pure and Applied Chemistry code for amino acids, entering first the one-letter code of the amino acid of the reference sequence, followed by the position and the amino acid substitution of interest. On the results page, users can see the predicted classification (either benign or pathogenic) and the probability of pathogenicity (PoP). This probability ranges from 0 (highest probability for the variant to be benign) to 1 (highest probability for the variant to be pathogenic). Probability values close to 0.5 indicate less confidence in the prediction.

Additionally, the results page displays the structure of the MDR3 protein (PDB ID: 6S7P) with the NGL Viewer,[41,42] including the membrane localization obtained from the Orientations of Proteins in Membranes database[43] as a red and blue plane. The substituted residue is colored according to the predicted effect either in red (pathogenic) or green (benign). The user can download a zip archive containing a high-resolution image of the complete protein, PDB files of the reference sequence and the variant protein, and high-resolution images of the position with the reference sequence residue or the substituted one.

Code availability

The code for Vasor was written in Python 3.9 and is provided for download at https://cpclab.uni-duesseldorf.de/index.php/Software.

RESULTS

Generation of a data set with informative features and good overall coverage of the MDR3 protein

To establish an MDR3-specific prediction tool, we prepared a data set of benign and pathogenic MDR3 variants. Relevant literature on MDR3-associated diseases was screened. Variants with unclear association to effects were omitted to avoid misclassified variants. Additionally, the gnomAD database[26] was screened for MDR3 variants, and the results were subjected to filtering by VarSome[27] using ACMG-AMP rules[28] to remove variants with a high potential for a pathogenic effect. This step was necessary as pathogenic MDR3 variants on a single allele with a potential late-onset or mild phenotype might have been included in the gnomAD database. Next, we used general predictors (EVE,[18] EVmutation,[24] PolyPhen-2,[19] I-Mutant2.0,[20] MUpro,[21] MAESTRO,[22] and PON-P2[23]) and descriptors of the variant site, namely, the disruption of secondary structure, possible PTM site disturbance, and changes in the relative solvent accessibility and half-sphere exposure of the position in question, as features in the data set. Projecting the variant locations from the data set onto the known cryogenic electron microscopy structure of MDR3 (PDB ID: 6S7P)[4] revealed a broad coverage of the structure with benign and pathogenic variants (Figure 2A). No functional domain is devoid of variants, and we do not observe large clusters of benign or pathogenic variants, which may indicate a potential bias within the data set. Such a bias might prevent applying the tool to areas of low coverage. Hence, we expect that our tool can generalize predictions to every position of MDR3.

To further probe for domains of low applicability, we mapped variants misclassified by Vasor to the MDR3 structure. Misclassified variants from the data set tend to occur on the solvent-exposed surface of the protein rather than within buried regions of the protein (Figure S3). As solvent-exposed residues are less evolutionary conserved than buried residues,[44] the obtained trend might visualize the underlying increased

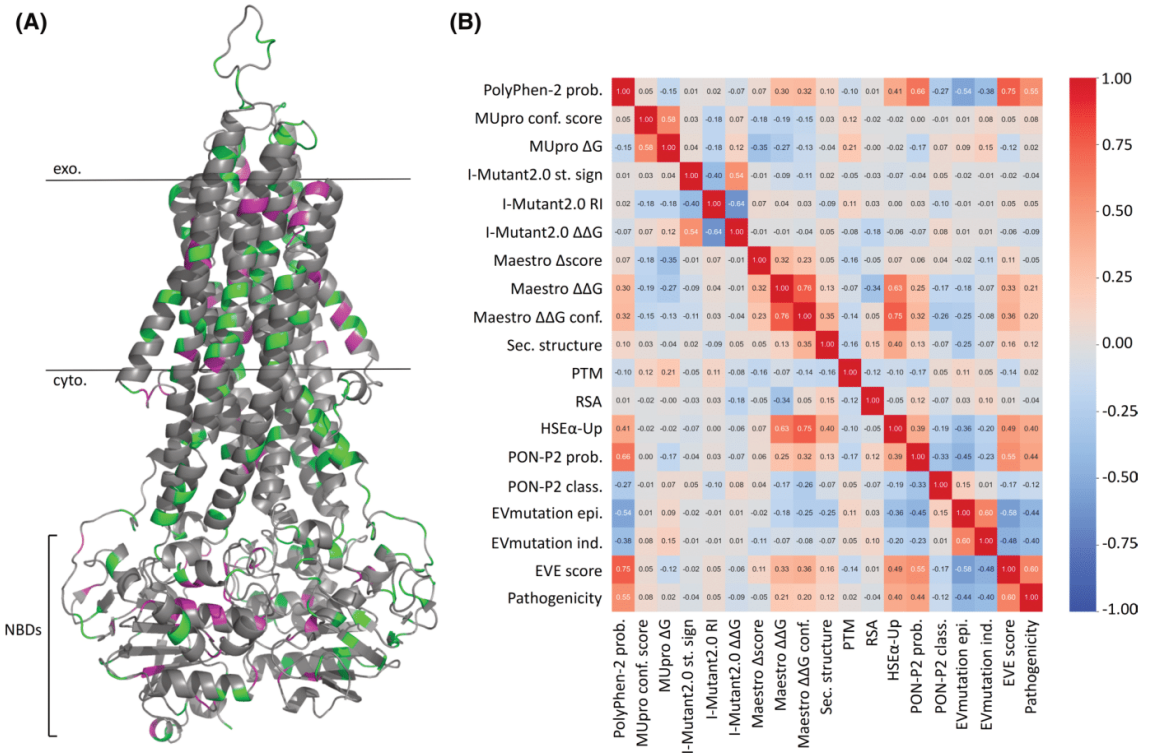


FIGURE 2 Coverage of MDR3 by the data set and correlation analysis of features. (A) Mapping of data set variants onto the MDR3 structure. Benign variants are marked in green and pathogenic variants in magenta. (B) Spearman rank correlation matrix of features computed for the data set. conf., confidence; cyto., cytosolic; epi., epistatic; EVE, Evolutionary Models of Variant Effects; exo., extracellular; HSE, half-sphere exposure; ind., independent; MDR3, multidrug resistance protein 3; NBD, nucleotide-binding domain; prob., probability; PTM, posttranslational modification; RI, reliability index; RSA, relative solvent accessibility; Sec. structure, secondary structure; st. sign, stability sign.

uncertainty of those integrated general predictors that are based on evolutionary sequence conservation. Overall, also given the small number of misclassifications, we do not see indications of domains of increased uncertainty for MDR3 predictions. The correlation coefficients between input features range from −0.64 to 0.76 (RMS value, 0.25) over the 18 features (Figure 2B), indicating that each feature adds information that does not overlap with information from another feature.

Generating Vasor: training the XGBoost algorithm on the data set

For machine-learning models to function reliably, it is vital to estimate potential overfitting or underfitting of the trained model. One of the most important techniques in that respect is the hold-out method, where a subsection of the entire data set is split off as an external test set. Ideally, the test set has a similar probability distribution as the entire data set[45]; however, this is not certain if a test set is randomly drawn. Therefore, we paid attention to drawing our test set with a similar distribution of amino acids as to both reference sequence and variant amino acid distributions by minimizing the RMSD-based difference in amino acid distributions to the overall data set; the test set contained 20 benign and pathogenic variants each (Figure S1).

Next, for the remaining data set, SMOTE[38] was used to create synthetic examples of the minority class (pathogenic variants) to balance the classes. The final training data set consisted of 259 data points for each class, benign and pathogenic, on which an XGBoost algorithm was trained. To evaluate the most important features, we measured and visualized feature importance (Figure S4) and removed the four consistently least important features (Figure S5) without reducing performance. Of note, EVE is highly important for the prediction outcome of the model, indicating that Vasor primarily relies on EVE's predictions compared to other features.

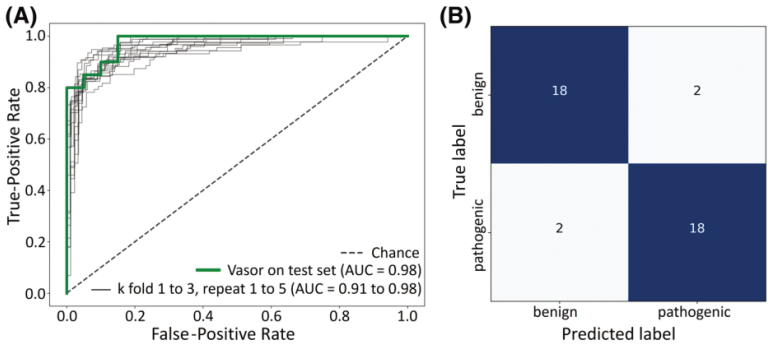


FIGURE 3 Performance of Vasor on the test set. (A) ROC curve of Vasor performance on the test set (green line) compared to performance estimates from repeated *k*-fold cross-validation (black lines). (B) Confusion matrix of Vasor performance on the test set. AUC, area under the curve; ROC, receiver operating characteristic; Vasor, variant assessor of MDR3.

Performance estimates were visualized within a repeated *k*-fold cross-validation and compared to the performance against the held-out test set (Figure 3A). The trained model performs on the test set with an accuracy of 90%, with 18 out of 20 variants being predicted correctly, both for the benign and the pathogenic class (Figure 3B). Notably, the performance based on the *k*-fold cross-validation does not differ from that on the independent test set, indicating a well-fit model without overfitting or underfitting.

Vasor outperforms integrated general predictors and the external general predictor MutPred2

We compared the performance of Vasor with general predictors on the entire data set. We compared Vasor to EVE, PolyPhen-2, and PON-P2, integrated as features into the data set on which Vasor was trained. Vasor should outperform each predictor due to the additional information gathered from the other features. Additionally, we compared Vasor to MutPred2[25] as an external prediction tool; the predecessor tool MutPred was indicated to perform well on MDR3 classification problems.[16] Vasor outperformed EVE, PolyPhen-2, PON-P2, and MutPred2 according to ROC (Figure 4A) and precision-recall curves (Figure 4C), with an area under the curve (AUC) of 0.98 for Vasor against 0.90 for EVE, 0.89 for MutPred2, 0.87 for PolyPhen2, and 0.81 for PON-P2 for the ROC and an AUC of 0.94 for Vasor against an AUC of 0.86 for EVE, 0.74 for MutPred2, 0.72 for PolyPhen2, and 0.55 for PON-P2 for the precision-recall curves. Precision-recall curves have been shown to be more robust and accurate for binary classifiers on imbalanced data sets.[46]

Noteworthy, the second best performing predictor, EVE, was the most important feature for Vasor, suggesting that the machine-learning model recognized the information contained within this feature as highly correlated with the true output and its value in predicting the output correctly. However, EVE could only predict 85.7% of the variants in the data set, whereas Vasor, by design, predicted an outcome for every possible missense variant of MDR3 (Figure 4B; Table 1).

Additional performance measures are summarized in Table 1, indicating that Vasor outperforms existing prediction tools according to the weighted measures F1-score (0.85) and MCC (0.80). Specifically, Vasor achieved a low number of FNs. Comparable low values in FNs were achieved by PolyPhen2 and MutPred2 (but at the cost of an increased number of FPs) and PON-P2, but only at coverage of 45.1% of the variants in the MDR3 protein and an increased number of FPs.

When comparing the performance of the missense predictors on the test set (Table S2), our tool reached the best scores in F1-score and MCC (0.90 and 0.80, respectively) compared to other predictors with full coverage of the test set. EVE showed F1-score and MCC values of 0.91 and 0.83, respectively, on a subset (82.5%) of variants where it reached a prediction. By contrast, MutPred2 was able to predict every pathogenic variant as pathogenic, albeit at the cost of predicting almost half of the benign variants as pathogenic, resulting in a high number of FPs.

Overall, Vasor outperformed other predictors consistently according to ROC and precision-recall curves, revealing a well-balanced prediction with few FNs and FPs, both on the entire data set and the test set.

Vasor classifies the majority of variants with high certainty

Additionally, we investigated the distribution of Vasor's output, the PoP values. Vasor assigns the majority of benign cases low probability values (74% of benign variants <0.24 PoP), whereas the majority of

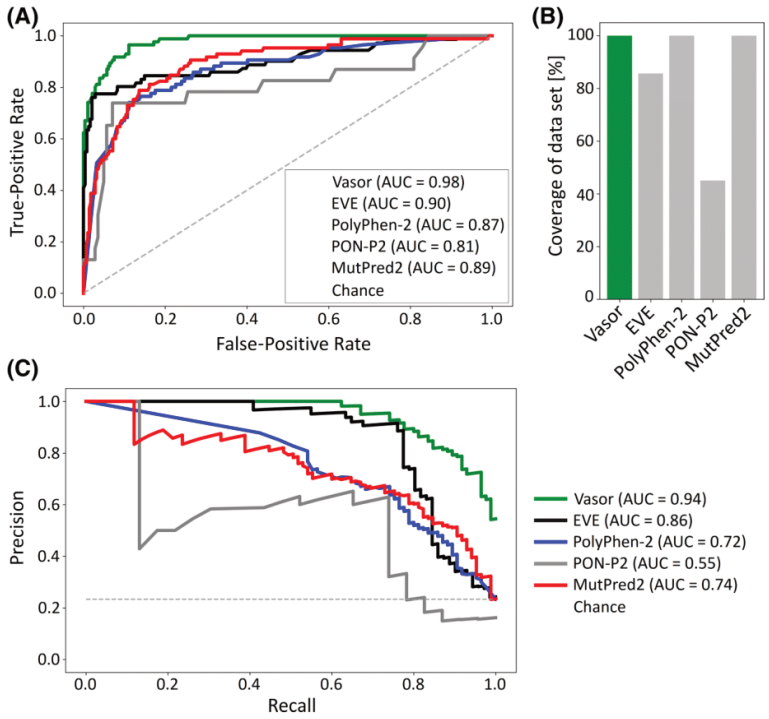


FIGURE 4 Performance of Vasor in comparison to established general predictors. (A) ROC curve of the performance of Vasor, EVE, PolyPhen-2, PON-P2, and MutPred2 on the variants of the entire data set. Note that the performance was determined for those variants each predictor was able to make a prediction for (see [B]). (B) Coverage of data set variants by the predictors. (C) Precision-recall curves of the predictors. Performance was determined for those variants each predictor was able to make a prediction for. AUC, area under the curve; EVE, Evolutionary Models of Variant Effects; ROC, receiver operating characteristic; Vasor, variant assessor of MDR3.

TABLE 1 Detailed performance measurements of Vasor in comparison to EVE, PolyPhen-2, PON-P2, and MutPred2 on the entire data set

	Vasor	EVE	PolyPhen-2	PON-P2	MutPred2
Recall	0.84	0.73	0.84	0.74	0.93
Specificity	0.96	0.98	0.74	0.89	0.67
Precision	0.86	0.91	0.49	0.52	0.46
NPV	0.95	0.93	0.94	0.95	0.97
Accuracy	0.93	0.92	0.76	0.87	0.73
F1-score	0.85	0.81	0.62	0.61	0.61
MCC	0.80	0.77	0.50	0.54	0.51
TP	71	52	71	17	79
FN	14	19	14	6	6
TN	267	236	206	125	186
FP	12	5	73	16	93
Coverage (%)	100	85.7	100	45.1	100

Abbreviations: EVE, Evolutionary Models of Variant Effects; FN, false negative; FP, false positive; MCC, Matthew's correlation coefficient; NPV, negative predictive value; TN, true negative; TP, true positive; Vasor, variant assessor of MDR3.

pathogenic cases are assigned a high probability value (75% of pathogenic variants >0.74 PoP) (Figure 5). Furthermore, Vasor showed no misclassifications of variants in the data set for values below 0.23 and above 0.84, indicating high certainty for benign variant predictions in the range 0–0.23 (74% of the benign variants) and pathogenic variant predictions in the range 0.84–1 (60% of the pathogenic variants).

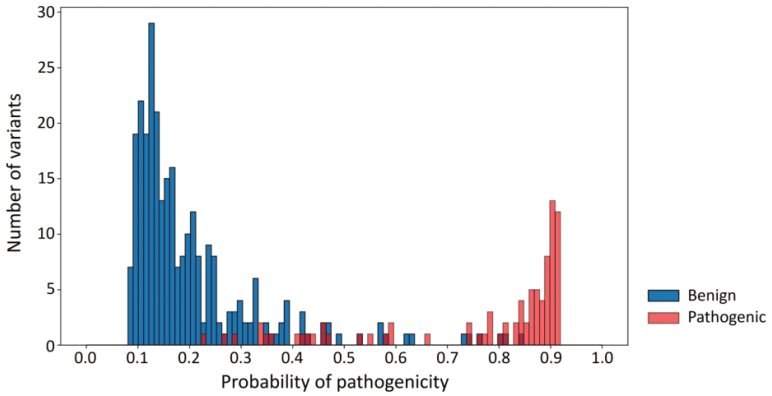


FIGURE 5 Distribution of probability of pathogenicity values over the entire data set. Distribution of Vasor's probability of pathogenicity output for benign (blue) and pathogenic (red) variants. Vasor classified 74% of benign variants into the benign category with values below 0.22, which is below the lowest probability value of any pathogenic variant (0.23) within the data set; 60% of pathogenic variants were classified into the pathogenic category with values above 0.85, which is greater than the highest probability value of any benign variant (0.84) within the data set; 75% of pathogenic variants were classified with probability values greater than 0.74. Vasor, variant assessor of MDR3.

We further investigated the use of SMOTE to generate data points for the minority class (i.e., pathogenic variants). Due to the method underlying SMOTE, SMOTE-generated data points are expected to follow the distribution of pathogenic variants within the PoP curve. Accordingly, no SMOTE data point was predicted with a lower value of PoP than 0.28, and data points mainly clustered within the high certainty zone (Figure S6).

Overall, Vasor showed a robust separation of PoP values of both variant classes, indicating that Vasor classified most variants within the data set with high certainty.

Easy accessibility of Vasor as a webserver tool

Using Vasor, we precalculated the effect of every possible amino acid substitution for MDR3, resulting in a heatmap of 1286 × 20 probabilities of pathogenicity (Figure 6; Table S3). We mapped the average PoP of each position onto the MDR3 protein structure to visualize positions that are functionally more sensitive to substitutions (Figure 7). As expected, areas near the ATP-binding site within the nucleotide-binding domain displayed a high average PoP. Similarly, buried residues within the helices forming the TM part showed high sensitivity as several missense mutations may lead to a disruption of the helical structure. More exposed residues located on the outsides of helices or in flexible regions, such as the small extracellular loops, displayed less sensitivity. However, this trend does not exclude that specific variants at seemingly less sensitive sites can be pathogenic and vice versa.

To indicate the usage of the webserver more specifically, we exemplarily predicted the effect of two variants, V428D and N902D, identified in Dröge et al.[9] These variants were identified in patients without further *in vitro* analysis and not used in the data set for creating Vasor. The variant V428D is predicted to be pathogenic by Vasor with a PoP of 0.77, indicating a good level of certainty for a correct prediction of the pathogenic effect as only four out of 12 variants from the data set were falsely predicted with a similarly high score (Figure 5). V428D is located directly before the Walker A motif, which is important for correctly coordinating the adenosine and the phosphate moiety of ATP in combination with the Walker B motif. Accordingly, the variant might disturb this recognition, resulting in a distorted functionality of MDR3. The variant N902D is predicted to be pathogenic by Vasor with a PoP of 0.90, indicating a high level of certainty for a correct prediction as no false predictions within the data set were observed at such high values (Figure 5). N902D is located in the cytosol-facing part of TM10, with the potential to interact with residues of the X loop of nucleotide-binding domain 1, especially R529. As the X loop is likely involved in relaying the ATP-binding event to the TM domains through conformational change,[47] N902D might exert its effect by hindering this transmission.

We also used the precomputed heatmap for rapid lookup and output generation of the webserver tool, thus eliminating waiting time for users needing a prediction for a specific MDR3 variant. The webserver can be accessed at https://cpclab.uni-duesseldorf.de/mdr3_predictor/. It requires as input an MDR3 variant (with the amino acid of the reference sequence in the one-letter format, its position within the canonical sequence of Uniprot ID: P21439-1, and the substituted amino acid

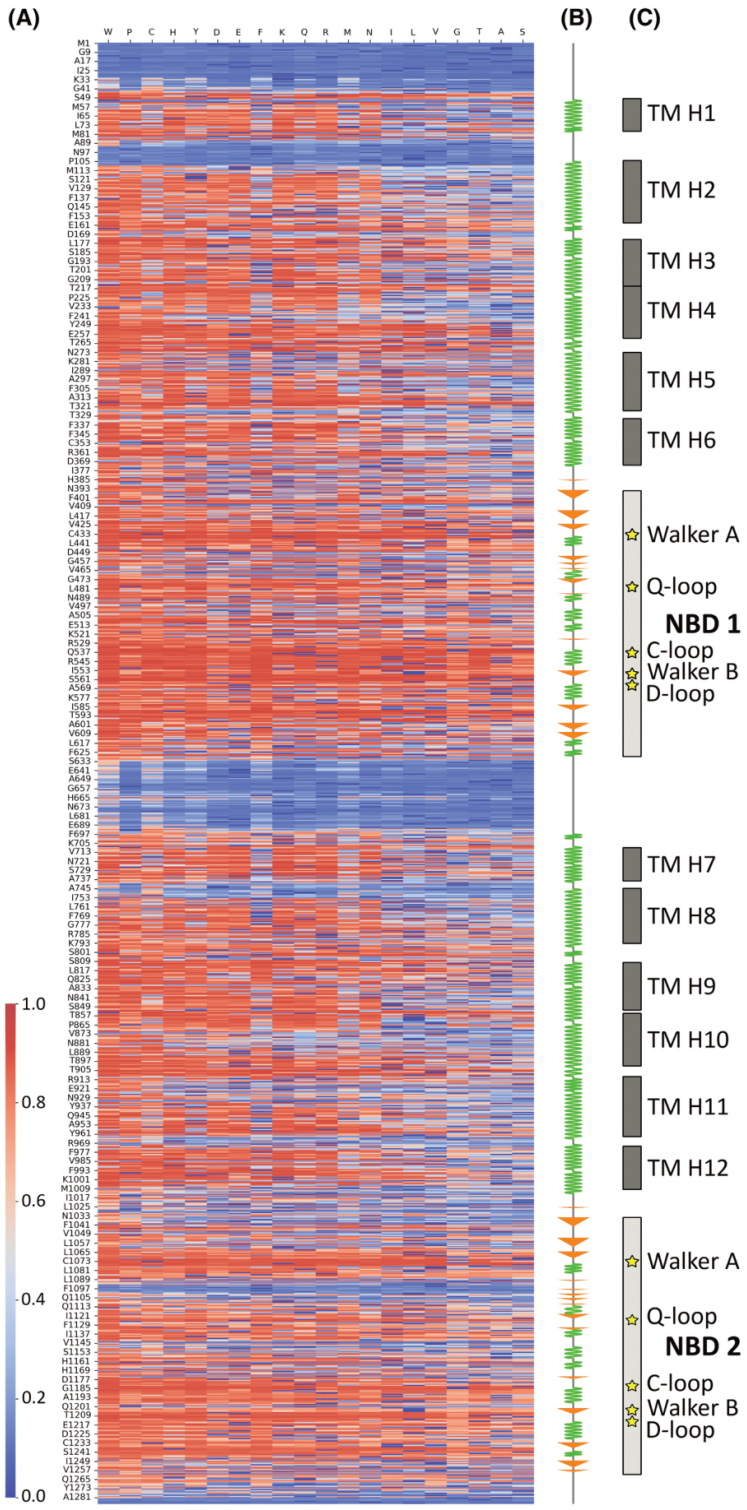


FIGURE 6 Heatmap of predictions for every possible amino acid substitution in MDR3. (A) Color-coded predictions for every position (displayed on the *y* axis) within the MDR3 protein and every possible amino acid substitution (*x* axis). Prediction values range from likely benign (blue) to likely pathogenic (red). (B) Secondary structure of MDR3. α-helical stretches are depicted as green zig-zag curves, β-sheet stretches as orange arrows. (C) Domains, secondary structure elements, and characteristic motives are indicated on the right. MDR3, multidrug resistance protein 3; NBD, nucleotide-binding domain; TM H, transmembrane helix.

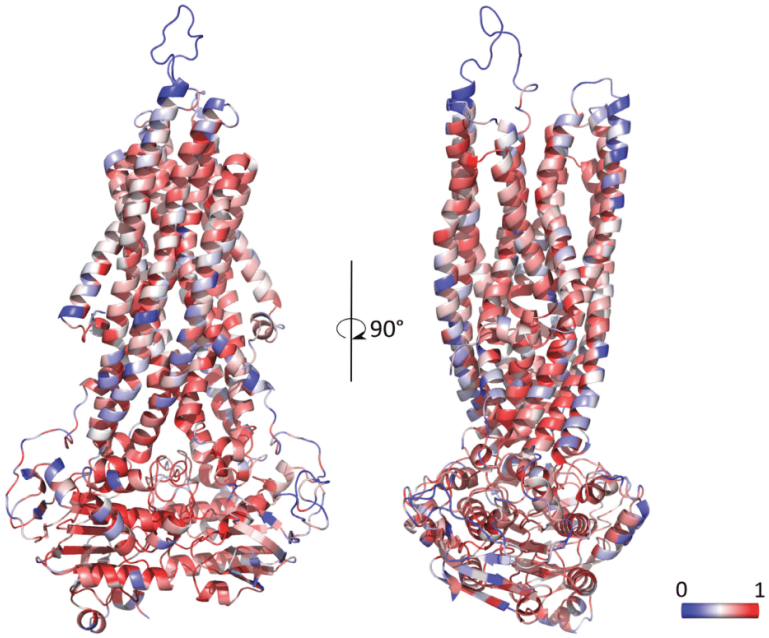


FIGURE 7 Mapping the average pathogenicity onto the structure of MDR3. Prediction values for each position were averaged over all possible substitutions. Values closer to 0 (most likely benign) correspond to blue, values closer to 1 (most likely pathogenic) correspond to red residues. MDR3, multidrug resistance protein 3.

in the one-letter format) and yields the predicted effect of the entered variant, either benign or pathogenic, together with the PoP. Additionally, the variant position is depicted in the three-dimensional structure of MDR3, and high-quality images of reference sequence amino acid, variant, and the overall MDR3 structure can be downloaded. The heatmap is also downloadable from the webserver for implementation in other applications.

DISCUSSION

Although recent years have resulted in many general predictors for protein properties, their performance on specific proteins of interest can differ greatly.[48] While existing state-of-the-art tools to predict substitution effects perform admirably on the MDR3 protein, especially EVE,[18] the potential for improvement is given both for the performance on and coverage of the MDR3 data set because not every general predictor can classify each MDR3 variant. To improve predictions, we created what is to our knowledge the largest data set specific for pathogenic and benign variants of MDR3, obtained from the literature and gnomAD database and comprising 85 pathogenic and 279 benign variants. As the generation of a high-quality data set is a critical first step for any machine-learning approach,[45,49] we carefully screened the literature specifically for MDR3 variants, filtering out variants with unclear disease associations. To counteract the bias that mainly pathogenic variants are chosen for detailed *in vitro* or *in vivo* analysis, we obtained variants from the gnomAD database.[26] Because there may be potentially disease-associated variants in the database, we implemented an additional filtering step of removing variants categorized as likely pathogenic or pathogenic as evaluated by VarSome[27] to exclude FN variants. The data set resulting from this strategy was then kept as is (i.e., no variants were added or removed), thus eliminating the potential to introduce bias from the researcher. Using established general predictors and variant site properties, we trained an MDR3-specific machine-learning model, termed Vasor, to classify

protein missense variants into benign or pathogenic. Vasor outperformed general predictors. Over the entire data set, Vasor showed F1-score and MCC values of 0.85 and 0.80, respectively; the second best method, EVE, followed with scores of 0.81 and 0.77, respectively, but coverage of only 85.7%. By contrast, Vasor ensured high-quality predictions for all MDR3 missense variants. As machine-learning models trained on a specific data set exhibit a bias toward overperformance on this data set, Vasor has an inherent advantage when evaluated on the entire data set over other predictors. Notably, the superior performance of Vasor was also present on the independent test set where Vasor only misclassified two (5%) benign and two (5%) pathogenic variants, leading to the highest performance compared to other predictors, as indicated by F1-score and MCC of 0.9 and 0.8, respectively. Although EVE and PON-P2 achieved similar performances for the test set, they only covered a fraction of the variants (82.5% and 37.5%, respectively). Overall, no other analyzed predictor provided a similarly good balance of consistently low FN and FP predictions. Both measures have important implications for using Vasor within a clinical setting. Predictors with a high number of FNs will lead to variants found within patients being falsely given no attention, whereas a high number of FPs will result in a predictor raising too often a false alarm for an actually benign variant.

We established an easily accessible webserver for reliable and fast predictions of novel MDR3 variants based on Vasor. It can serve as an important step for deciding which variants to study and to provide the first indication of a variant effect. It does not eliminate the need for classical *in vitro* studies for mutational impact, however, and in a clinical setting, the ACMG-AMP guidelines[28] should be followed. The webserver classifies single-site amino acid substitutions into the categories benign or pathogenic. Truncation, insertion, and deletion variants of MDR3 cannot be assessed. However, the PoP for such variants is often more definite.[50] Of note, the effect of a single missense variant within the biological context might not always be a clear-cut pathogenic or benign effect. Therefore, the PoP provided by the webserver can act as an indicator of prediction reliability.

As a limitation, the exact mechanism underlying a pathogenic variant cannot be inferred from the current tool. MDR3 missense variants may impact protein folding and maturation, activity, or stability,[13] and several of these categories can be influenced. Information on mechanistic dysfunction may aid in targeted therapy. In terms of machine learning, such a multiclass classification problem might be solved—with the premise of a sizeable data set of quality-assured variants. Unfortunately, we are unaware of such a data set for MDR3. The currently employed data set strived for such quality-assured variants; however, especially lacking large-scale functional studies of benign variants, variants indicated by VarSome as of unclear significance were included. Thus, we encourage the scientific community to submit novel MDR3 variants with a proven effect on folding, maturation, activity, and stability to the authors to be added to the data set to improve and develop Vasor further.

ACKNOWLEDGMENTS

We are grateful for computational support and infrastructure provided by the "Zentrum für Informations-und Medientechnologie" at the Heinrich Heine University Düsseldorf and the computing time provided by the John von Neumann Institute for Computing to Holger Gohlke on the supercomputer JUWELS at Jülich Supercomputing Centre (user ID: HKF7, VSK33, FIC). Open Access funding enabled and organized by Projekt DEAL. WOA Institution: HEINRICH-HEINE-UNIVERSITAET DUESSELDORF Consortia Name: Projekt DEAL.

FUNDING INFORMATION

BMBF through HiChol (translational network on hereditary intrahepatic cholestasis); Grant Numbers: 01GM1904A, 01GM1904B

CONFLICT OF INTEREST

Verena Keitel is on the speakers' bureau of Falk Foundation and Albireo and advises Astra Zeneca. The other authors have nothing to report.

ORCID

Holger Gohlke https://orcid.org/0000-0001-8613-1447

REFERENCES

1. Smith AJ, Timmermans-Hereijgers JL, Roelofsen B, Wirtz KW, van Blitterswijk WJ, Smit JJ, et al. The human MDR3 P-glycoprotein promotes translocation of phosphatidylcholine through the plasma membrane of fibroblasts from transgenic mice. FEBS Lett. 1994;354(3):263–6.
2. van Helvoort A, Smith AJ, Sprong H, Fritzsche I, Schinkel AH, Borst P, et al. MDR1 P-glycoprotein is a lipid translocase of broad specificity, while MDR3 P-glycoprotein specifically translocates phosphatidylcholine. Cell. 1996;87(3):507–17.
3. Oude Elferink RPJ, Paulusma CC. Function and pathophysiological importance of ABCB4 (MDR3 P-glycoprotein). Pflugers Arch. 2007;453:601–10.
4. Olsen JA, Alam A, Kowal J, Stieger B, Locher KP. Structure of the human lipid exporter ABCB4 in a lipid environment. Nat Struct Mol Biol. 2020;27(1):62–70.
5. Prescher M, Bonus M, Stindt J, Keitel-Anselmino V, Smits SHJ, Gohlke H, et al. Evidence for a credit-card-swipe mechanism in the human PC floppase ABCB4. Structure. 2021;29(10):1144–55.e5.
6. Rosmorduc O, Hermelin B, Poupon R. MDR3 gene defect in adults with symptomatic intrahepatic and gallbladder cholesterol cholelithiasis. Gastroenterology. 2001;120(6):1459–67.
7. Deleuze J, Jacquemin E, Dubuisson C, Cresteil D, Dumont M, Erlinger S, et al. Defect of multidrug-resistance 3 gene expression in a subtype of progressive familial intrahepatic cholestasis. Hepatology. 1996;23(4):904–8.
8. Lang C, Meier Y, Stieger B, Beuers U, Lang T, Kerb R, et al. Mutations and polymorphisms in the bile salt export pump

and the multidrug resistance protein 3 associated with drug-induced liver injury. Pharmacogenet Genomics. 2007; 17(1):47–60.
9. Dröge C, Bonus M, Baumann U, Klindt C, Lainka E, Kathemann S, et al. Sequencing of FIC1, BSEP and MDR3 in a large cohort of patients with cholestasis revealed a high number of different genetic variants. J Hepatol. 2017;67(6):1253–64.
10. Pauli-Magnus C, Lang T, Meier Y, Zodan-Marin T, Jung D, Breymann C, et al. Sequence analysis of bile salt export pump (ABCB11) and multidrug resistance p-glycoprotein 3 (ABCB4, MDR3) in patients with intrahepatic cholestasis of pregnancy. Pharmacogenetics. 2004;14:91–102.
11. Gudbjartsson DF, Helgason H, Gudjonsson SA, Zink F, Oddson A, Gylfason A, et al. Large-scale whole-genome sequencing of the Icelandic population. Nat Genet. 2015;47(5):435–44.
12. Dong C, Condat B, Picon-Coste M, Chrétien Y, Potier P, Noblinski B, et al. Low-phospholipid-associated cholelithiasis syndrome: prevalence, clinical features, and comorbidities. JHEP Rep. 2020;3(2):100201.
13. Delaunay JL, Durand-Schneider AM, Dossier C, Falguières T, Gautherot J, Davit-Spraul A, et al. A functional classification of ABCB4 variations causing progressive familial intrahepatic cholestasis type 3. Hepatology. 2016;63(5):1620–31.
14. Hassan MS, Shaalan AA, Dessouky MI, Abdelnaiem AE, ElHefnawi M. A review study: computational techniques for expecting the impact of non-synonymous single nucleotide variants in human diseases. Gene. 2019;680:20–33.
15. Niroula A, Vihinen M. Variation interpretation predictors: principles, types, performance, and choice. Hum Mutat. 2016;37:579–97.
16. Khabou B, Durand-Schneider AM, Delaunay JL, Aït-Slimane T, Barbu V, Fakhfakh F, et al. Comparison of in silico prediction and experimental assessment of ABCB4 variants identified in patients with biliary diseases. Int J Biochem Cell Biol. 2017;89:101–9.
17. Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, et al. Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics. 2009;25(21):2744–50.
18. Frazer J, Notin P, Dias M, Gomez A, Min JK, Brock K, et al. Disease variant prediction with deep generative models of evolutionary data. Nature. 2021;599(7883):91–5. Erratum in: Nature. 2022;601(7892):E7.
19. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7:248–9.
20. Capriotti E, Fariselli P, Casadio R. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res. 2005;33:W306–10.
21. Cheng J, Randall A, Baldi P. Prediction of protein stability changes for single-site mutations using support vector machines. Proteins. 2006;62(4):1125–32.
22. Laimer J, Hofer H, Fritz M, Wegenkittl S, Lackner P. MAESTRO - multi agent stability prediction upon point mutations. BMC Bioinformatics. 2015;16:116.
23. Niroula A, Urolagin S, Vihinen M. PON-P2: Prediction method for fast and reliable identification of harmful variants. PLoS One. 2015;10(2):e0117380.
24. Hopf TA, Ingraham JB, Poelwijk FJ, Schärfe CPI, Springer M, Sander C, et al. Mutation effects predicted from sequence co-variation. Nat Biotechnol. 2017;35(2):128–35.
25. Pejaver V, Urresti J, Lugo-Martinez J, Pagel KA, Lin GN, Nam HJ, et al. Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. Nat Commun. 2020;11(1):5918.
26. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al.; Genome Aggregation Database Consortium. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434–43. Erratum in: Nature. 2021;590(7846):E53.
27. Kopanos C, Tsiolkas V, Kouris A, Chapple CE, Albarca Aguilera M, Meyer R, et al. VarSome: the human genomic variant search engine. Bioinformatics. 2019;35(11):1978–80.
28. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015;17(5):405–24.
29. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. 2021;49:D480–9.
30. Amanchy R, Periaswamy B, Mathivanan S, Reddy R, Tattikota SG, Pandey A. A curated compendium of phosphorylation motifs. Nat Biotechnol. 2007;25(3):285–6.
31. Hornbeck PV, Zhang B, Murray B, Kornhauser JM, Latham V, Skrzypek E. PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. Nucleic Acids Res. 2015;43:D512–20.
32. Blom N, Gammeltoft S, Brunak S. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. J Mol Biol. 1999;294(5):1351–62.
33. Kumar M, Gouw M, Michael S, Sámano-Sánchez H, Pancsa R, Glavina J, et al. ELM—the eukaryotic linear motif resource in 2020. Nucleic Acids Res. 2020;48:D296–306.
34. Joosten RP, te Beek TAH, Krieger E, Hekkelman ML, Hooft RWW, Schneider R, et al. A series of PDB related databases for everyday needs. Nucleic Acids Res. 2011;39:D411–9.
35. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983;22(12):2577–637.
36. Tien MZ, Meyer AG, Sydykova DK, Spielman SJ, Wilke CO. Maximum allowed solvent accessibilites of residues in proteins. PLoS One. 2013;8(11):e80635.
37. Hamelryck T. An amino acid has two sides: a new 2D measure provides a different view of solvent exposure. Proteins. 2005;59(1):38–48.
38. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–57.
39. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. 2016. Available from: https://arxiv.org/abs/1603.02754. Accessed March 1, 2021.
40. Vihinen M. How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. BMC Genomics. 2012;13(Suppl 4):S2.
41. Rose AS, Bradley AR, Valasatava Y, Duarte JM, Prlić A, Rose PW. NGL viewer: web-based molecular graphics for large complexes. Bioinformatics. 2018;34:3755–8.
42. Rose AS, Hildebrand PW. NGL Viewer: a web application for molecular visualization. Nucleic Acids Res. 2015;43(W1):W576–9.
43. 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;40:D370–6.
44. Echave J, Spielman SJ, Wilke CO. Causes of evolutionary rate variation among protein sites. Nat Rev Genet. 2016;17(2):109–21.
45. Raschka S. Model evaluation, model selection, and algorithm selection in machine learning. 2018. Available from: https://arxiv.org/abs/1811.12808. Accessed May 15, 2021.
46. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015;10(3):e0118432.
47. Dawson RJP, Locher KP. Structure of a bacterial multidrug ABC transporter. Nature. 2006;443(7108):180–5.
48. Riera C, Padilla N, de la Cruz X. The complementarity between protein-specific and general pathogenicity predictors for amino acid substitutions. Hum Mutat. 2016;37(10):1013–24.
49. Walsh I, Pollastri G, Tosatto SCE. Correct machine learning on protein sequences: a peer-reviewing perspective. Brief Bioinform. 2016;17(5):831–40.

50. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536(7616):285–91.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Behrendt A, Golchin P, König F, Mulnaes D, Stalke A, Dröge C, et al. Vasor: Accurate prediction of variant effects for amino acid substitutions in multidrug resistance protein 3. Hepatol Commun. 2022;6:3098–3111. https://doi.org/10.1002/hep4.2088

Supporting Information

Vasor: Accurate prediction of variant effects for amino acid substitutions in MDR3

Annika Behrendt[1], Pegah Golchin[2], Filip König[1], Daniel Mulnaes[1], Amelie Stalke[3,4], Carola Dröge[5,6], Verena Keitel[5,6], Holger Gohlke[1,7,*]

[1]Institute for Pharmaceutical and Medicinal Chemistry, Heinrich Heine University Düsseldorf, Germany
[2]Department of Electrical Engineering and Information Technology, Technische Universität Darmstadt
[3]Department of Human Genetics, Hannover Medical School, Hannover, Germany
[4]Department of Pediatric Gastroenterology and Hepatology, Division of Kidney, Liver and Metabolic Diseases, Hannover Medical School, Hannover, Germany
[5]Department for Gastroenterology, Hepatology and Infectious Diseases, Medical Faculty, Otto von Guericke University, Magdeburg, Germany
[6]Department for Gastroenterology, Hepatology and Infectious Diseases, University Hospital, Medical Faculty, Heinrich Heine University Düsseldorf, Germany
[7]John-von-Neumann-Institute for Computing (NIC), Jülich Supercomputing Centre (JSC), Institute of Biological Information Processing (IBI-7: Structural Biochemistry), and Institute of Bio- and Geosciences (IBG-4: Bioinformatics), Forschungszentrum Jülich GmbH, 52428 Jülich

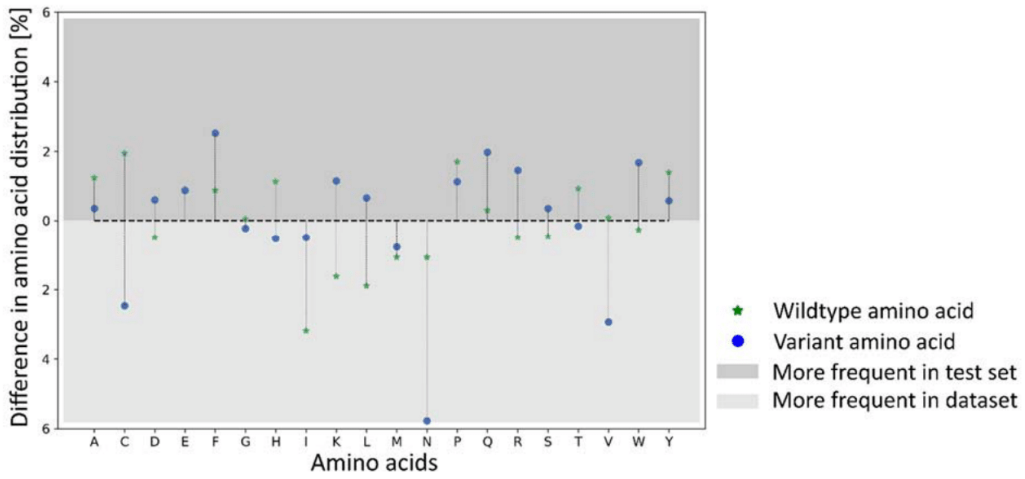
Supplemental Tables

SI Tables 1 and 3 are provided as separate .xlsx files.

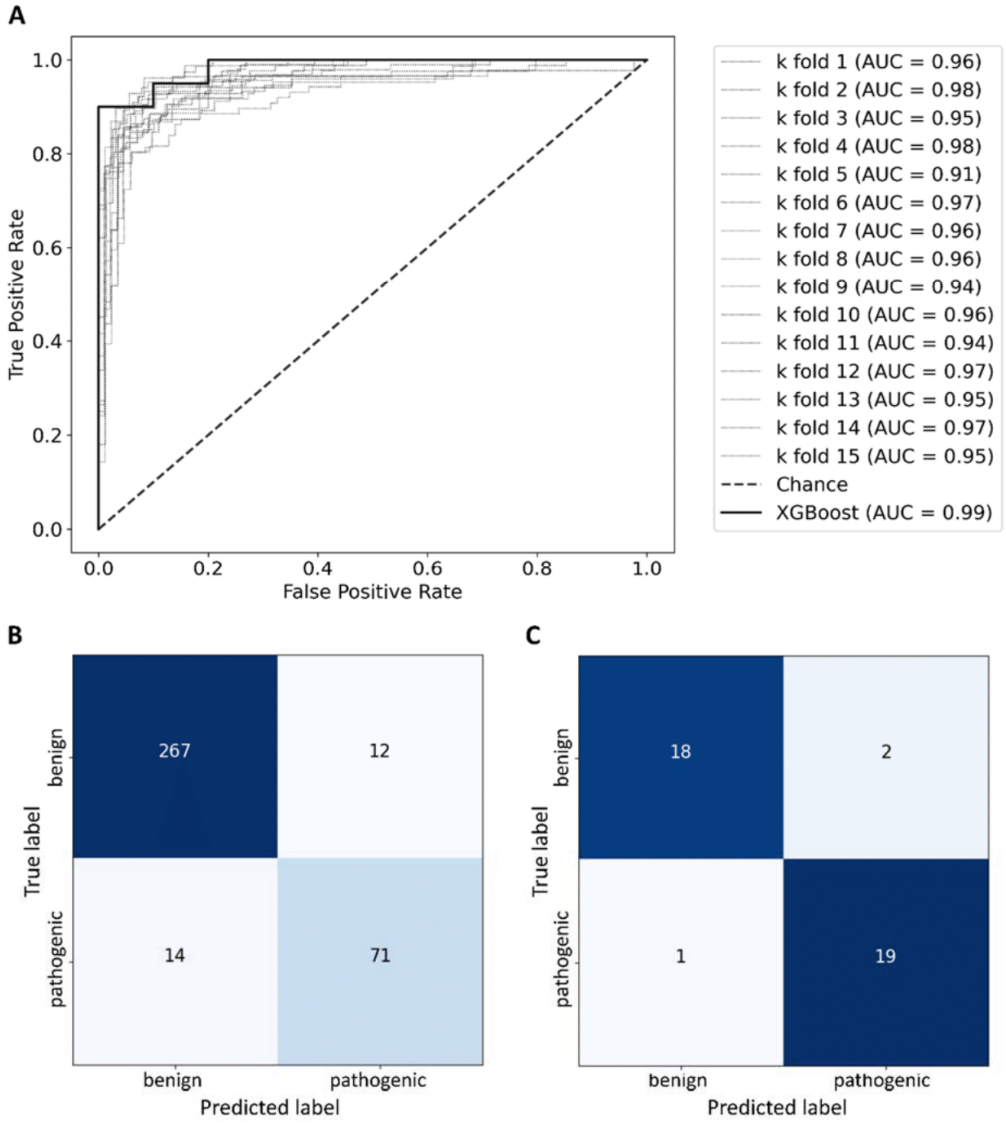
SI Table 2: Detailed performance measurements of Vasor in comparison to EVE, PolyPhen-2, PON-P2, and MutPred2 on the independent test set.

	Vasor	EVE	PolyPhen-2	PON-P2	MutPred2
Recall	0.90	0.83	0.95	0.75	1.00
Specificity	0.90	1.00	0.80	1.00	0.55
Precision	0.90	1.00	0.83	1.00	0.69
NPR	0.90	0.83	0.94	0.92	1.00
Accuracy	0.90	0.91	0.88	0.93	0.78
F1-Score	0.90	0.91	0.88	0.86	0.82
MCC	0.80	0.83	0.76	0.83	0.62
TP	18	15	19	3	20
FN	2	3	1	1	0
TN	18	15	16	11	11
FP	2	0	4	0	9
Coverage [%]	100	82.5	100	37.5	100

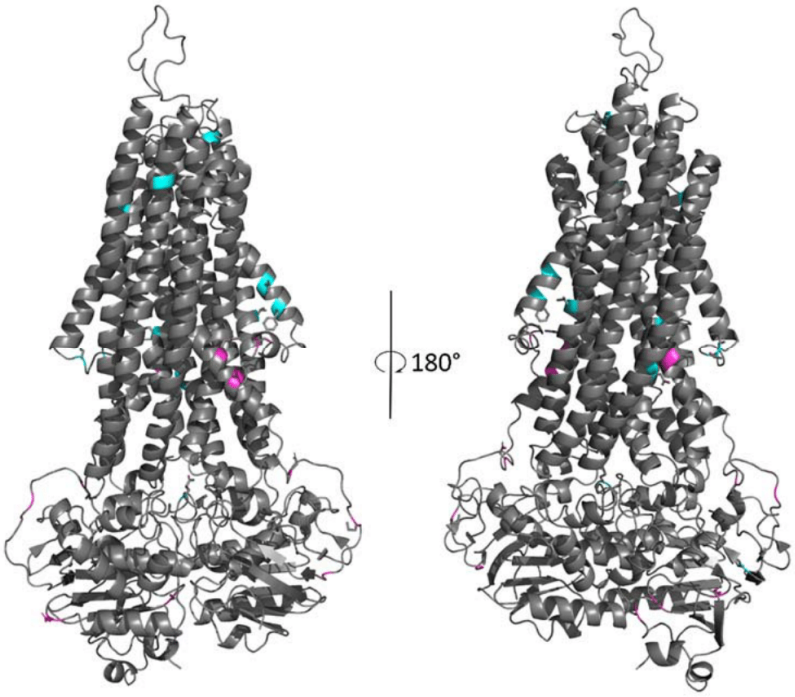
Supplemental Figures



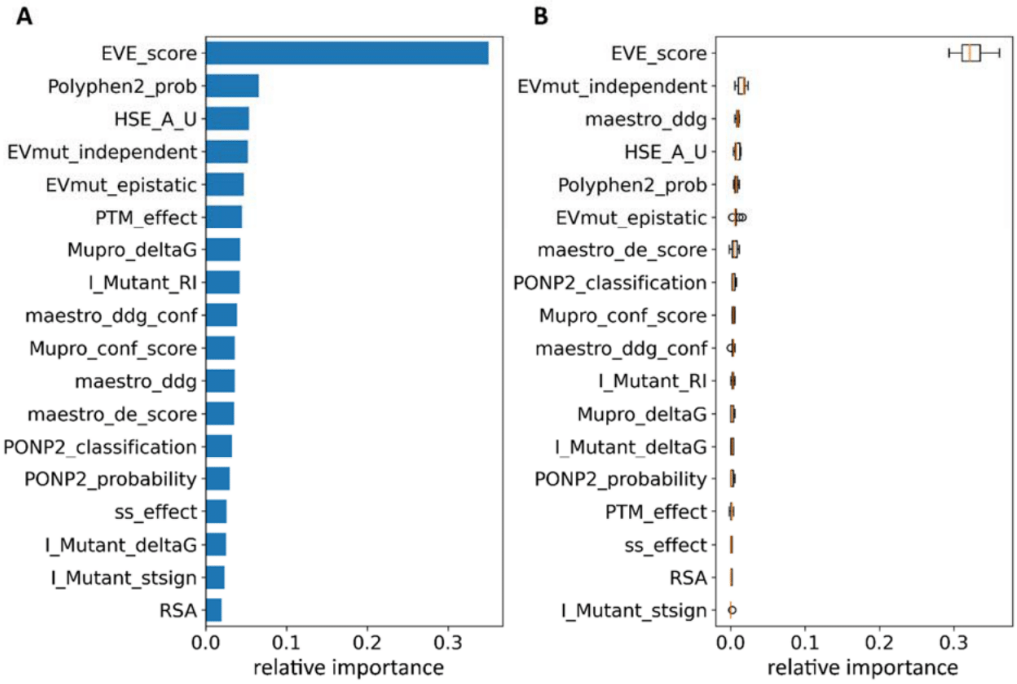
SI Fig. 1: Agreement of the distributions of amino acids between the entire dataset and the test set. The distribution differences of reference sequence and variant amino acids between dataset and test set was computed as RMSD differences.



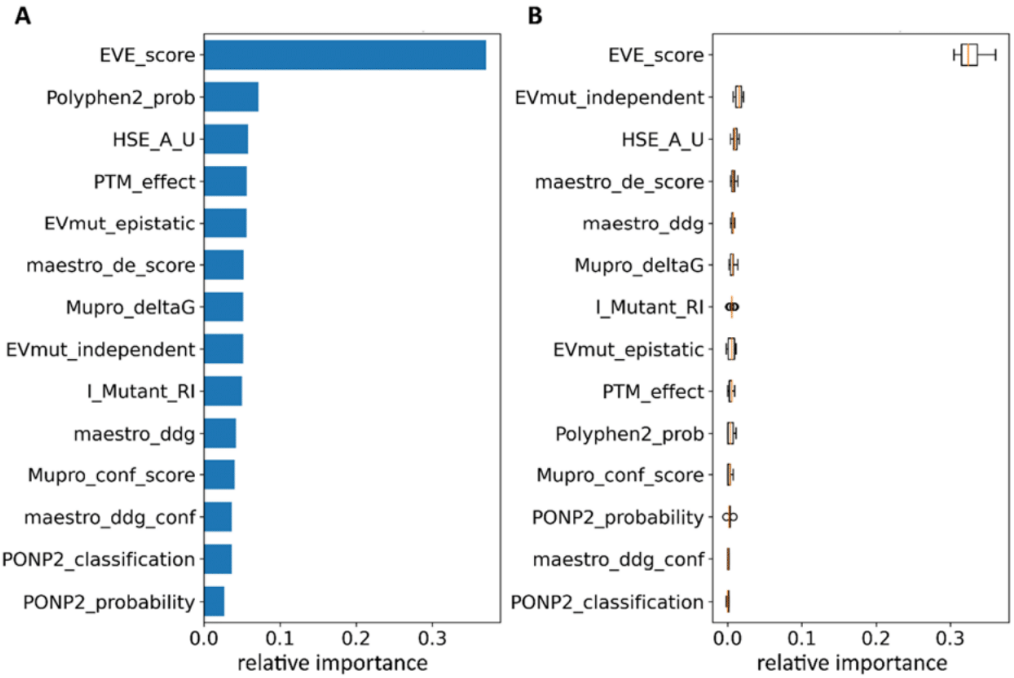
SI Fig. 2: XGBoost model performance without feature selection. [A] ROC curve of the performance of an XGBoost model trained on every feature within the dataset. The performance of the model on the test set (solid line) is compared to the performances during the repeated *k*-fold cross-validation (dotted lines). [B] Confusion matrix of the model on the entire dataset. [C] Confusion matrix of the model on the test set.



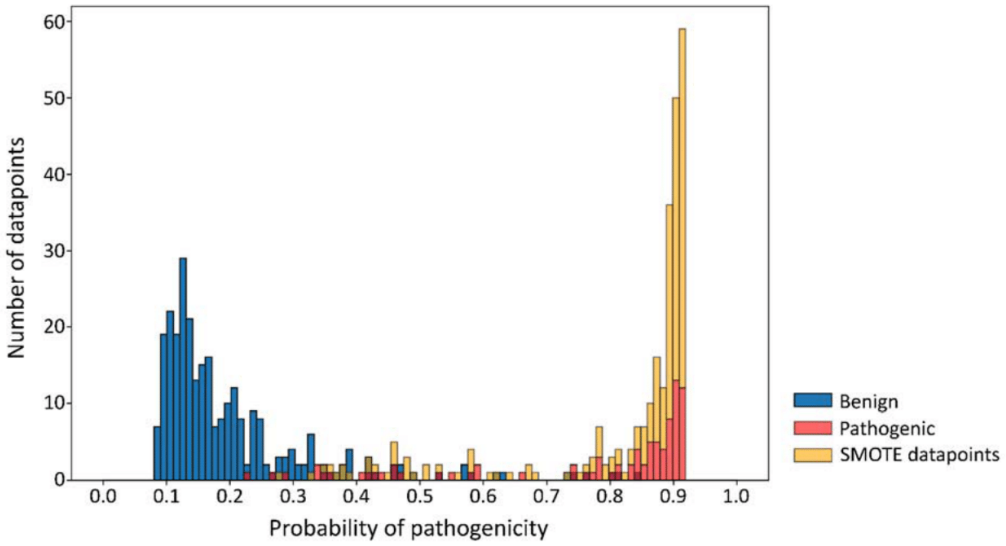
SI Fig. 3: Distribution of misclassified variants. Misclassified variants were mapped to the MDR3 structure. Vasor-misclassified False Negatives are depicted in cyan and False Positives in pink.



SI Fig. 4: Importance of the features. [A] Tree-based feature importance. [B] Permutation importance. Each feature was subjected to permutation for 10 repeats. Mean values of those repeats are depicted as orange lines, with the box ranging from the first to the third quartile of the data. The whiskers extend 1.5 times the inter-quartile range from the box. Outlier points located further than the whiskers are depicted as points if present.



SI Fig. 5: Importance of the features in Vasor. [A] Tree-based feature importance. [B] Permutation importance. Each feature was subjected to permutation for 10 repeats. Mean values of those repeats are depicted as orange lines, with the box ranging from the first to the third quartile of the data. The whiskers extend 1.5 times the inter-quartile range from the box. Outlier points located further than the whiskers are depicted as points if present.



SI Fig. 6: Distribution of probability of pathogenicity values over the entire dataset including SMOTE-generated data points. Distribution of Vasor's probability of pathogenicity output for benign (blue) and pathogenic (red) variants, and SMOTE-generated data points for the pathogenic class (orange). Pathogenic and SMOTE data points are represented as stacked bars. Vasor classified 74 % of benign variants into the benign category with values below 0.22, which is below the lowest probability value of any pathogenic variant (0.23) within the dataset. 70 % of pathogenic variants and SMOTE data points were classified into the pathogenic category with values above 0.84, which is greater than the highest probability value of any benign variant (0.84) within the dataset. 75 % of pathogenic variants and SMOTE data points were classified with probability values > 0.80.

Publication II *Page 103 to 148*

Reprinted from

Impaired transitioning of the FXR ligand binding domain to an active state underlies a PFIC5 phenotype

A. Behrendt, J. Stindt, E.-D. Pfister, K. Grau, S. Brands, C. Dröge, A. Stalke, M. Bonus, M. Sgodda, T. Cantz, A. Bastianelli, U. Baumann, V. Keitel, H. Gohlke.

bioRxiv preprint, DOI: 10.1101/2024.02.08.579530

Impaired transitioning of the FXR ligand binding domain to an active state underlies a PFIC5 phenotype

Annika Behrendt[1], Jan Stindt[2], Eva-Doreen Pfister[3], Kathrin Grau[1], Stefanie Brands[1], Alex Bastianelli[4], Carola Dröge[2,4], Amelie Stalke[5], Michele Bonus[1], Malte Sgodda[6], Tobias Cantz[6,7], Sabine Franke[8], Ulrich Baumann[3], Verena Keitel[2,4]* and Holger Gohlke[1,9]*

[1] Institute for Pharmaceutical and Medicinal Chemistry, Heinrich Heine University, Düsseldorf, Germany

[2] Department of Gastroenterology, Hepatology and Infectious Diseases, Medical Faculty University Hospital Düsseldorf, Heinrich Heine University, Düsseldorf, Germany

[3] Pediatric Gastroenterology and Hepatology, Department for Pediatric Kidney, Liver and Metabolic Diseases, Hannover Medical School, Hannover, Germany

[4] Department of Gastroenterology, Hepatology and Infectious Diseases, Medical Faculty, University Hospital Magdeburg, Otto von Guericke University, Magdeburg, Germany

[5] Department of Human Genetics, Hannover Medical School, Hannover, Germany

[6] Research Group Translational Hepatology and Stem Cell Biology, Department of Gastroenterology, Hepatology, Infectious Diseases and Endocrinology, Hannover Medical School, Hannover, Germany

[7] REBIRTH-Research Center for Translational Regenerative Medicine, Hannover Medical School, Hannover, Germany

[8] Institute of Pathology, University Hospital Magdeburg, Otto von Guericke University, Magdeburg, Germany

[9] John-von-Neumann-Institute for Computing, Jülich Supercomputing Center, Institute of Biological Information Processing (IBI-7: Structural Biochemistry), and Institute of Bio- and Geosciences (IBG-4: Bioinformatics), Forschungszentrum Jülich GmbH, Jülich, Germany

*corresponding authors

Contact information / Correspondence

Prof. Dr. rer. nat. Holger Gohlke

Institute for Pharmaceutical and Medicinal Chemistry,

Heinrich-Heine-Universität Düsseldorf

Universitätsstr. 1

40225 Düsseldorf

Germany

Phone: +49 211 81 13662

Email: gohlke@uni-duesseldorf.de

Prof. Dr. med. Verena Keitel

Department of Gastroenterology, Hepatology and Infectious Diseases, Medical Faculty of Otto-von-Guericke University Magdeburg, University Hospital Magdeburg AöR

Leipziger Str. 44

39120 Magdeburg

Germany

Fax: +49 391 6713105

Phone: +49 391 6713100

Email: verena.keitel-anselmino@med.ovgu.de

Keywords: farnesoid X receptor, *NR1H4*, progressive familial intrahepatic cholestasis, missense variant, molecular dynamics simulations

Abbreviations

FXR farnesoid X receptor, LBD ligand binding domain; PFIC progressive familial intrahepatic cholestasis; BA bile acids; RXR retinoic X receptor; MD Molecular Dynamics; NR nuclear receptor; RMSD root mean square deviation; H12 helix 12; AF activation function; OCA obeticholic acid; CDCA chenodeoxycholic acid; NCoA2 nuclear receptor coactivator 2;

Funding

This study was supported by the BMBF through HiChol (01GM1904A and 01GM2204A to V.K. and C.D.; 01GM1904A and 01GM2204B to H.G.; 01GM1904B and 01GM2204C to U.B., A.S., E.P., TC).

Acknowledgment

Expert technical assistance by Paulina Philippski and Aileen Nötzold is gratefully acknowledged. We are grateful for computational support by the “Zentrum für Informations und Medientechnologie” at the Heinrich-Heine-Universität Düsseldorf and the computing time provided by the John von Neumann Institute for Computing (NIC) to H.G. on the supercomputer JUWELS at Jülich Supercomputing Centre (JSC) (user IDs: VSK33; FIC1).

Conflict of interest statement

The authors declare that there is no potential conflict of interest.

Abstract

Nuclear receptor farnesoid X receptor (FXR) acts as a key regulator of bile acid pool homeostasis and metabolism. Within the enterohepatic circulation, reabsorbed bile acids act as agonists on FXR, which transcriptionally controls the synthesis and transport of bile acids. Binding occurs in the ligand binding domain (LBD), favoring a conformational change to the active state in which helix 12 interacts with the LBD to form an interaction surface for nuclear co-activators. The homozygous missense variant T296I, identified in a PFIC5 patient, is located close to the critical helix 12 interaction. Here, we identified reduced transcriptional activity of the variant protein on the downstream targets BSEP and SHP *in vitro* and within the patient's liver. Analysis of the structural dynamics of the conformational change from an inactive to an active state of the FXR LBD with molecular dynamics simulations revealed that while the wildtype protein frequently transitions into the active state, this movement and the necessary perfect placement of helix 12 was significantly impeded in the T296I mutated protein. To our knowledge, this is the first study to describe the conformational change from an inactive to an active state of the FXR LBD. This might be useful for new therapeutic approaches targeting the activation of FXR. Overall, combining *in vivo* data with *in vitro* and *in silico* experiments, we suggest a molecular mechanism underlying the PFIC phenotype of a patient with an FXR missense variant.

Introduction

Progressive familial intrahepatic cholestasis (PFIC) is a rare group of genetic disorders that affect the liver's ability to excrete bile constituents, resulting in impaired bile flow, subsequent intrahepatic cholestasis, and progressive liver damage and failure (1, 2).

Farnesoid X receptor (FXR), encoded by the *NR1H4* gene, is a nuclear receptor (NR) responsive to bile acids (BA) and a key regulator of BA metabolism, playing a pivotal role in maintaining BA homeostasis by controlling BA synthesis, transport, and detoxification (3, 4). *NR1H4* variants associated with PFIC (subtype 5) were characterized by coagulopathy and a rapid progression toward end-stage liver disease (5-7). While most patients carried bi-allelic protein-truncating variants (5-8), only two *NR1H4*-associated PFIC patients carrying homozygous missense variants have been identified (7, 9). While one patient died on the transplant waiting list due to end-stage liver disease at the age of 9 months (c.557G>A) (7), the other patient was successfully transplanted at the age of 8 months (c.887C>T, p.(Thr296Ile), referred to as T296I in the following) and is currently 10 years old (9). FXR and BSEP staining was found negative in the liver tissue of PFIC patients with protein-truncating NR1H4 variants (5, 8). To determine the contribution of the homozygous *NR1H4* T296I missense variant to the PFIC phenotype of our patient, we studied the localization and transcriptional activity of the mutated protein *in vitro*.

Molecular dynamics (MD) simulations have proven useful in elucidating the functional mechanisms of protein activity (10). In particular, nuclear receptors (NRs) have benefited from this in-depth analysis as their functions are often diverse, and subtle changes in ligands can lead to altered conformations and, thus, protein activity (11-13). The positioning of helix 12 (H12), forming part of the activation function 2 (AF2) surface, is pivotal for NR activity via the recruitment of coregulatory proteins. Coactivators interact with the AF2 surface using a conserved LXXLL motif (14), while antagonist-bound NRs favor corepressor binding to the AF2 surface with a larger hydrophobic motif and blocking the active positioning of H12 (15). Several MD studies of the LBD of FXR have underlined the importance of H12 positioning (16-18). However, the transitioning from the inactive to the active conformation as well as the effect of

single-site missense variants on the function of FXR has so far not been analyzed by MD studies. Thus, we employed MD simulations to analyze the conformational change from an inactive to an active state and evaluated the impact of the T296I variant both with a localized distance measurement and with regard to its influence on H12 positioning.

Materials and Methods

Plasmids, cloning and mutagenesis

The BSEP promoter plasmid based on pGL3-basic (BSEP^{prom}-Luc) was a kind gift from Roche. The human *SHP* promoter (bases -572 to +10, GenBank Accession Number AF044316) (19) was amplified by PCR from a healthy human liver genomic DNA pool. DNA sequencing was performed for all cDNAs used (Eurofins). Note that the numbering of the protein variant (T296I) is based on the alpha1 isoform (Uniprot acc. Q96RI1-1). For details on the cloning strategies see SI.

Immunofluorescence staining of HEK293 cells

HEK293 cells seeded onto glass coverslips in 12-well plates were transiently transfected with 1µg each of FXR and retinoic X receptor (RXR) α expression constructs for 48h. After 24h, cells were stimulated with obeticholic acid (OCA, INT-747, 10µM) and 9-cis-retinoic acid (9-cis-RA, 1µM). Cells on coverslips were washed with PBS before fixation with ice-cold methanol (30sec). After blocking in UltraVision protein block (ThermoFisher Scientific) for 30min, cells were stained for 1h at 1:100 with rabbit anti-FXR (H-130; sc-13063, Santa Cruz Biotechnology) followed by staining at 1:250 with goat anti-rabbit-IgG-FITC (Jackson ImmunoResearch) and DAPI at 1:20.000. Coverslips were mounted on microscopic slides using Dako fluorescence mounting medium.

Western Blot

HEK293 cells seeded into 6-well plates were transiently transfected for 48h with 2µg per well of either WT or mutant *pno*Cherry-FXR as described above and in SI. Membranes were blocked with 5% BSA in TBS-T for 1h before overnight incubation with rabbit anti-FXR (H-130; 1:2.000) and mouse anti-β-actin (ab6276, Abcam; 1:10.000) followed by incubation with goat anti-rabbit-IgG-AlexaFluor 647 and goat anti-mouse-IgG-AlexaFluor 488 (both at 1:5.000). Fluorescent signals were detected using a ChemiDoc MP imaging system (Biorad).

Luciferase Assay

Luciferase assays were performed using the Dual Luciferase reporter assay (Promega) according to the manufacturer´s instructions. Briefly, HEK293 cells kept in DMEM containing 10% fetal calf serum (FCS) were seeded onto 12-well plates at 150.000 cells per well and transfected the next morning with 1µg of the $BSEP^{prom}$-Luc or SHP^{prom}-Luc plasmid and 100ng each of FXR and RXR expression plasmids using Fugene HD (Promega) at a ratio of 2.5:1 (reagent:DNA). Cells transfected with FXRα1/2 expression plasmid were stimulated with 10µM OCA (INT-747), cells transfected with RXRα were stimulated with 1µM 9-cis-retinoic acid (9-cis-RA), cells transfected with both FXRα1/2 and RXRα expression plasmids were stimulated with both ligands.

RNA preparation, reverse transcription, pre-amplification, and PCR analysis

Total RNA was extracted and purified using the AmoyDx FFPE DNA/RNA Kit, (Amoy Diagnostics Co.) according to the manufacturer's instructions. 100ng of RNA was reverse transcribed using the High Capacity cDNA Reverse Transcription Kit (Applied Biosystems). After pre-amplification, qPCR was carried out with different TaqMan™ Gene Expression Assays (Applied Biosystems). Relative quantification of mRNA was performed according to the comparative $2^{-\Delta\Delta CT}$ method with SDHA as an endogenous control (see SI for detailed information).

Structure modeling and molecular dynamics simulations

To analyze the impact of the variant T296I, the ligand binding domain (LBD) structure of FXR (Q96RI1-1, residues 248 to 476) was modeled based on the chenodeoxycholic acid (CDCA)- and NCoA2 peptide-bound X-ray crystal structure of the FXR LBD (PDB ID 6HL1) (13), representing the active state of FXR, using SWISS-MODEL (20). To model the inactive state with H12 not interacting with the LBD core, the loop between helix11 and H12 was remodeled within PyMOL (Schrödinger, LLC, New York). In detail, residues 460 to 466 (^{460}VNDHKFT^{466}) were removed and readded, pointing away from the LBD core, followed by the α-helix H12.

MD simulations were performed for both the active and inactive states in the presence of the endogenous ligand CDCA and a short peptide sequence from the NCoA2 protein (sequence KENALLRYLLDKD), containing the signature motif LXXLL for binding to an NR (14). The structural models were prepared for molecular dynamics (MD) simulations using the AMBER21 package (21). Overall, four different systems were prepared: FXR wildtype in the active state (hereafter termed “active WT”), FXR T296I variant in the active state (“active T296I”), FXR wildtype in the inactive state (“inactive WT”), and FXR T296I variant in the inactive state (“inactive T296I”). Postprocessing and analysis of the MD trajectories were performed with CPPTRAJ (22) implemented in AmberTools21 (21). For further details, please see the SI Methods.

Statistical Analysis

Significance tests were performed using the Mann-Whitney U test or Student’s t-test if not indicated otherwise. The indicated significance levels are n.s. (not significant, $p > 0.05$), *: $p \leq 0.05$, **: $p \leq 0.01$, ***: $p \leq 0.001$, ****: $p \leq 0.0001$.

Results

The T296I variant is located within the LBD of FXR

The ligand binding domain (LBD) of FXR is critical in regulating the protein's activity. Residue 296 is located on helix 3 with its side chain facing toward the AF2 interaction surface formed partly by H12 (Fig. 1A and B). We thus hypothesized that variant T296I impacts FXR's ability to transition from the inactive to the active state. Accordingly, we investigated the effect of variant T296I in MD simulations starting from one of the four configurations: FXR WT in the active state ("active WT"), FXR T296I variant in the active state ("active T296I"), FXR WT in the inactive state ("inactive WT"), and FXR T296I variant in the inactive state ("inactive T296I") (Fig. 1C). All systems contained the LBD of FXR, the agonist chenodeoxycholic acid (CDCA) (23, 24), and a short peptide sequence of the nuclear receptor coactivator 2 (NCoA2). The inactive state was created from the active state through repositioning of the loop region between H11 and H12 such that H12 pointed away from the LBD core and had a distance > 45Å to it (distance in the active state 16Å). This setup allowed us to study if the substitution impacts the active state and/or the transition from the inactive to the active state.

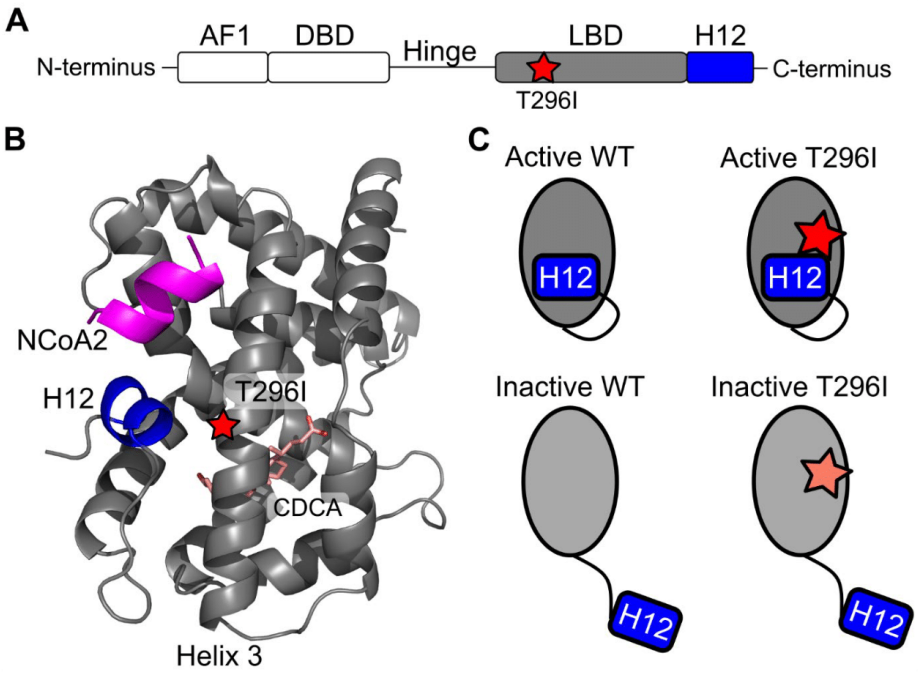


Fig. 1: Overview of the protein structure of FXR and the variant site within the LBD. A) Schematic of the domain arrangement of the FXR protein. The N-terminal activation function 1 (AF1) motif is followed by the DNA-binding domain (DBD), which is connected via a flexible hinge region to the LBD and the C-terminal H12. The variant T296I (red star) is located within the LBD. B) Protein structure of the LBD of FXR. The protein systems were modeled based on the crystal structure of agonist-bound FXR LBD (13) and used for MD simulations, containing additionally CDCA as ligand (pink, shown as sticks) and a short peptide of NCoA2 (magenta). H12 is highlighted in blue and H3, containing the variant site T296I (red star), is labeled. C) Overview of the four systems used as input to MD simulations to study the variant's impact on the active and the inactive state of the LBD. To differentiate between the different systems, we have consistently used the following color scheme: active WT in darker grey, inactive WT in lighter grey (corresponding to the depicted color of the LBD), active T296I in red and inactive T296I in faded red (corresponding to the depicted color of the star indicating the variant position).

T296I decreases the transcriptional activity of FXR

To determine the consequences of the FXR missense variant T296I on expression, subcellular localization and target gene induction, human FXRα1 and FXRα2 were cloned from human liver and co-transfected with RXRα into HEK293 cells. Both wildtype (WT) as well as the missense variant were detected within the nucleus of transfected cells (Fig. 2A). Furthermore, protein amounts as determined by western blotting were similar in WT and T296I transfected

cells (Fig. 2B). For functional analysis, we used a luciferase expression vector containing the BSEP promoter sequence (BSEP^{prom}-Luc), which was co-transfected with several combinations of RXRα and either FXRα1, FXRα2, FXRα1^{T296I}, or FXRα2^{T296I} and subsequently stimulated with an FXR ligand (OCA, 10μM) and an RXR ligand (9-cis-RA, 1μM). The highest BSEP transactivation was observed when both RXR and WT FXRα1/2 were co-transfected, represented as 100% luciferase activity (Fig. 2C). Transfection of FXRα1 or FXRα2 alone resulted in luciferase activity of 19.24% and 34.45% respectively, in comparison to co-transfection with both RXRα and either WT FXRα1/2. However, when the FXRα1/2^{T296I} variant was transfected alone, there was a significant decrease in luciferase activity to 11.72% and 4.89%, respectively, when compared to either WT FXRα1/2. Similarly, the co-transfection of RXRα with the FXRα1/2^{T296I} variant resulted in a significant reduction in luciferase activity, with decreases to 26.06% and to 13.14%, respectively, in comparison to the co-transfection of RXRα and either WT FXRα1/2. We obtained similar results by using the SHP promoter sequence in the same luciferase expression vector (SHP^{prom}-Luc) and under the same experimental conditions (Fig. S2). In summary, even though the subcellular localization and protein expression levels were unaffected, the presence of the T296I missense variant resulted in a substantial decrease in FXR target gene transactivation *in vitro*.

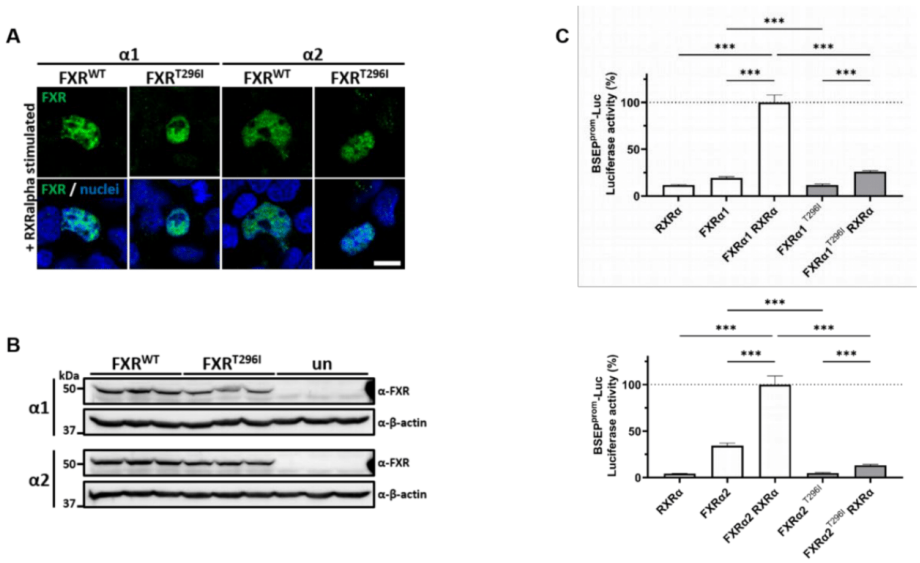


Fig. 2: FXR T296I reduces transcriptional activity of the BSEP promoter in transfected HEK293 cells. A) HEK293 transiently co-transfected with RXRα and either FXRα1, FXRα2, FXRα1^{T296I}, or FXRα2^{T296I} showed correct nuclear localization of the wildtype (WT) and mutant protein. Bar = 10µm. B) Western blot analysis revealed similar protein amounts in cells transfected with the different FXR cDNA constructs, un=untransfected controls. C) Analysis of luciferase enzymatic activity after transfection of HEK293 cells with a luciferase reporter gene downstream of the BSEP promoter (BSEP^{prom}-Luc) as well as different combinations of RXRα and either FXRα1, FXRα2, FXRα1^{T296I}, or FXRα2^{T296I}, as indicated on the x-axis. The plasmid pRL-TK was included in each transfection for normalization. Cells were stimulated with the FXR ligand OCA (10µM) and the RXR ligand 9-cis-RA (1µM). Values were obtained from three independent experiments, in which each condition was tested in duplicates. Values on the y-axis represent the mean and SD, expressed as % luciferase activity. The asterisks indicate a significant difference analyzed by a two-tailed Student t-test, *** = $p \leq 0.001$.

FXR T296I leads to an increased distance between H12 and the substitution site, indicative of a less favorable active state

Using MD, we investigated the molecular mechanisms underlying the decreased activity of the T296I variant. All simulation systems showed minor structural variability with respect to the binding of CDCA and NCoA2 and the FXR LBD structure up to and including H11 (Fig. S1).

Based on the crystal structure of the FXR LBD (13), representing the agonist-bound active state, the WT T296 likely interacts with T466 preceding H12 (Fig. 3A). Accordingly, we measured the distance between residue T296 and T466 during MD simulations and compared it to the reference distance in the crystal structure. For the active states, the variant showed an increase in the distance (Fig. 3B). We determined the frequency of occurrence when the reference cutoff distance was reached (Fig. 3C, Table S1). Across the 15 replicas, the active WT system was found in approx. 27% of the time in the active state according to distances below the reference distance. The frequency of occurrence was significantly lower for active T296I (0.40%). The inactive systems showed initially high distance values, as expected. Inactive WT reached below the reference distance in 6 out of 15 replicas and often stayed within this active state for the remainder of the simulation time, indicating that the active state is the preferred one under the simulation conditions (Fig. S3, Table S1). Inactive T296I only reached the reference distance in one replica (Fig. S3, Table S1) and, accordingly, the frequency of reaching the reference value was significantly reduced in the inactive T296I system compared to the inactive WT system (Fig. 3C).

Overall, we observed an increase in the distance between residue T296 and T466 for the T296I variant in the active state. In line, for systems started either from the active or inactive state, T296I led to a significant decrease in the frequency of occurrence in reaching the reference distance compared to the WT. This data indicates that the active state is destabilized in the variant.

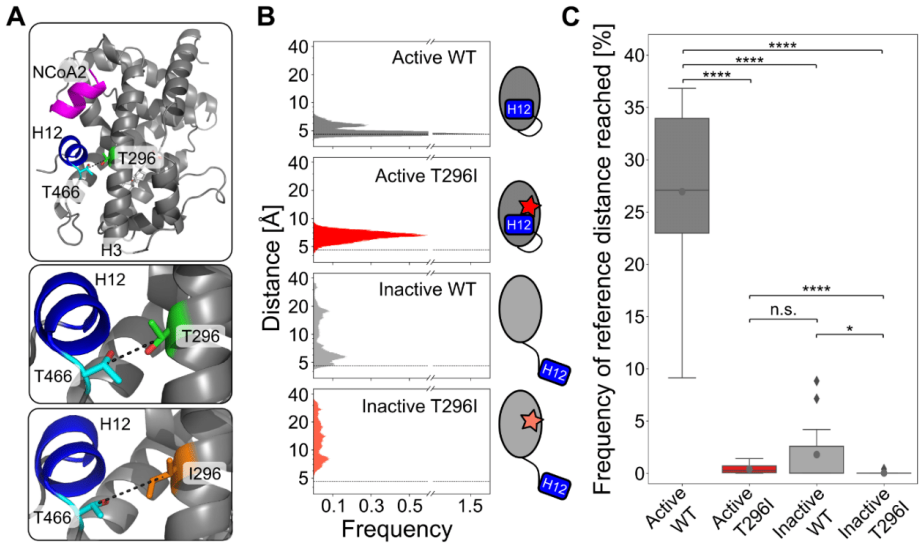


Fig. 3: Distance analysis between residues T296 and T466 over MD simulation time. A) Depiction of the distance measured within the LBD of FXR between the C_β atoms of residue T296 and T466. The mean distance is increased in the active T296I (6.6Å) compared to the active WT (5.0Å). B) Histograms of measured distances over all 15 MD runs (see Fig. S3) for each analyzed state. The reference distance (4.6Å) as measured in the agonist-bound crystal structure (13) is depicted as dashed lines. C) Frequency of occurrence that the respective system is in the active state. For each replica (Fig. S3), the percentage of reaching a distance below the reference distance (4.6Å) is depicted as a boxplot. Individual values are shown in Table S1. Boxes depict the quartiles of the data with the median (straight black lines) and mean (grey dots) indicated; the whiskers indicate the minimum and the maximal values, outlier points are depicted as rhombus. Differences in the mean values were statistically evaluated using a two-sided Mann-Whitney U test (N = 15, n.s.: not significant; *: $p \leq 0.05$, ****: $p \leq 0.0001$).

The correct positioning of H12 in the active state is reduced in the T296I variant

As the correct positioning of H12 within NRs is crucial for activity (25-28), we visually analyzed the simulation trajectories of the inactive systems and exemplarily show the conformational change from the inactive to the active state for one out of several MD replicas of the inactive WT system showing this transitioning (Fig. 4A and Movie S1). For the inactive T296I, we show the MD replica where H12 positioning was closest to the conformation in the active state (Fig. 4B and Movie S2). Further, we analyzed the impact of T296I on the positioning of H12 using the root mean square deviation (RMSD) of residues 466 to 473 (H12 and preceding T466) over the MD simulation time (Fig. 4C and D). As a reference state, we employed the active

conformation based on the crystal structure of the agonist-bound FXR LBD (13). For the active WT, a skewed Gaussian function was fitted to the RMSD histogram, yielding a mean of 1.9Å (Fig. S4 and Fig. 4C). Over the entire MD simulations, more than half of the time (~56%, Table S2 and Fig. 4D) the active WT had an RMSD below this mean value, which we used as a further reference to indicate reaching the active state. Comparing the active systems revealed a significant shift in the distributions, with a larger mean of active T296I indicating a higher deviation from the active positioning (active WT mean: 1.9Å, active T296I mean: 2.6Å, p = 0.0022, two-sided t-test, Fig. S4). The frequency of reaching the active state was significantly lower for active T296I, inactive WT, and inactive T296I systems compared to active WT (Fig. 4D). While the inactive WT reached the reference RMSD value in four out of 15 replicas, the inactive T296I reached it in one replica (Fig. S5). Furthermore, while the inactive T296I did not stay in the active state long (frequency: 0.01% in replica no. 6), the inactive WT – once reaching the active state – showed often prolonged persistence times (frequency: 15.40% in replica no. 2, 0.26% in replica no. 6, 10.17% in replica no. 8, 14.29% in replica no. 13) (Fig. S5 and Table S2). The comparison between inactive WT and inactive T296I indicated a similar trend as observed for the distance analysis but did not reach the significance level (Fig. 4D).

Overall, H12 positioning is significantly structurally deviating with respect to the reference active state for all systems compared to the active WT. While the active T296I could reach the reference cutoff, it did so for a significantly decreased amount of time compared to the active WT, again indicating that the active conformation is less favorable in the variant. Although inactive WT and inactive T296I could both reach the cutoff, the inactive WT reached it more frequently and for a longer time. However, the differences to inactive T296I are not significant. Our data indicate that unbiased MD simulations on the μs-scale can sample the transition from the inactive to the active state (see Table 1) and that this conformational change is less frequent in the variant.

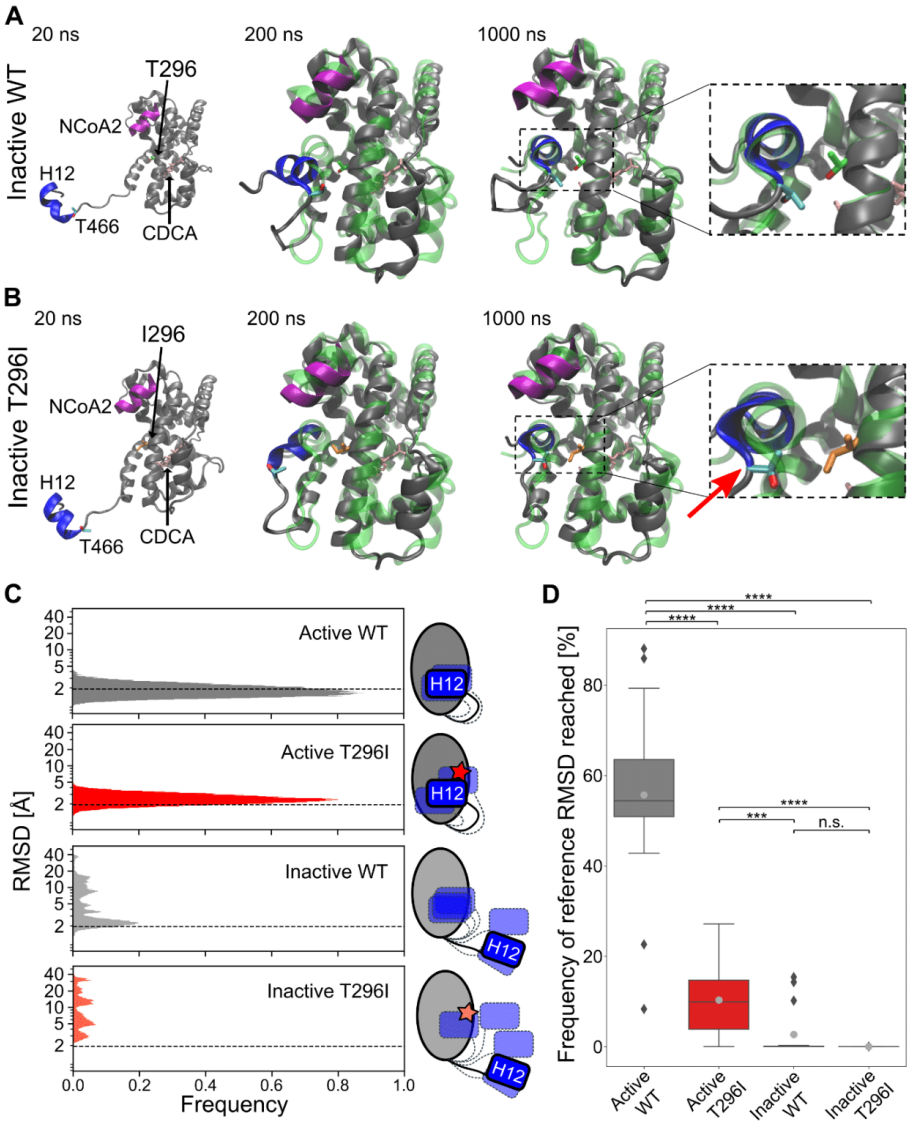


Fig. 4: Positioning of H12 during MD simulations. A) Conformational transitioning of the inactive WT (replica no. 2, Fig. S3 and S5) over the MD simulation time. The initial active state (based on the crystal structure of agonist-bound FXR LBD (13) is depicted as a green, translucent cartoon structure. B) Transitioning of the inactive T296I (replica no. 6, Fig. S3 and S5) over the MD simulation time. The initial active state is depicted as a green, translucent cartoon structure. Important residues (variant site 296 [green or orange, shown as sticks], T466 [cyan, shown as sticks], H12 [blue], NCoA2 [magenta], and bound ligand CDCA [pink, shown as sticks]) are additionally depicted in A and B. After 1000ns, H12 of the inactive WT almost perfectly overlayed with that of the active structure, while H12 of the inactive T296I showed

structural deviations (red arrow). C) Histograms of RMSD values of H12 and the preceding T466 over all 15 MD replicas for each analyzed state. Data for each MD replica are depicted in Fig. S5. A skewed Gaussian function was fitted to the distribution of active WT (Fig. S4), and the obtained mean (1.9Å) was used as a reference cutoff (dashed lines). D) Frequency of occurrence a system spends in the active state, i.e., when the reference cutoff is reached. For each replica (Fig. S5), the frequency of occurrence was determined and depicted within boxplots. Individual values are shown in Table S2. Boxes depict the quartiles of the data with the median (straight black lines) and mean (grey dots) indicated; the whiskers depict the minimum and the maximum values, outlier points are depicted as rhombus. Differences in the mean values were statistically evaluated using a two-sided Mann-Whitney U test (N = 15, n.s.: not significant; *: ***: $p \leq 0.001$, ****: $p \leq 0.0001$).

Table 1: Overview of the MD simulation results.

	Distance criteria	RMSD criteria
Inactive WT reaching active state	6 out of 15 runs	4 out of 15 runs
Inactive T296I reaching active state	1 out of 15 runs	1 out of 15 runs
Significance between WT and T296I[a]	* (p = 0.016)	n.s. (p = 0.076)

[a] Using one-sided t-test.

The T296I variant is associated with reduced expression of FXR target genes

To investigate the mRNA expression of FXR and two of its targets (BSEP, SHP) in the PFIC5 patient carrying the T296I variant, we performed qPCR analysis using FFPE samples from the patient´s liver taken at the time of transplantation. Additionally, FFPE-liver samples from two cirrhotic adult patients and a healthy adult control were included in the analysis. To address the putatively low RNA integrity after isolation from FFPE samples, we employed TaqMan Gene Expression Assays targeting different regions of the FXR and BSEP transcripts. While FXR mRNA expression was similar between the patient and control and the two cirrhosis livers (Fig. 5A), BSEP mRNA and SHP mRNA expression was strongly reduced in comparison to the healthy control but also the two cirrhotic liver samples (reduction to 3.03, 1.82 and 10.91% of healthy control for BSEP Taq1, BSEP Taq2, and SHP, respectively) (Fig. 5B). These

findings further demonstrate that our *in vitro* and *in silico* data authentically reflect the impaired transcriptional target gene activation by the FXR T296I variant.

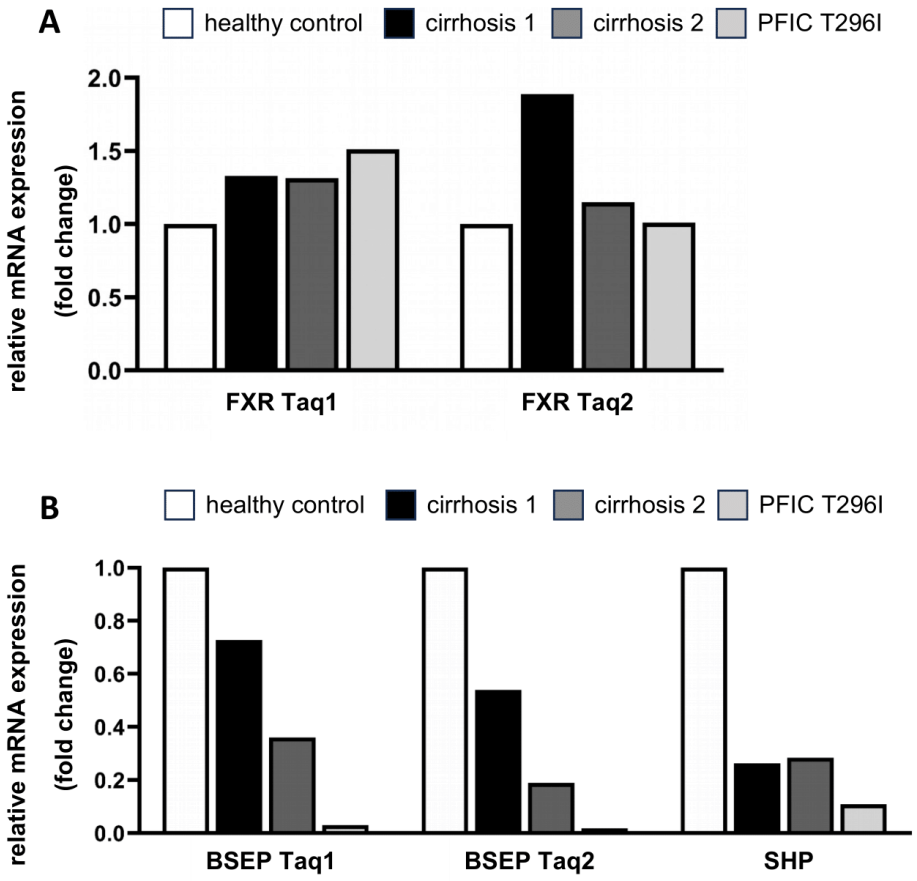


Fig. 5: FXR, BSEP, and SHP expression in the patient's liver tissue. A) Relative mRNA expression of FXR using two different TaqMan probes (Taq1 and Taq2) showed similar levels in the liver of a healthy control (white bars), two samples from cirrhotic livers (black bar and dark grey bar) as well as the patient (light grey bar). B) FXR target gene expression of BSEP and SHP were lower in the two cirrhotic livers as compared to the healthy control but were further reduced in the sample from the T296I PFIC patient.

Discussion

In this study, we combined *in vitro* experiments with computational studies to analyze the impact of the PFIC5-associated NR1H4 T296I variant on FXR protein expression, subcellular localization, and function. While the introduction of the missense variant into the α1 or α2 FXR isoform did not affect protein expression and nuclear localization *in vitro*, it significantly reduced activation of the FXR target genes BSEP and SHP. A strong reduction of BSEP and SHP was also observed in the patient's liver at the time of transplantation. Using our computational approach, we elucidated a detailed mechanism for the effect of the variant on the conformational transition of the LBD from the inactive to the active state. The variant showed a significantly reduced tendency to reach the active state, which can explain the *in vitro*-identified decreased target gene expression and thus the PFIC phenotype of the patient (9).

We describe conformational changes from an inactive to the known active state of the FXR LBD in unbiased MD simulations. To drive the system towards the active state, we used a coactivator peptide and the most potent *in vivo* agonist CDCA (29) within the MD systems as both coactivator peptide and ligand binding have been shown to induce and stabilize the active state *in vitro* (13). Depending on the analysis, the inactive WT system reached the active state in 27% or 40% (4/15 replicas for H12 RMSD analysis and 6/15 replicas for distance analysis, respectively) of the simulations (Table 1). A dynamic movement from the inactive to the active state (and potentially reverse, at least in ligand-free states) may occur in the nanosecond time scale as indicated by time-resolved fluorescence anisotropy decay studies on PPARγ (27). Chrisman et al. showed conformational changes of the H12 within the NR PPARγ towards an almost-active state within the µs to ms time scale range in unbiased MD simulations (30). Within our MD simulations, the LBD of FXR might sample conformational spaces usually not available due to sterical hindrances either by the not considered parts of FXR (DNA binding domain and linker sequences) or due to hetero-dimer binding partners. Thus, it is not surprising that in several replicas, H12 did not move into the active position within the 1µs of simulation time due to being trapped in other energy minimas. Still, fast transitioning from the inactive to the active state was observable in some replicas. Comparing the difference of inactive WT and

inactive T296I systems in reaching the active state, we observed a decrease by a factor of 3.8 to 5.7 (inactive WT: 26.7% or 40%; inactive T296I: 7%). This is in good agreement with the transcriptional activity reduction of T296I compared to WT as shown in the luciferase assay for BSEP as well as the patient's liver tissue (Fig. 2, Fig. 5).

Overall, the variant T296I, while also impacting the active state, likely exerts its negative impact on protein activity due to a change in the structural dynamics of the inactive-to-active state transition. Our results indicate that the T296I protein does not reach the active state fully and less frequently compared to the WT protein.

Furthermore, from the analysis of the MD simulations, insights into the activation mechanism of the LBD were gained. The presence of the ligand and co-activation peptide allows FXR to switch into an active state and stably stay within this state. This is in line with previous NMR or MD studies in RORγ (11), PPARγ (30, 31), and FXR or FXR/RXR heterodimers (16-18). By contrast, the transition from inactive to active FXR has previously not been observed in MD simulations. Our setup of the MD simulations can be useful to predict the impact of other missense variants on FXR function and potentially strengthen studies on FXR targeting, enabling detailed evaluations of the molecular mechanism of drugs based on their impact on the activation transitioning.

References

1. Bull LN, Thompson RJ. Progressive Familial Intrahepatic Cholestasis. Clin Liver Dis 2018;22:657-669.
2. Felzen A, Verkade HJ. The spectrum of Progressive Familial Intrahepatic Cholestasis diseases: Update on pathophysiology and emerging treatments. Eur J Med Genet 2021;64:104317.
3. Sinal CJ, Tohkin M, Miyata M, Ward JM, Lambert G, Gonzalez FJ. Targeted disruption of the nuclear receptor FXR/BAR impairs bile acid and lipid homeostasis. Cell 2000;102:731-744.
4. Cariello M, Piccinin E, Garcia-Irigoyen O, Sabba C, Moschetta A. Nuclear receptor FXR, bile acids and liver damage: Introducing the progressive familial intrahepatic cholestasis with FXR mutations. Biochim Biophys Acta Mol Basis Dis 2018;1864:1308-1318.
5. Gomez-Ospina N, Potter CJ, Xiao R, Manickam K, Kim MS, Kim KH, Shneider BL, et al. Mutations in the nuclear bile acid receptor FXR cause progressive familial intrahepatic cholestasis. Nat Commun 2016;7:10713.
6. Himes RW, Mojarrad M, Eslahi A, Finegold MJ, Maroofian R, Moore DD. NR1H4-related Progressive Familial Intrahepatic Cholestasis 5: Further Evidence for Rapidly Progressive Liver Failure. J Pediatr Gastroenterol Nutr 2020;70:e111-e113.
7. Mehta S, Kumar K, Bhardwaj R, Malhotra S, Goyal N, Sibal A. Progressive Familial Intrahepatic Cholestasis: A Study in Children From a Liver Transplant Center in India. J Clin Exp Hepatol 2022;12:454-460.
8. Czubkowski P, Thompson RJ, Jankowska I, Knisely AS, Finegold M, Parsons P, Cielecka-Kuszyk J, et al. Progressive familial intrahepatic cholestasis - farnesoid X receptor deficiency due to NR1H4 mutation: A case report. World J Clin Cases 2021;9:3631-3636.
9. Pfister ED, Dröge C, Liebe R, Stalke A, Buhl N, Ballauff A, Cantz T, et al. Extrahepatic manifestations of progressive familial intrahepatic cholestasis syndromes: presentation of a case series and literature review. Liver International 2022.
10. Hollingsworth SA, Dror RO. Molecular Dynamics Simulation for All. Neuron 2018;99:1129-1143.
11. Saen-Oon S, Lozoya E, Segarra V, Guallar V, Soliva R. Atomistic simulations shed new light on the activation mechanisms of RORγ and classify it as Type III nuclear hormone receptor regarding ligand-binding paths. Scientific Reports 2019;9:17249.
12. Chrisman IM, Nemetchek MD, de Vera IMS, Shang J, Heidari Z, Long Y, Reyes-Caballero H, et al. Defining a conformational ensemble that directs activation of PPARgamma. Nat Commun 2018;9:1794.
13. Merk D, Sreeramulu S, Kudlinzki D, Saxena K, Linhard V, Gande SL, Hiller F, et al. Molecular tuning of farnesoid X receptor partial agonism. Nat Commun 2019;10:2915.
14. Heery DM, Kalkhoven E, Hoare S, Parker MG. A signature motif in transcriptional co-activators mediates binding to nuclear receptors. Nature 1997;387:733-736.
15. Xu HE, Stanley TB, Montana VG, Lambert MH, Shearer BG, Cobb JE, McKee DD, et al. Structural basis for antagonist-mediated recruitment of nuclear co-repressors by PPARalpha. Nature 2002;415:813-817.
16. Kumari A, Mittal L, Srivastava M, Pathak DP, Asthana S. Conformational Characterization of the Co-Activator Binding Site Revealed the Mechanism to Achieve the Bioactive State of FXR. Front Mol Biosci 2021;8:658312.
17. Kumari A, Mittal L, Srivastava M, Pathak DP, Asthana S. Deciphering the Structural Determinants Critical in Attaining the FXR Partial Agonism. J Phys Chem B 2023;127:465-485.

18. Díaz-Holguín A, Rashidian A, Pijnenburg D, Monteiro Ferreira G, Stefela A, Kaspar M, Kudova E, et al. When Two Become One: Conformational Changes in FXR/RXR Heterodimers Bound to Steroidal Antagonists. ChemMedChem 2023;18:e202200556.
19. Lee HK, Lee YK, Park SH, Kim YS, Park SH, Lee JW, Kwon HB, et al. Structure and expression of the orphan nuclear receptor SHP gene. J Biol Chem 1998;273:14398-14402.
20. Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, Heer FT, et al. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res 2018;46:W296-W303.
21. Case DA, Aktulga HM, Belfon K, Ben-Shalom IY, Borzell SR, Cerutti DS, Cheatham TE, 3rd, et al. Amber. In. University of California, San Francisco; 2021.
22. Roe DR, Cheatham TE, III. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. Journal of Chemical Theory and Computation 2013;9:3084-3095.
23. Makishima M, Okamoto AY, Repa JJ, Tu H, Learned RM, Luk A, Hull MV, et al. Identification of a nuclear receptor for bile acids. Science 1999;284:1362-1365.
24. Parks DJ, Blanchard SG, Bledsoe RK, Chandra G, Consler TG, Kliewer SA, Stimmel JB, et al. Bile acids: natural ligands for an orphan nuclear receptor. Science 1999;284:1365-1368.
25. Mi LZ, Devarakonda S, Harp JM, Han Q, Pellicciari R, Willson TM, Khorasanizadeh S, et al. Structural basis for bile acid binding and activation of the nuclear receptor FXR. Mol Cell 2003;11:1093-1100.
26. Renaud JP, Rochel N, Ruff M, Vivat V, Chambon P, Gronemeyer H, Moras D. Crystal structure of the RAR-gamma ligand-binding domain bound to all-trans retinoic acid. Nature 1995;378:681-689.
27. Kallenberger BC, Love JD, Chatterjee VK, Schwabe JW. A dynamic mechanism of nuclear receptor activation and its perturbation in a human disease. Nat Struct Biol 2003;10:136-140.
28. Nolte RT, Wisely GB, Westin S, Cobb JE, Lambert MH, Kurokawa R, Rosenfeld MG, et al. Ligand binding and co-activator assembly of the peroxisome proliferator-activated receptor-γ. Nature 1998;395:137-143.
29. Wang H, Chen J, Hollister K, Sowers LC, Forman BM. Endogenous bile acids are ligands for the nuclear receptor FXR/BAR. Mol Cell 1999;3:543-553.
30. Chrisman IM, Nemetchek MD, de Vera IMS, Shang J, Heidari Z, Long Y, Reyes-Caballero H, et al. Defining a conformational ensemble that directs activation of PPARγ. Nature Communications 2018;9:1794.
31. Heidari Z, Chrisman IM, Nemetchek MD, Novick SJ, Blayo A-L, Patton T, Mendes DE, et al. Definition of functionally and structurally distinct repressive states in the nuclear receptor PPARγ. Nature Communications 2019;10:5825.

Impaired transitioning of the FXR ligand binding domain to an active state underlies a PFIC5 phenotype

Annika Behrendt[1], Jan Stindt[2], Eva-Doreen Pfister[3], Kathrin Grau[1], Stefanie Brands[1], Alex Bastianelli[4], Carola Dröge[2,4], Amelie Stalke[5], Michele Bonus[1], Malte Sgodda[6], Tobias Cantz[6,7], Sabine Franke[8], Ulrich Baumann[3], Verena Keitel[2,4]* and Holger Gohlke[1,9]*

[1] Institute for Pharmaceutical and Medicinal Chemistry, Heinrich Heine University, Düsseldorf, Germany

[2] Department of Gastroenterology, Hepatology and Infectious Diseases, Medical Faculty University Hospital Düsseldorf, Heinrich Heine University, Düsseldorf, Germany

[3] Pediatric Gastroenterology and Hepatology, Department for Pediatric Kidney, Liver and Metabolic Diseases, Hannover Medical School, Hannover, Germany

[4] Department of Gastroenterology, Hepatology and Infectious Diseases, Medical Faculty, University Hospital Magdeburg, Otto von Guericke University, Magdeburg, Germany

[5] Department of Human Genetics, Hannover Medical School, Hannover, Germany

[6] Research Group Translational Hepatology and Stem Cell Biology, Department of Gastroenterology, Hepatology, Infectious Diseases and Endocrinology, Hannover Medical School, Hannover, Germany

[7] REBIRTH-Research Center for Translational Regenerative Medicine, Hannover Medical School, Hannover, Germany

[8] Institute of Pathology, University Hospital Magdeburg, Otto von Guericke University, Magdeburg, Germany

[9] John-von-Neumann-Institute for Computing, Jülich Supercomputing Center, Institute of Biological Information Processing (IBI-7: Structural Biochemistry), and Institute of Bio- and Geosciences (IBG-4: Bioinformatics), Forschungszentrum Jülich GmbH, Jülich, Germany

*corresponding authors

Supplementary Results

Structural variability during the MD simulations

The structural variability of the systems was analyzed by root mean square deviations (RMSD) with respect to the first production frame over the MD trajectories (Fig. S1). The results indicate that the CDCA ligand remains in its initial binding mode (Fig. S1A, B) and show minor structural differences between the active T296I and the inactive WT system (Fig. S1C-E). The LBD of FXR, excluding the flexible region of helix 12 (H12) and the preceding loop, showed a constant RMSD over the simulation time with no significant differences between the four systems (Fig. S1C), further indicating that the LBD without H12 did not undergo significant conformational changes over the simulation time. Expectedly, the inactive WT and inactive T296I systems showed higher structural variability of H12 and the preceding loop region compared to the active systems, indicating higher mobility of this part of FXR in the inactive systems. Furthermore, in the active T296I, the mobility of H12 and the preceding loop region is significantly higher than in the active WT system (Fig. S1D), in line with further analyses of H12 mobility (see Fig. 4). The mobility of the NCoA2 peptide is similarly low in all systems, indicated by generally low RMSD values. Larger values, especially visible in the inactive systems, are indicative of higher mobility, and the displacement of the peptide was also visually observed in several replicas. Except for the comparison of active T296I to inactive T296I, which reveals a significantly increased mobility, the differences between the systems were not significant.

FXR T296I reduces transcriptional activity of the SHP promoter

For further functional analysis, we used a luciferase expression vector containing the SHP promoter sequence (SHP^{prom}-Luc), which was co-transfected with several combinations of RXRα and either FXR1α, FXRα2, $FXRα1^{T296I}$, or $FXRα2^{T296I}$. The highest SHP transactivation was observed when both RXR and either WT FXRα1/2 were co-transfected, which we represented as 100% luciferase activity (Fig. S2). Transfection of FXRα1 or FXRα2 alone resulted in luciferase activity of 17.72% and 18.46%, respectively, in comparison to co-transfection with both RXRα and WT FXRα1/2. However, when the $FXRα1/2^{T296I}$ variant was transfected alone, there was a significant decrease in luciferase activity to 7.39% and 9.53%, respectively, when compared to WT FXRα1/2. Similarly, the co-transfection of RXRα with the $FXRα1/2^{T296I}$ variant led to a significant reduction in luciferase activity, with decreases to 43.65% and 71.67%, respectively, in comparison to the co-transfection of RXRα and WT

FXRα1/2. In conclusion, the presence of T296I resulted in a substantial decrease in SHP transactivation in the transfected cells.

Geometric analyses of the MD trajectories

The distance between residue T296, the mutation site, and the threonine preceding H12, T466, was measured over the MD simulation time (Fig. S3). The distance was measured between the C_β atoms of both residues to avoid biases due to rotations of the side chains. For each replica, the time the system has a distance below the reference distance cutoff (taken from the initial active structure) was calculated (Table S1).

The structural variability of H12 was measured using the RMSD of all atoms of the H12 residues and the preceding T466 (^{466}TPLLCEIW^{473}). Beforehand, the least mobile part of the FXR LBD was identified, and the trajectories were fitted to this core. As a reference, the initial active WT system was used. The RMSD was determined over the entire MD simulations time; a histogram of the values revealed a skewed Gaussian curve for the active WT and active T296I system (Fig. S4). Since the distribution of active WT can be expected in a physiological, uninhibited, active system, the mean of a fitted skewed Gaussian curve was used as a reference for the active state. The histogram of the active T296I is significantly shifted (assessed by a two-sided *t*-test) towards higher RMSD values compared to the active WT, indicating increased mobility of H12. The RMSD values over the MD trajectories of the individual replicas are depicted in Fig. S5. The inactive WT reached the reference RMSD value of the active WT in four out of 15 replicas, whereas the inactive T296I only reached the reference in one out of 15 replicas. For each replica, the time the system showed an RMSD below the reference RMSD cutoff (mean of the fitted skewed Gaussian curve on active WT) was calculated (Table S2).

Melting temperature measurement of WT and variant protein in the absence or presence of ligand

To exclude the possibility that the variant‘s decreased protein activity identified in luciferase assays is due to changed ligand binding, we measured the thermostability of the FXR protein in the presence and absence of the agonist OCA (INT-747). FXR WT and FXR T296I variant proteins were expressed in *E. coli* with a His-tag and SUMO-tag to aid in purification and solubility (Fig. S6). In the absence of ligand, both FXR WT and variant FXR T296I had comparable melting temperatures, indicating that the protein structure is not significantly changed due to the amino acid substitution (Fig. S7 and Fig. S8, Welch’s *t*-test WT+DMSO vs. T296I+DMSO: n.s.). As expected, the presence of the ligand induced a shift in the melting

temperature curve and led to a significantly lower melting temperature for the WT FXR protein by 2.32+/-0.70°C (Welch's t-test WT+DMSO vs. WT+OCA: *). For the variant protein FXR T296I, the presence of the ligand induced a similar shift by 2.78+/-0.49°C (Welch's t-test T296I+DMSO vs. T296I+OCA: ***) and resulted in a lower melting temperature, comparable to the FXR WT in the presence of ligand (Welch's t-test WT+OCA vs. T296I+OCA: n.s.). Accordingly, agonist OCA binding to the protein was not disturbed by the variant, in line with previous studies on the NR retinoic-acid related-orphan-receptor-C (RORγ) suggesting that the ligand entry and exit pathway occurs via the so-called "backdoor" pathway and, thus, away from the variant site (1). Of note, other NRs such as estrogen receptor, androgen receptor, and glucocorticoid receptor can be classified by a different ligand entry via a Helix 3/Helix 7/Helix 11 interface (2), while FXR has been grouped with RORγ and PPARγ (1, 3). Similarly, unbinding MD simulation studies on the FXR LBD indicated egress pathways in line with the backdoor pathway, facing away from the Helix 12 and the variant site, as most favorable for the agonistic ligand GW4064 (4).

Visualization of the conformational change from the inactive to the active state

The trajectories of the inactive WT replica no. 2 (Movie S1) and the inactive T296I replica no. 6 (Movie S2) were chosen as representative trajectories of the respective system reaching the active state. In the case of inactive T296I, the system moved close to the active state but did not entirely reach it and/or showed higher mobility there. The reference state of the initial active WT is repeatedly shown in the movie as a green translucent representation to aid in the judgment of H12 positioning.

Supplementary Figures

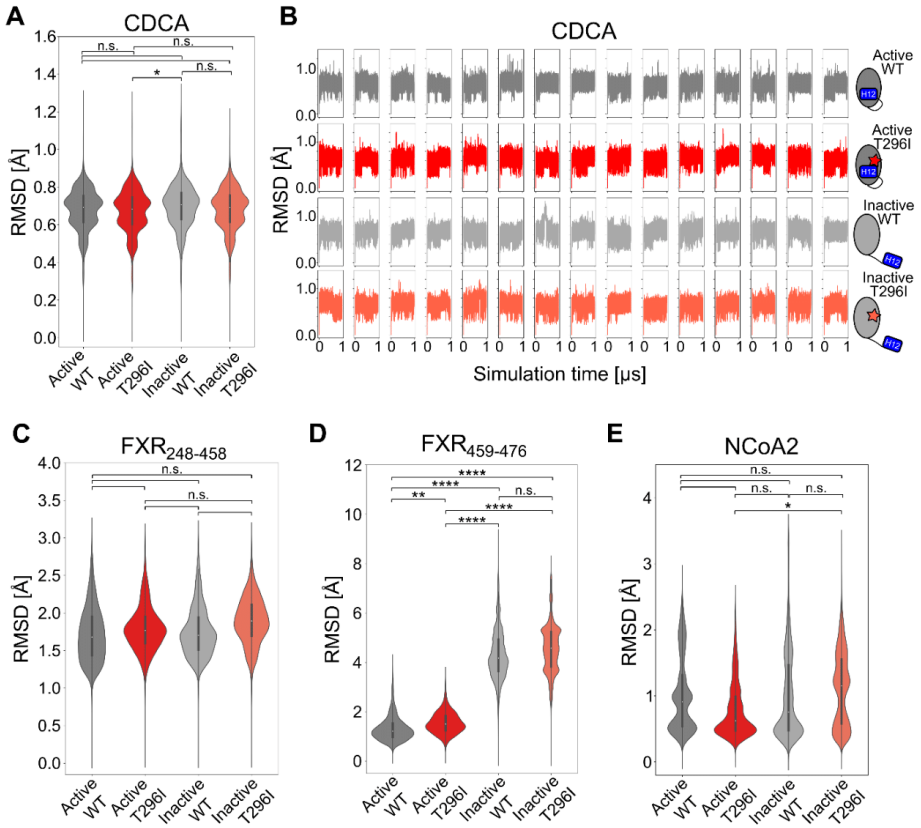


Fig. S1: RMSD over MD simulation time of ligand, protein, and co-activation peptide. (A) RMSD analysis of CDCA ligand (all atoms) over the MD simulation time (pooled over 15 replicas). (B) RMSD of CDCA ligand (all atoms) over the MD simulation time, shown for each of the 15 replicas. (C) RMSD of FXR LBD (residues 248-458) (C_α atoms), excluding the loop between helix (H) 11 and H12 and H12 itself (pooled over 15 replicas). (D) RMSD of FXR (residues 459-476) (C_α atoms), corresponding to the loop between helix 11 and helix 12 and helix 12 itself, over the MD simulation time (pooled over 15 replicas). (E) RMSD of NCoA2 co-activation peptide (C_α atoms) over the MD simulation time (pooled over 15 replicas). Significance tests were performed based on the means of the 15 replicas, respectively, using the Mann-Whitney U test. Violin plots were plotted using the Seaborn library (5), with the 1st to 3rd quartile within the box and the median marked as a white dot.

5

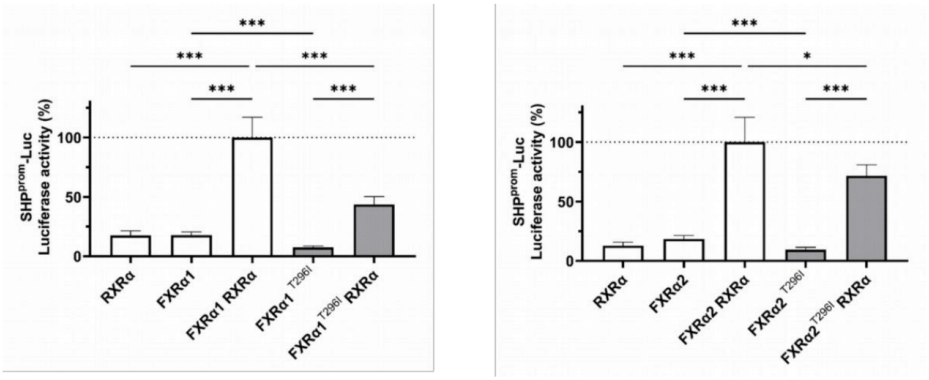


Fig. S2: FXR T296I reduces transcriptional activity of the SHP promoter in transfected HEK293 cells. Analysis of the luciferase enzymatic activity after transfection of HEK293 cells transfected with a luciferase reporter gene downstream of the SHP promotor (SHP^{prom}-Luc) as well as different combinations of RXRα and either FXRα1, FXRα2, FXRα1T296I, or FXRα2T296I, as indicated on the x-axis. The plasmid pRL-TK was included in each transfection for normalization. Cells were stimulated with an FXR and RXR ligand (OCA, 10µM and 9-cis-RA, 1µM). Values were obtained from six independent experiments, in which each condition was tested in duplicate. Values on the y-axis represent the mean and SD, expressed as % luciferase activity. The asterisks indicate a significant difference analyzed by a two-tailed Student t-test, * = $p \leq 0.05$, *** = $p \leq 0.001$.

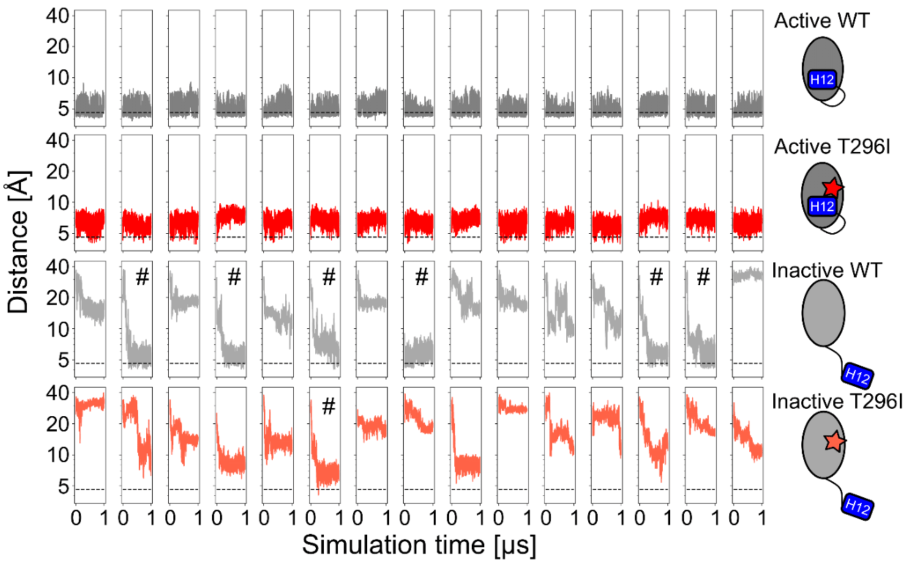


Fig. S3: Distance between C_{β} atoms of residues 296 and T466 over the MD simulations time. The distances over the 15 independent replicas for the four different systems of FXR LBD are shown. The reference distance cutoff of 4.6Å, based on the crystal structure of the agonist-bound FXR LBD (6), is shown as a dashed line. For the inactive WT and inactive T296I systems, replicas are marked (#) where the reference distance cutoff was reached. Histograms and frequencies of occurrence shown in Fig. 3 were calculated based on this data.

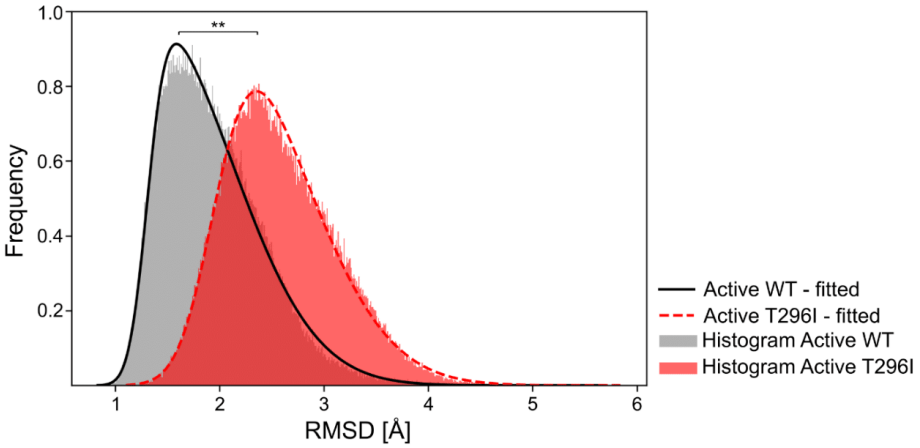


Fig. S4: Skewed Gaussian fits on histograms of RMSD values of H12. RMSD values of H12 and preceding T466 residue over 15 replicas of MD simulations with respect to the crystal structure of the agonist-bound FXR LBD are shown for active WT (grey) and active T296I (red). Using the scipy.stats module (7), skewed Gaussian functions were fitted (black and red lines). The fitted distributions are significantly different (assessed by a two-sided *t*-test based on the mean (active WT: 1.9Å, active T296I: 2.6Å) and the standard deviation (active WT: 0.5Å, active T296I: 0.5Å). The mean of the active WT was further used as a reference cutoff value for the expected RMSD fluctuations of H12 to the initial active structure. **: *p*-value = 0.0022.

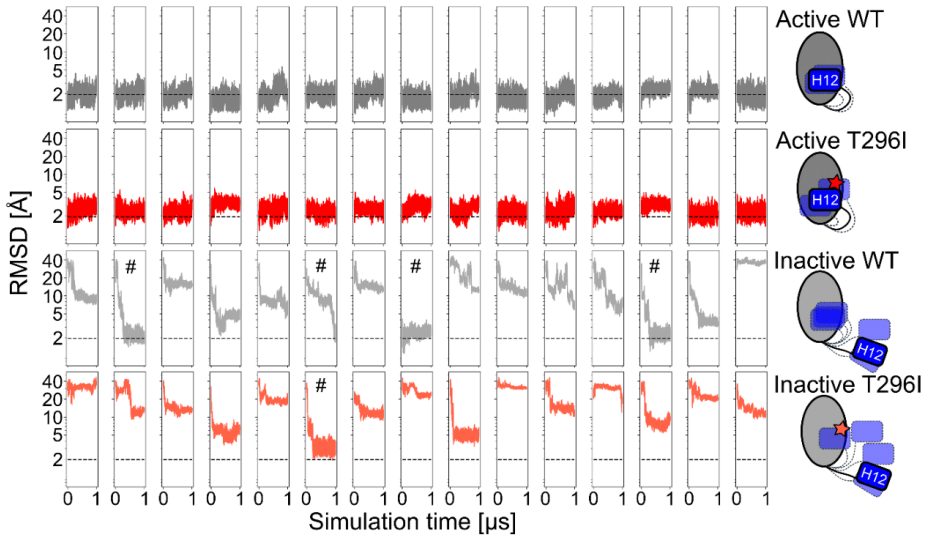


Fig. S5: RMSD values of H12 over the MD simulation time. The initial active WT structure, based on the crystal structure of agonist-bound FXR LBD (6), was used as a reference state for the RMSD analysis. The RMSD over the 15 independent replicas for the four different systems of FXR LBD is shown. The reference RMSD cutoff of 1.9Å, based on the mean of the fitted skewed Gaussian function for active WT values (Fig. S3), is shown as dashed lines. For the inactive WT and inactive T296I systems, replicas are marked (#) where the reference RMSD cutoff was reached. Histograms and frequencies of occurrence shown in Fig. 4 were calculated based on this data.

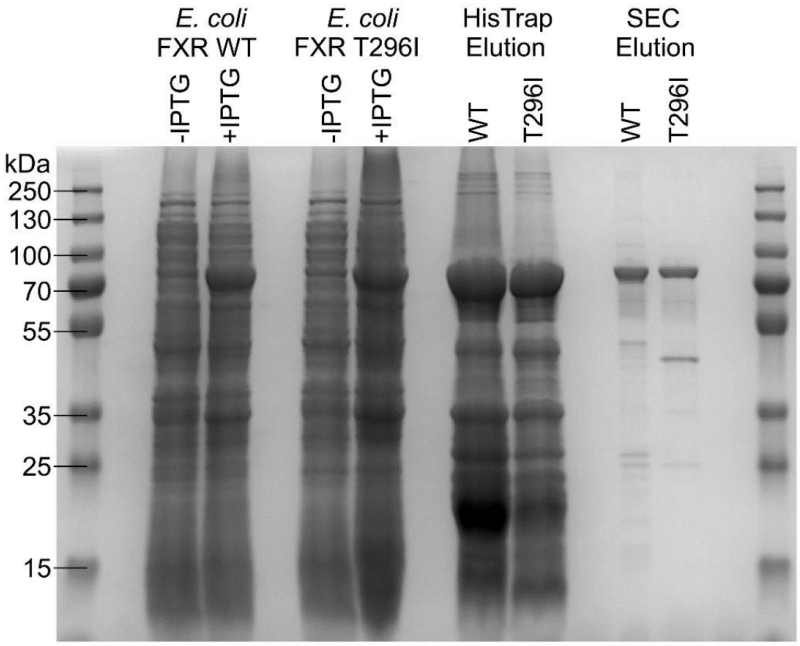


Fig. S6: Coomassie staining of FXR WT and FXR T296I protein expression and purification. The molecular weight of FXR protein with the SUMO and His-tag is ~70kDa. SEC: Size exclusion chromatography.

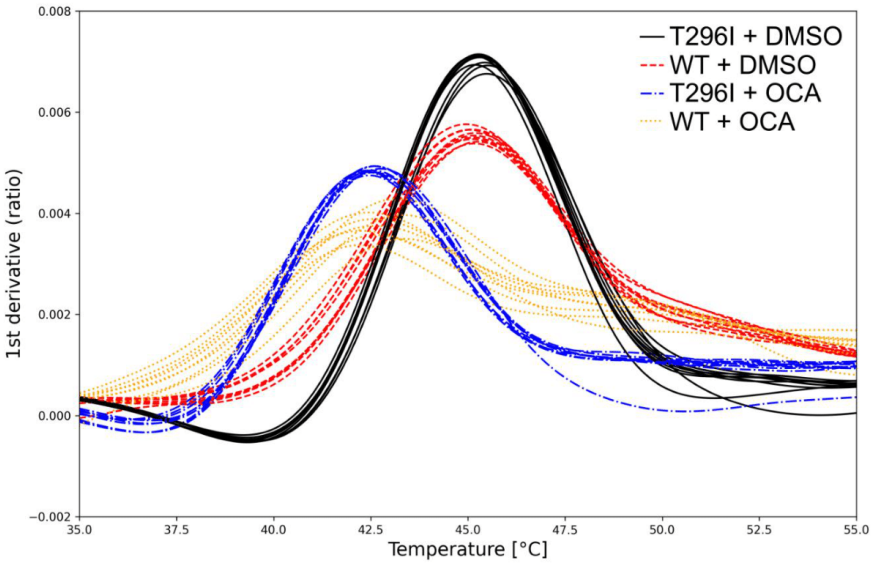
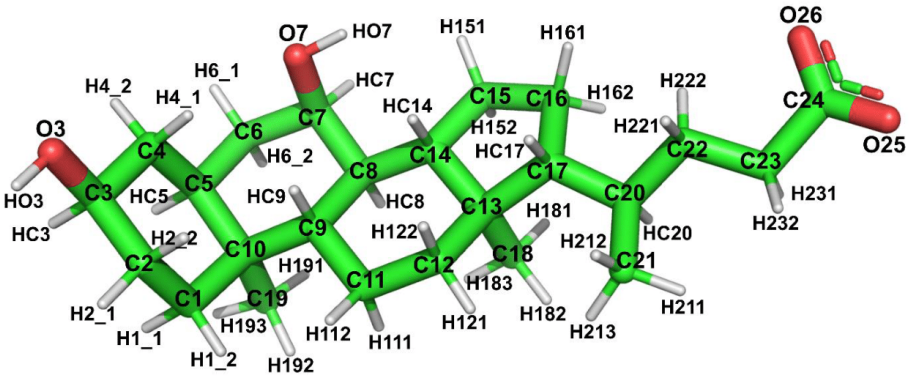


Fig. S7: Thermostability assay of purified FXR WT and FXR T296I revealing binding of agonist OCA. First derivative of fluorescence ratio 350nm over 330nm measured via nanoDSF. Measured melting temperatures and significance testing are shown Fig. S8.

		Mean [°C]	Standard deviation [°C]
*** * n.s.	FXR WT + DMSO	45.17	0.09
*** * n.s.	FXR T296I + DMSO	45.36	0.12
*** n.s.	FXR WT + OCA	42.84	0.41
*** n.s.	FXR T296I + OCA	42.58	0.12

Fig. S8: Melting point measurements of FXR WT and FXR T296I. Tm was measured via nanoDSF (Fig. S7) and statistical significance was tested using Welch's *t*-test. Protein samples were measured either in the absence (DMSO) or presence of agonist OCA.



Atom	Charge	Atom	Charge	Atom	Charge	Atom	Charge
C1	-0.0476	C7	0.1181	C15	-0.1328	C21	-0.0746
H1_1	0.0044	HC7	0.0410	H151	0.0228	H211	0.0044
H1_2	0.0044	O7	-0.6947	H152	0.0228	H212	0.0044
C2	-0.0483	HO7	0.4468	C16	-0.0703	H213	0.0044
H2_1	0.0171	C8	-0.0346	H161	0.0529	C22	0.0418
H2_2	0.0171	HC8	0.0557	H162	0.0529	H221	-0.0266
C3	0.1878	C9	0.0268	C17	0.0099	H222	-0.0266
HC3	0.0129	HC9	0.0527	HC17	0.0100	C23	-0.0168
O3	-0.7098	C10	0.0880	C18	-0.1016	H231	-0.0319
HO3	0.4297	C11	-0.0192	H181	0.0165	H232	-0.0319
C4	-0.0620	H111	0.0115	H182	0.0165	C24	0.7860
H4_1	0.0852	H112	0.0115	H183	0.0165	O25	-0.8101
H4_2	0.0852	C12	-0.0520	C19	-0.0961	O26	-0.8101
C5	-0.0533	H121	-0.0008	H191	0.0107		
HC5	0.0144	H122	-0.0008	H192	0.0107		
C6	-0.0347	C13	0.0863	H193	0.0107		
H6_1	0.0210	C14	0.0017	C20	0.0470		
H6_2	0.0210	HC14	0.0098	HC20	-0.0038		

Fig. S9: Atomic point charges of the CDCA ligand used for MD simulations. Top: Visualization of ligand structure and atom names. Bottom: RESP-derived atomic point charges.

Supplementary Tables

Table S1: Frequencies of occurrence of the distance between residue T296 and T466 (C_β atoms) below the cutoff of 4.6Å, calculated for each MD simulations replica.

Replica #	Active WT[a]	Active T296I[a]	Inactive WT[a]	Inactive T296I[a]
1	25.52	0.72	0.00	0.00
2	28.41	0.27	4.17	0.00
3	16.60	1.01	0.00	0.00
4	35.54	0.05	2.87	0.00
5	25.91	0.67	0.00	0.00
6	32.81	0.04	2.27	0.40
7	17.34	0.01	0.00	0.00
8	31.91	0.18	8.82	0.00
9	25.83	0.00	0.00	0.00
10	36.84	1.39	0.00	0.00
11	35.10	0.25	0.00	0.00
12	27.08	1.04	0.00	0.00
13	9.13	0.00	1.57	0.00
14	20.44	0.05	7.13	0.00
15	35.84	0.31	0.00	0.00
Mean	26.95	0.40	1.79	0.03
STD	8.20	0.45	2.85	0.10

[a] In %.

Table S2: Frequencies of occurrence of the RMSD of H12 with respect to the initial active reference structure below the cutoff of 1.9Å, calculated for each MD simulations replica.

Replica #	Active WT[a]	Active T296I[a]	Inactive WT[a]	Inactive T296I[a]
1	52.92	5.97	0.00	0.00
2	51.83	9.88	15.40	0.00
3	54.43	12.72	0.00	0.00
4	88.11	0.46	0.00	0.00
5	49.94	9.75	0.00	0.00
6	60.23	14.64	0.26	0.01
7	42.79	1.39	0.00	0.00

8	85.92	1.84	10.17	0.00
9	53.63	12.91	0.00	0.00
10	79.32	14.81	0.00	0.00
11	59.12	19.89	0.00	0.00
12	66.80	5.80	0.00	0.00
13	8.33	0.02	14.29	0.00
14	22.62	27.15	0.00	0.00
15	59.88	17.60	0.00	0.00
Mean	55.72	10.32	2.67	0.00
STD	21.15	7.90	5.59	0.00

[a] In %.

Table S3: List of TaqMan™ Gene Expression Assays used for pre-amplification qPCR analysis.

Accession number	Gene Symbol	Taq ID	Amplicon length (base pairs)	Exon boundary	RefSeq
Hs01026590_m1	NR1H4	FXR1	78	10-11	NM_001206979.1
Hs01026592_m1	NR1H4	FXR2	84	2-3	NM_001206979.1
Hs00994824_m1	ABCB11	BSEP1	93	3-4	NM_003742.2
Hs00994811_m1	ABCB11	BSEP2	77	16-17	NM_003742.2
Hs00222677_m1	NR0B2	SHP	87	1-2	NM_021969.2
Hs99999909_m1	HPRT1	HPRT1	100	6-7	NM_000194.2
Hs00188166_m1	SDHA	SDHA	70	5-6	NM_001294332.1
Hs02786624_g1	GAPDH	GAPDH	156	7	NM_001256799.2

Supplementary Movies

Movie S1: Trajectory of inactive WT (replica no. 2), visualized over the entire MD simulations time. The trajectory displays the conformational change from the inactive to the active state. For an easier judgment of the protein conformational state, the initial active state (based on the agonist-bound crystal structure) is depicted as green translucent representation. Important residues and motifs are highlighted (T296 as green sticks, T466 as cyan sticks, H12 in dark blue, NCoA2 peptide in magenta, and CDCA as pink sticks).

Movie S2: Trajectory of inactive T296I (replica no. 6), visualized over the entire MD simulations time. The trajectory displays the conformational change from an inactive to an almost active state. For an easier judgment of the protein conformational state, the initial active state (based on the agonist-bound crystal structure) is depicted as green translucent representation. Important residues and motifs are highlighted (I296 as orange sticks, T466 as cyan sticks, H12 in dark blue, NCoA2 peptide in magenta, and CDCA as pink sticks).

Supplementary Methods

Plasmids, cloning and mutagenesis

The BSEP promoter plasmid based on pGL3-basic ($BSEP^{prom}$-Luc) was a kind gift from Roche. The human *SHP* promoter (bases -572 to +10, GenBank Accession Number AF044316) (8) was amplified by PCR using forward (5′- agggtaccTCCTAGACTGGACAGTGGGCAAAG-3′) and reverse (5′- gtgctagcCTTCCAGCTCTCTGGCTCTGTGTT-3′) to introduce, respectively, exogenous *KpnI* and *NheI* sites at 5´ends. Genomic DNA was extracted from human liver tissue, using the DNeasy® Blood & Tissue Kit (Qiagen) and was PCR amplified with the above primers. pCDNA3.1(+)-hRXR α (9) was obtained from Addgene (#135910). The FXR coding sequence was amplified from a human liver cDNA pool with the primer pair FXR-S1/-S2 (5´-ATGGGATCAAAAATGAATCTCATTGAACA-3´; 5´-TCACTGCACGTCCCAGATTTCACAGAG-3´) using Phusion HiFi DNA polymerase (ThermoFisher scientific). We thus obtained pCR2.1-FXR (α1 isoform, acc. no. NM_001206979.2). The FXR expression vector was constructed as follows: pmCherry-N1 (TaKaRa Bio) was linearized by PCR using primer pair pmCherry-tagdel-S1/-S2 (5´-CGGCCGCGACTCTAGATCATA-3´; 5´-GGTGGCGACCGGTGGATCCC-3´), removing the mCherry coding sequence in the process. FXR was amplified from pCR2.1-FXR using primer pair FXR-IFHD-S1/-S2 (5´-CCACCGGTCGCCACCATGGGATCAAAAATGAATCTCATTGAACA-3´; 5´-CTAGAGTCGCGGCCGTCACTGCACGTCCCAGATTTCACAGAG-3´), adding necessary terminal homologous overhangs. The α2 isoform (acc no. NM_005123.4) of FXR was generated from pnoCherry-FXR by inverse PCR using the primer pair FXR-MYTGΔ-S1/-S2 (5´-CTTGTTAACTGAAATTCAGTGTAAATCTAAGCGACTGAG-3´; 5´-CATTCAGCCAACATTCCCATCTCTTTGCATTTCC-3´) followed by phosphorylation of the 5´-termini and blunt-end self-ligation. The T296I variant was introduced into both FXR isoforms by site-directed mutagenesis using primer FXR-T296I-SDM (5´-TGACGGAAATGGCAATCAATCATGTACAGGTTCTT-3´) and the QuikChange Lightning Multi kit (Stratagene) according to the manufacturer´s instructions. DNA sequencing was performed for all cDNAs used (Eurofins). Note that the numbering of the protein variant (T296I) is based on the alpha1 isoform (Uniprot acc. Q96RI1-1).

Luciferase Assay

Luciferase assays were performed using the Dual Luciferase reporter assay (Promega) according to the manufacturer´s instructions. Briefly, HEK293 cells kept in DMEM containing 10% fetal calf serum (FCS) were seeded onto 12-well plates at 150.000 cells per well and

transfected the next morning with 1µg of the BSEP-Luc or SHP-Luc plasmid and 100ng each of FXR and RXR expression plasmids using Fugene HD (Promega) at a ratio of 2.5:1 (reagent:DNA). Where applicable, plasmids were substituted with equal amounts of their respective empty backbones as control, and each well additionally received 50ng of pRL-TK as internal assay control so that the total amount of DNA per well was always 1.25µg. Each condition was assayed in three independent replicates for $BSEP^{prom}$-Luc-based experiments and six independent replicates for SHP^{prom}-Luc-based experiments. 4h after transfection, cells were pre-starved overnight by a medium change to DMEM containing 1% FCS (starvation medium) before stimulation with ligands in starvation medium for 2h. Cells transfected with FXRα1/2 expression plasmid were stimulated with 10µM obeticholic acid (OCA, INT-747), cells transfected with RXRα were stimulated with 1µM 9-cis-retinoic acid (9-cis-RA), cells transfected with both FXRα1/2 and RXRα expression plasmids were stimulated with both ligands. Cells were washed with PBS, lysed in 80µL passive lysis buffer at RT for 20min and scraped into 1.5mL microcentrifuge tubes. Lysates were cleared by centrifugation at 16.000 g at 4°C for 10min, and supernatants were kept on ice. 10µL of each sample were assayed in duplicate using 50µL each of LARII and Stop & Glo reagents in a GloMax multi detection system (Promega).

RNA preparation, reverse transcription, pre-amplification, and PCR analysis

Total RNA was extracted from FFPE blocks of the patient's liver, one control liver, and two cirrhotic livers and purified using the AmoyDx FFPE DNA/RNA Kit (Amoy Diagnostics Co.) according to the manufacturer's instructions. The non-tumorous liver tissue from a patient undergoing liver metastasis resection was used as control tissue and histopathologically contained no signs of hepatitis or fibrosis (denoted as healthy control). Cirrhotic tissue was obtained from two patients undergoing resection of hepatocellular carcinoma. The cirrhotic tissue used was histopathologically free of HCC. The study was approved by the local ethics committee Magdeburg, Germany (33/01) and Hannover, Germany (10062_BO_K_2021). Purified RNA was eluted with 30µL nuclease-free water. RNA integrity was assessed by microcapillary electrophoresis on 2100 BioAnalyser (Agilent Technologies). 100ng of RNA was reverse transcribed for 120min at 37°C followed by 5min of enzyme inactivation at 85°C using the High Capacity cDNA Reverse Transcription Kit (Applied Biosystems) with the following components: 1xRT buffer, 100mM dNTP mix, 1xRandom Primers, 50 U of MultiScribe Reverse Transcriptase and 1 U of RNAse inhibitor. cDNA was then diluted 1:3 in nuclease-free water. Before qPCR, sequences of interest in the cDNA were pre-amplified through 10 cycles using 1xTaqMan PreAmp Master Mix (Applied Biosystems) and a pool of TaqMan Gene Expression Assays (Table S3), diluted 1:200 in nuclease-free water, in a final volume of 50µL (0.05x each

Assay). Relative quantification of mRNA was performed according to the comparative $2^{-\Delta\Delta CT}$ method with SDHA as an endogenous control. HPRT1 and GAPDH were excluded as endogenous controls due to high Ct values. All TaqMan Gene Expression Assays used for cDNA pre-amplification and qPCR amplification were ordered from Applied Biosystems.

Setup of MD simulations and production replicas

The structural models were prepared for molecular dynamics (MD) simulations using the AMBER21 package (10). Overall, four different systems were prepared: FXR wildtype in the active state (hereafter termed “active WT”), FXR T296I variant in the active state (“active T296I”), FXR wildtype in the inactive state (“inactive WT”), and FXR T296I variant in the inactive state (“inactive T296I”). Maestro (Schrödinger, LLC, New York) was used for assigning protonation states with PROPKA (11) at pH 7.0; histidine HIP states were reverted to HIE. Parameters for the protein were taken from the ff14SB force field (12), and the TIP3P (13) parameters were used for the water and ions. The protein was solvated in a cubic water box, and Na^+ ions were added to neutralize the protein charges using tleap (10). The CDCA ligand was parametrized in its physiologically relevant deprotonated form. Electrostatic point charges were obtained with the RESP method (14) to represent the electrostatic potential of the ligand using the R.E.D. server (PyRED version April 2022) (14-17) with the electrostatic potential calculated at the 6-31G(d) level of theory using Gaussian16 vC.01 (Gaussian, Inc., Wallingford, USA). In line with the deprotonated state of the carboxylic acid group, the overall molecule charge was kept at -1. The ligand, as well as the derived point charges, are depicted in Fig. S9. Applying the SHAKE algorithm (18) to constrain bond lengths of hydrogen atoms to heavy atoms enabled a time step of 2fs. Long-range electrostatic interactions were considered using the particle mesh Ewald algorithm (19). Fifteen independent replicas were set up for each system, and each system was minimized for 1000 steps using the steepest descent algorithm, followed by 1000 steps using the conjugate gradient algorithm, applying harmonic positional restraints with a force constant of 50 kcal mol^{-1} $Å^{-2}$ to the system excluding the protein hydrogens, water molecules, protein side chains, protein, and ligand atoms in consecutive runs. The system was heated to 300 K for 25ps in the NVT ensemble using the Langevin thermostat with a collision frequency of $2.0ps^{-1}$ (20) with the protein atoms restrained with a force constant of 10 kcal mol^{-1} $Å^{-2}$. Within the following seven consecutive equilibration steps performed in the NPT ensemble using the Berendsen barostat (21) at 1 bar, the restraints were removed (after 100ps of the overall 4975ps simulation time). Using hydrogen mass repartitioning (22), the time step was increased to 4fs for the production replicas (each replica was simulated for 1µs, giving a total length of 15µs per system). Coordinates were stored in time steps of 100ps.

Analysis of MD simulations

Postprocessing and analysis of the MD trajectories were performed with CPPTRAJ (23) implemented in AmberTools21 (10). Root mean square deviations (RMSD) of the systems over the production time were based on C_α atom positions (for FXR and NCoA2) or all atoms (for CDCA) to analyze the structural variability of the systems. Analysis of the RMSD of H12 and the preceding T466 (residues 466 to 473) was performed against the reference of the active state of the initial FXR WT structure, first fitting the conformations along a trajectory on the most stable core of the four different FXR systems (active WT, active T296I, inactive WT, inactive T296I) over the production time. In detail, frames were extracted from all trajectories every 10ns and analyzed using the BIO3D package (24-26) to identify the least mobile residues throughout the simulations, which resulted in a part of helix 4 (residues 328 to 335) located at the core of the FXR LBD. The distance between the C_β atoms of residue 296 and T466 was measured to avoid a bias due to side-chain motions. Visualization was done using PyMOL v2.4.0. (Schrödinger, LLC, New York) or VMD v1.9.3. (27). The distance between H12 (residues 467 to 473) and LBD core (residues 248 to 459) was measured using the center of mass function within PyMOL.

Statistical Analysis of MD simulations

Significance tests were performed using the Mann-Whitney U test if not indicated otherwise. The indicated significance levels are n.s. (not significant): $p > 0.05$, *: $p \leq 0.05$, **: $p \leq 0.01$, ***: $p \leq 0.001$, ****: $p \leq 0.0001$.

Bacterial expression plasmid construction

In order to express the FXRα2 isoform (residues 1 to 472, uniprot acc. Q96RI1-2) in *Escherichia coli* (*E. coli*), the FXRα2 gene was PCR-amplified (Q5 Polymerase, NEB, Ipswich, USA) from the pnoCherry::fxr_alpha2_wt plasmid with the following primers: FXR_fw 5'-GAACAGATTGGTGGTATGGGATCAAAAATGAATCTC-3' and FXR_rv 5'-CAGCCGGATCTCACTGCACGTCCCAGATTTC-3'. The pET SUMO vector backbone (Invitrogen, Waltham, USA) was PCR-amplified using the following primers: pET SUMO_fw 5'-TGAGATCCGGCTGCTAACAAAGCC-3' and pET SUMO_rv 5'-ACCACCAATCTGTTCTCTGTGAGC-3'. PCR products were run on agarose gels for expected size verification, DpnI (NEB) digested for 1h at 37°C and purified using NucleoSpin Gel and PCR clean-up kit (Macherey-Nagel, Dueren, Germany). For homologous recombination, vector backbone and insert were used at a ratio of 1:3 for transformation of competent *E. coli* DH5alpha (DE3) (Invitrogen, Waltham, USA). Transformed cells were plated on Luria-Bertani

agar plates containing 50µg/mL kanamycin (Sigma Aldrich, St. Louis, USA) and incubated overnight at 37°C. Plasmid isolation of single colonies was performed using the NucleoSpin Plasmid kit (Macherey-Nagel) according to manufacturer's instruction and sequenced (Eurofins, Luxembourg). Recombinant plasmid (pET SUMO::fxr_alpha2_wt) was further used for the transformation of competent *E. coli* Rosetta (DE3) pLysS cells (Sigma Aldrich) for protein expression. The T296I mutation was introduced via a modified QuikChange protocol taking the pET-SUMO-FXR_alpha2_wt as template, using the primers: FXR_T296I_fw 5'-GACGGAAATGGCAattAATCATGTACAGG-3' and FXR_T296I_rv 5'-CATGATTaatTGCCATTTCCGTCAAAATG-3' (small letters indicating codon exchange). Similar to the wildtype, variant recombinant plasmid pET-SUMO::fxr_alpha2_T296I was first subcloned in *E. coli* DH5alpha cells, sequenced to verify the site-specific exchange, and used for transformation of competent *E. coli* Rosetta (DE3) pLysS for protein expression.

Protein expression and purification

FXR WT and T296I protein expression was performed with *E. coli* Rosetta(DE3) pLysS pET-SUMO::fxr_alpha2_wt or pET-SUMO::fxr_alpha2_T292I grown in terrific broth (TB) medium with 50µg/mL kanamycin (Sigma Aldrich). Cultures were grown at 37°C and shaking at 180 rpm until reaching an OD600 of 0.9. Protein expression was induced with 1mM of IPTG (Sigma Aldrich, St. Louis, USA) and cultures were further incubated at 20°C for 20h at 180 rpm. Bacterial cells were harvested by centrifugation (9,150xg for 12min at 4°C (Avanti JXN-26, Beckman Coulter, Brea, USA)) and cell pellets either stored at -80°C or directly used for subsequent cell lysis. For cell lysis, the bacterial pellet was resuspended in 10mL/g Lysis Buffer (50mM Na_2HPO_4, 300mM NaCl, pH 8.0, supplemented with protease and phosphatase inhibitors (Complete tablets, Roche, Basel, Switzerland)), 1mg/mL of lysozyme (Sigma Aldrich) and 5µg/mL of DNAse 1 (Roche, Basel, Switzerland) were added. The mixture was then sonicated on ice three consecutive times for 15min each with an interval pulse of 2 seconds (UP200St Sonicator, Hielscher Ultrasonics, Teltow, Germany). Cell debris was removed by centrifugation at 58,540xg for 60min at 4°C (Avanti JXN-26, Beckman Coulter, Brea, USA) before loading the supernatant on a pre-equilibrated HisTrap column (HisTrap HP, Cytiva, Marlborough, USA). Pre-equilibration was performed with binding buffer (50mM Na_2HPO_4, 300mM NaCl, 30mM imidazole (Thermo Fisher Scientific, Waltham, USA), pH 8.0). After protein extract application, the column was washed with 4 column volumes of binding buffer, before eluting the recombinant protein with elution buffer (50mM Na_2HPO_4, 300mM NaCl, 500mM imidazole, pH 8.0). The purity of the recombinant protein was further improved via size exclusion chromatography (SEC) using a HiLoad 16/600 Superdex 200pg column (Cytiva, Marlborough, USA, product number 28989335) in SEC buffer (50mM Na_2HPO_4, 300mM NaCl,

pH 8.0). Samples for SDS-PAGE analysis (12% gels, MiniProtean Gel, BioRad, Hercules, USA) were taken before and after induction, after HisTrap elution, and after SEC elution (Fig. S6). Fractions containing purified protein were concentrated and stored at -80°C in 50µM aliquots.

Thermostability assay

Recombinant FXR WT and FXR T296I protein samples were analyzed for their thermostability using the nanoDSF technology (Prometheus, Nanotemper, Munich, Germany). Protein samples were analyzed at a concentration of 25µM in SEC buffer (50mM Na_2HPO_4, 300mM NaCl, pH 8.0) either in the presence of 10-fold excess of the agonist OCA (INT-747, AbMole, Houston, USA) (250µM in 2.5% DMSO) or with a respective 2.5% DMSO control. Prometheus Standard capillaries (Nanotemper, Munich, Germany; with three technical replicates per sample) were used and three runs from 20°C to 90°C with an increase of 1°C per minute were performed (Fig. S7 and Fig. S8).

Supplementary References

1. Saen-Oon S, Lozoya E, Segarra V, Guallar V, Soliva R. Atomistic simulations shed new light on the activation mechanisms of RORγ and classify it as Type III nuclear hormone receptor regarding ligand-binding paths. Scientific Reports 2019;9:17249.
2. Grebner C, Lecina D, Gil V, Ulander J, Hansson P, Dellsen A, Tyrchan C, et al. Exploring Binding Mechanisms in Nuclear Hormone Receptors by Monte Carlo and X-ray-derived Motions. Biophys J 2017;112:1147-1156.
3. Fischer A, Smiesko M. Ligand Pathways in Nuclear Receptors. J Chem Inf Model 2019;59:3100-3109.
4. Li W, Fu J, Cheng F, Zheng M, Zhang J, Liu G, Tang Y. Unbinding pathways of GW4064 from human farnesoid X receptor as revealed by molecular dynamics simulations. J Chem Inf Model 2012;52:3043-3052.
5. Waskom ML. seaborn: statistical data visualization. Journal of Open Source Software 2021;6.
6. Merk D, Sreeramulu S, Kudlinzki D, Saxena K, Linhard V, Gande SL, Hiller F, et al. Molecular tuning of farnesoid X receptor partial agonism. Nat Commun 2019;10:2915.
7. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 2020;17:261-272.
8. Lee HK, Lee YK, Park SH, Kim YS, Park SH, Lee JW, Kwon HB, et al. Structure and expression of the orphan nuclear receptor SHP gene. J Biol Chem 1998;273:14398-14402.
9. Zolfaghari R, Mattie FJ, Wei CH, Chisholm DR, Whiting A, Ross AC. CYP26A1 gene promoter is a useful tool for reporting RAR-mediated retinoid activity. Anal Biochem 2019;577:98-109.
10. Case DA, Aktulga HM, Belfon K, Ben-Shalom IY, Borzell SR, Cerutti DS, Cheatham TE, 3rd, et al. Amber. In. University of California, San Francisco; 2021.
11. Bas DC, Rogers DM, Jensen JH. Very fast prediction and rationalization of pKa values for protein-ligand complexes. Proteins 2008;73:765-783.
12. Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. J Chem Theory Comput 2015;11:3696-3713.
13. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of simple potential functions for simulating liquid water. The Journal of Chemical Physics 1983;79:926-935.
14. Bayly CI, Cieplak P, Cornell W, Kollman PA. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. The Journal of Physical Chemistry 1993;97:10269-10280.
15. Vanquelef E, Simon S, Marquant G, Garcia E, Klimerak G, Delepine JC, Cieplak P, et al. R.E.D. Server: a web service for deriving RESP and ESP charges and building force field libraries for new molecules and molecular fragments. Nucleic Acids Res 2011;39:W511-517.
16. Dupradeau F-Y, Pigache A, Zaffran T, Savineau C, Lelong R, Grivel N, Lelong D, et al. The R.E.D. tools: advances in RESP and ESP charge derivation and force field library building. Physical Chemistry Chemical Physics 2010;12:7821-7839.
17. Wang F, Becker JP, Cieplak P, Dupradeau FY. R.E.D. Python: Object oriented programming for Amber force fields. Université de Picardie-Jules Verne, Sanford Burnham Prebys Medical Discovery Institute 2013.

18. Ryckaert J-P, Ciccotti G, Berendsen HJC. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. Journal of Computational Physics 1977;23:327-341.
19. Darden T, York D, Pedersen L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. The Journal of Chemical Physics 1993;98:10089-10092.
20. Pastor RW, Brooks BR, Szabo A. An analysis of the accuracy of Langevin and molecular dynamics algorithms. Molecular Physics 1988;65:1409-1419.
21. Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR. Molecular dynamics with coupling to an external bath. The Journal of Chemical Physics 1984;81:3684-3690.
22. Hopkins CW, Le Grand S, Walker RC, Roitberg AE. Long-Time-Step Molecular Dynamics through Hydrogen Mass Repartitioning. Journal of Chemical Theory and Computation 2015;11:1864-1874.
23. Roe DR, Cheatham TE, III. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. Journal of Chemical Theory and Computation 2013;9:3084-3095.
24. Grant BJ, Rodrigues APC, ElSawy KM, McCammon JA, Caves LSD. Bio3d: an R package for the comparative analysis of protein structures. Bioinformatics 2006;22:2695-2696.
25. Grant BJ, Skjærven L, Yao X-Q. The Bio3D packages for structural bioinformatics. Protein Science 2021;30:20-30.
26. Skjærven L, Yao X-Q, Scarabelli G, Grant BJ. Integrating protein structural dynamics and evolutionary analysis with Bio3D. BMC Bioinformatics 2014;15:399.
27. Humphrey W, Dalke A, Schulten K. VMD: Visual molecular dynamics. Journal of Molecular Graphics 1996;14:33-38.

Chapter 10 Bibliography

Abrams, C., & Bussi, G. (2013). Enhanced Sampling in Molecular Dynamics Using Metadynamics, Replica-Exchange, and Temperature-Acceleration. *Entropy*, *16*(1), 163–199. https://doi.org/10.3390/e16010163

Adorini, L., Pruzanski, M., & Shapiro, D. (2012). Farnesoid X receptor targeting to treat nonalcoholic steatohepatitis. *Drug Discovery Today*, *17*(17–18), 988–997. https://doi.org/10.1016/j.drudis.2012.05.012

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., & Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. In *Nature Methods* (Vol. 7, Issue 4, pp. 248–249). https://doi.org/10.1038/nmeth0410-248

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2007). *Molecular Biology of the Cell*. W.W. Norton & Company. https://doi.org/10.1201/9780203833445

Albrecht, S., Fleck, A.-K., Kirchberg, I., Hucke, S., Liebmann, M., Klotz, L., & Kuhlmann, T. (2017). Activation of FXR pathway does not alter glial cell function. *Journal of Neuroinflammation*, *14*(1), 66. https://doi.org/10.1186/s12974-017-0833-6

Alder, B. J., & Wainwright, T. E. (1959). Studies in Molecular Dynamics. I. General Method. *The Journal of Chemical Physics*, *31*(2), 459–466. https://doi.org/10.1063/1.1730376

Alrefai, W. A., & Gill, R. K. (2007). Bile Acid Transporters: Structure, Function, Regulation and Pathophysiological Implications. *Pharmaceutical Research*, *24*(10), 1803–1823. https://doi.org/10.1007/s11095-007-9289-1

Amanchy, R., Periaswamy, B., Mathivanan, S., Reddy, R., Tattikota, S. G., & Pandey, A. (2007). A curated compendium of phosphorylation motifs. *Nature Biotechnology*, *25*(3), 285–286. https://doi.org/10.1038/nbt0307-285

Amigo, L., Mendoza, H., Zanlungo, S., Miquel, J. F., Rigotti, A., González, S., & Nervi, F. (1999). Enrichment of canalicular membrane with cholesterol and sphingomyelin prevents bile salt-induced hepatic damage. *Journal of Lipid Research*, *40*(3), 533–542. https://doi.org/10.1016/S0022-2275(20)32458-5

Ananthanarayanan, M., Balasubramanian, N., Makishima, M., Mangelsdorf, D. J., & Suchy, F. J. (2001). Human Bile Salt Export Pump Promoter Is Transactivated by the Farnesoid X Receptor/Bile Acid Receptor. *Journal of Biological Chemistry*, *276*(31), 28857–28865. https://doi.org/10.1074/jbc.M011610200

Anbalagan, M., Huderson, B., Murphy, L., & Rowan, B. G. (2012). Post-Translational Modifications of Nuclear Receptors and Human Disease. *Nuclear Receptor Signaling*, *10*(1), nrs.10001. https://doi.org/10.1621/nrs.10001

Andress, E. J., Nicolaou, M., McGeoghan, F., & Linton, K. J. (2017). ABCB4 missense mutations D243A, K435T, G535D, I490T, R545C, and S978P significantly impair the lipid floppase and likely predispose to secondary pathologies in the human population. *Cellular and Molecular Life Sciences*, *74*(13), 2513–2524. https://doi.org/10.1007/s00018-017-2472-6

Andress, E. J., Nicolaou, M., Romero, M. R., Naik, S., Dixon, P. H., Williamson, C., & Linton, K. J. (2014). Molecular mechanistic explanation for the spectrum of cholestatic disease caused by the S320F variant of ABCB4. *Hepatology*, *59*(5), 1921–1931. https://doi.org/10.1002/hep.26970

Ansorge, W., Sproat, B. S., Stegemann, J., & Schwager, C. (1986). A non-radioactive automated method for DNA sequence determination. *Journal of Biochemical and Biophysical Methods*, *13*(6), 315–323. https://doi.org/10.1016/0165-022X(86)90038-2

Appelman, M. D., van der Veen, S. W., & van Mil, S. W. C. (2021). Post-Translational Modifications of FXR; Implications for Cholestasis and Obesity-Related Disorders. *Frontiers in Endocrinology*, *12*(September), 1–13. https://doi.org/10.3389/fendo.2021.729828

Aranda, A., & Pascual, A. (2001). Nuclear Hormone Receptors and Gene Expression. *Physiological Reviews*, *81*(3), 1269–1304. https://doi.org/10.1152/physrev.2001.81.3.1269

Asada, T., Doi, K., Inokuchi, R., Hayase, N., Yamamoto, M., & Morimura, N. (2019). Organ system network analysis and biological stability in critically ill patients. *Critical Care*, *23*(1), 83. https://doi.org/10.1186/s13054-019-2376-y

Attinkara, R., Mwinyi, J., Truninger, K., Regula, J., Gaj, P., Rogler, G., Kullak-Ublick, G. A., & Eloranta, J. J. (2012). Association of genetic variation in the NR1H4 gene, encoding the nuclear bile acid receptor FXR, with inflammatory bowel disease. *BMC Research Notes*, *5*(1), 461. https://doi.org/10.1186/1756-0500-5-461

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., … Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Bartsch, R. P., Liu, K. K. L., Bashan, A., & Ivanov, P. C. (2015). Network physiology: How organ systems dynamically

interact. *PLoS ONE*, *10*(11), 1–36. https://doi.org/10.1371/journal.pone.0142143

Bauer, E., & Kohavi, R. (1999). Empirical comparison of voting classification algorithms: bagging, boosting, and variants. *Machine Learning*, *36*(1), 105–139. https://doi.org/10.1023/a:1007515423169

Bayly, C. I., Cieplak, P., Cornell, W., & Kollman, P. A. (1993). A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *The Journal of Physical Chemistry*, *97*(40), 10269–10280. https://doi.org/10.1021/j100142a004

Bishop-Bailey, D., Walsh, D. T., & Warner, T. D. (2004). Expression and activation of the farnesoid X receptor in the vasculature. *Proceedings of the National Academy of Sciences*, *101*(10), 3668–3673. https://doi.org/10.1073/pnas.0400046101

Blaut, M., Collins, M. D., Welling, G. W., Doré, J., van Loo, J., & de Vos, W. (2002). Molecular biological methods for studying the gut microbiota: the EU human gut flora project. *British Journal of Nutrition*, *87*(6), 203–211. https://doi.org/10.1079/BJNBJN/2002539

Blesl, A., & Stadlbauer, V. (2021). The gut-liver axis in cholestatic liver diseases. *Nutrients*, *13*(3), 1–32. https://doi.org/10.3390/nu13031018

Blom, N., Gammeltoft, S., & Brunak, S. (1999). Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *Journal of Molecular Biology*, *294*(5), 1351–1362. https://doi.org/10.1006/jmbi.1999.3310

Böck, M., Malle, J., Pasterk, D., Kukina, H., Hasani, R., & Heitzinger, C. (2022). Superhuman performance on sepsis MIMIC-III data by distributional reinforcement learning. *PLOS ONE*, *17*(11), e0275358. https://doi.org/10.1371/journal.pone.0275358

Bonus, M., Sommerfeld, A., Qvartskhava, N., Görg, B., Ludwig, B. S., Kessler, H., Gohlke, H., & Häussinger, D. (2020). Evidence for functional selectivity in TUDC- and norUDCA-induced signal transduction via α5β1 integrin towards choleresis. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-020-62326-y

Botstein, D., & Risch, N. (2003). Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature Genetics*, *33*(S3), 228–237. https://doi.org/10.1038/ng1090

Bottaro, S., & Lindorff-Larsen, K. (2018). Biophysical experiments and biomolecular simulations: A perfect match? *Science*, *361*(6400), 355–360. https://doi.org/10.1126/science.aat4010

Boyer, J. L. (2013). Bile Formation and Secretion. In *Comprehensive Physiology* (Vol. 3, Issue 3, p.). John Wiley & Sons, Inc. https://doi.org/10.1002/cphy.c120027

Braun, E., Gilmer, J., Mayes, H. B., Mobley, D. L., Monroe, J. I., Prasad, S., & Zuckerman, D. M. (2019). Best Practices for Foundations in Molecular Simulations [Article v1.0]. *Living Journal of Computational Molecular Science*, *1*(1), 1–28. https://doi.org/10.33011/livecoms.1.1.5957

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5–32. https://doi.org/https://doi.org/10.1023/A:1010933404324

Broom, A., Jacobi, Z., Trainor, K., & Meiering, E. M. (2017). Computational tools help improve protein stability but with a solubility tradeoff. *Journal of Biological Chemistry*, *292*(35), 14349–14361. https://doi.org/10.1074/jbc.M117.784165

Burris, T. P., Montrose, C., Houck, K. A., Osborne, H. E., Bocchinfuso, W. P., Yaden, B. C., Cheng, C. C., Zink, R. W., Barr, R. J., Hepler, C. D., Krishnan, V., Bullock, H. A., Burris, L. L., Galvin, R. J., Bramlett, K., & Stayrook, K. R. (2005). The Hypolipidemic Natural Product Guggulsterone Is a Promiscuous Steroid Receptor Ligand. *Molecular Pharmacology*, *67*(3), 948–954. https://doi.org/10.1124/mol.104.007054

Butt, T. R., Edavettal, S. C., Hall, J. P., & Mattern, M. R. (2005). SUMO fusion technology for difficult-to-express proteins. *Protein Expression and Purification*, *43*(1), 1–9. https://doi.org/10.1016/j.pep.2005.03.016

Cabitza, F., Campagner, A., Soares, F., García de Guadiana-Romualdo, L., Challa, F., Sulejmani, A., Seghezzi, M., & Carobene, A. (2021). The importance of being external. methodological insights for the external validation of machine learning models in medicine. *Computer Methods and Programs in Biomedicine*, *208*, 106288. https://doi.org/10.1016/j.cmpb.2021.106288

Calimet, N., Simoes, M., Changeux, J.-P., Karplus, M., Taly, A., & Cecchini, M. (2013). A gating mechanism of pentameric ligand-gated ion channels. *Proceedings of the National Academy of Sciences*, *110*(42). https://doi.org/10.1073/pnas.1313785110

Capriotti, E., Fariselli, P., & Casadio, R. (2005). I-Mutant2.0: Predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Research*, *33*(SUPPL. 2). https://doi.org/10.1093/nar/gki375

Carey, M. C., & Small, D. M. (1978). The physical chemistry of cholesterol solubility in bile. Relationship to gallstone formation and dissolution in man. *Journal of Clinical Investigation*, *61*(4), 998–1026. https://doi.org/10.1172/JCI109025

Carino, A., Cipriani, S., Marchianò, S., Biagioli, M., Santorelli, C., Donini, A., Zampella, A., Monti, M. C., & Fiorucci,

S. (2017). BAR502, a dual FXR and GPBAR1 agonist, promotes browning of white adipose tissue and reverses liver steatosis and fibrosis. *Scientific Reports*, *7*(1), 42801. https://doi.org/10.1038/srep42801

Carlton, V. E. H., Harris, B. Z., Puffenberger, E. G., Batta, A. K., Knisely, A. S., Robinson, D. L., Strauss, K. A., Shneider, B. L., Lim, W. A., Salen, G., Morton, D. H., & Bull, L. N. (2003). Complex inheritance of familial hypercholanemia with associated mutations in TJP2 and BAAT. *Nature Genetics*, *34*(1), 91–96. https://doi.org/10.1038/ng1147

Case, D. A., Aktulga. H.M., Belfon, K., Ben-Shalom, I. Y., Brozell, S. R., Cerutti, D. S., Cheatham III, T. E., Cisneros, G. A., Cruzeiro, V. W. D., Darden, T. A., Duke, R. E., Giambasu, G., Gilson, M. K., Gohlke, H., Goetz, A. W., Harris, R., Izadi, S., Izmailov, S. A., Jin, C., … Kollman, P. A. (2021). Amber 2021. *University of California, San Francisco.*

Case, D. A., Aktulga, H. M., Belfon, K., Cerutti, D. S., Cisneros, G. A., Cruzeiro, V. W. D., Forouzesh, N., Giese, T. J., Götz, A. W., Gohlke, H., Izadi, S., Kasavajhala, K., Kaymak, M. C., King, E., Kurtzman, T., Lee, T.-S., Li, P., Liu, J., Luchko, T., … Merz, K. M. (2023). AmberTools. *Journal of Chemical Information and Modeling*, *63*(20), 6183–6191. https://doi.org/10.1021/acs.jcim.3c01153

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., Jacobsen, A., Byrne, C. J., Heuer, M. L., Larsson, E., Antipin, Y., Reva, B., Goldberg, A. P., Sander, C., & Schultz, N. (2012). The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discovery*, *2*(5), 401–404. https://doi.org/10.1158/2159-8290.CD-12-0095

Cessie, S. L., & Houwelingen, J. V. (1992). Ridge estimators in logistic regression. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *41*(1), 191–201.

Chandra, V., Huang, P., Hamuro, Y., Raghuram, S., Wang, Y., Burris, T. P., & Rastinejad, F. (2008). Structure of the intact PPAR-γ–RXR-α nuclear receptor complex on DNA. *Nature*, *456*(7220), 350–356. https://doi.org/10.1038/nature07413

Chandra, V., Huang, P., Potluri, N., Wu, D., Kim, Y., & Rastinejad, F. (2013). Multidomain integration in the structure of the HNF-4α nuclear receptor complex. *Nature*, *495*(7441), 394–398. https://doi.org/10.1038/nature11966

Charilaou, P., & Battat, R. (2022). Machine learning models and over-fitting considerations. *World Journal of Gastroenterology*, *28*(5), 605–607. https://doi.org/10.3748/wjg.v28.i5.605

Chawla, N. V, Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. In *Journal of Artificial Intelligence Research* (Vol. 16).

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-August-2016*, 785–794. https://doi.org/10.1145/2939672.2939785

Cheng, J., Randall, A., & Baldi, P. (2006). Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins: Structure, Function and Genetics*, *62*(4), 1125–1132. https://doi.org/10.1002/prot.20810

Chiang, J. Y. L. (2013). Bile Acid Metabolism and Signaling. In *Comprehensive Physiology* (pp. 1191–1212). Wiley. https://doi.org/10.1002/cphy.c120023

Chiang, J. Y. L., Kimmel, R., Weinberger, C., & Stroup, D. (2000). Farnesoid X Receptor Responds to Bile Acids and Represses Cholesterol 7α-Hydroxylase Gene (CYP7A1) Transcription. *Journal of Biological Chemistry*, *275*(15), 10918–10924. https://doi.org/10.1074/jbc.275.15.10918

Choudhury, A., Mohammad, T., Anjum, F., Shafie, A., Singh, I. K., Abdullaev, B., Pasupuleti, V. R., Adnan, M., Yadav, D. K., & Hassan, M. I. (2022). Comparative analysis of web-based programs for single amino acid substitutions in proteins. *PloS One*, *17*(5), e0267084. https://doi.org/10.1371/journal.pone.0267084

Chrisman, I. M., Nemetchek, M. D., de Vera, I. M. S., Shang, J., Heidari, Z., Long, Y., Reyes-Caballero, H., Galindo-Murillo, R., Cheatham, T. E., Blayo, A.-L., Shin, Y., Fuhrmann, J., Griffin, P. R., Kamenecka, T. M., Kojetin, D. J., & Hughes, T. S. (2018). Defining a conformational ensemble that directs activation of PPARγ. *Nature Communications*, *9*(1), 1794. https://doi.org/10.1038/s41467-018-04176-x

Cipriani, S., Renga, B., D'Amore, C., Simonetti, M., De Tursi, A. A., Carino, A., Monti, M. C., Sepe, V., Zampella, A., & Fiorucci, S. (2015). Impaired Itching Perception in Murine Models of Cholestasis Is Supported by Dysregulation of GPBAR1 Signaling. *PLOS ONE*, *10*(7), e0129866. https://doi.org/10.1371/journal.pone.0129866

Claudel, T., Sturm, E., Duez, H., Torra, I. P., Sirvent, A., Kosykh, V., Fruchart, J.-C., Dallongeville, J., Hum, D. W., Kuipers, F., & Staels, B. (2002). Bile acid-activated nuclear receptor FXR suppresses apolipoprotein A-I transcription via a negative FXR response element. *Journal of Clinical Investigation*, *109*(7), 961–971. https://doi.org/10.1172/JCI14505

Claudel, T., & Trauner, M. (2020). Bile Acids as Signaling Molecules. In *The Liver* (pp. 299–312). Wiley. https://doi.org/10.1002/9781119436812.ch25

Clayton, R. J. (1969). Byler Disease. *American Journal of Diseases of Children*, *117*(1), 112. https://doi.org/10.1001/archpedi.1969.02100030114014

Collins, F. S., & Fink, L. (1995). The Human Genome Project. *Alcohol Health and Research World*, *19*(3), 190–195. http://www.ncbi.nlm.nih.gov/pubmed/31798046

Collins, S. L., Stine, J. G., Bisanz, J. E., Okafor, C. D., & Patterson, A. D. (2023). Bile acids and the gut microbiota: metabolic interactions and impacts on disease. *Nature Reviews Microbiology*, *21*(4), 236–247. https://doi.org/10.1038/s41579-022-00805-x

Colombo, C., Vajro, P., Degiorgio, D., Coviello, D. A., Costantino, L., Tornillo, L., Motta, V., Consonni, D., & Maggiore, G. (2011). Clinical features and genotype-phenotype correlations in children with progressive familial intrahepatic cholestasis type 3 related to ABCB4 mutations. *Journal of Pediatric Gastroenterology and Nutrition*, *52*(1), 73–83. https://doi.org/10.1097/MPG.0b013e3181f50363

Coppola, C., Gosche, J., Arrese, M., Ancowitz, B., Madsen, J., Vanderhoof, J., & Shneider, B. (1998). Molecular analysis of the adaptive response of intestinal bile acid transport after ileal resection in the rat. *Gastroenterology*, *115*(5), 1172–1178. https://doi.org/10.1016/S0016-5085(98)70088-5

Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., & Kollman, P. A. (1995). A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society*, *117*(19), 5179–5197. https://doi.org/10.1021/ja00124a002

Cornell, W. D., Cieplak, P., Bayly, C. I., & Kollman, P. A. (1993). Application of RESP Charges To Calculate Conformational Energies, Hydrogen Bond Energies, and Free Energies of Solvation. *Journal of the American Chemical Society*, *115*(21), 9620–9631. https://doi.org/10.1021/ja00074a030

Coronato, A., Naeem, M., De Pietro, G., & Paragliola, G. (2020). Reinforcement learning for intelligent healthcare applications: A survey. *Artificial Intelligence in Medicine*, *109*, 101964. https://doi.org/10.1016/j.artmed.2020.101964

Correia, J. C., Massart, J., de Boer, J. F., Porsmyr-Palmertz, M., Martínez-Redondo, V., Agudelo, L. Z., Sinha, I., Meierhofer, D., Ribeiro, V., Björnholm, M., Sauer, S., Dahlman-Wright, K., Zierath, J. R., Groen, A. K., & Ruas, J. L. (2015). Bioenergetic cues shift FXR splicing towards FXRα2 to modulate hepatic lipolysis and fatty acid metabolism. *Molecular Metabolism*, *4*(12), 891–902. https://doi.org/10.1016/j.molmet.2015.09.005

Crockett, D. K., Lyon, E., Williams, M. S., Narus, S. P., Facelli, J. C., & Mitchell, J. A. (2012). Utility of gene-specific algorithms for predicting pathogenicity of uncertain gene variants. *Journal of the American Medical Informatics Association*, *19*(2), 207–211. https://doi.org/10.1136/amiajnl-2011-000309

Cui, J., Heard, T. S., Yu, J., Lo, J.-L., Huang, L., Li, Y., Schaeffer, J. M., & Wright, S. D. (2002). The Amino Acid Residues Asparagine 354 and Isoleucine 372 of Human Farnesoid X Receptor Confer the Receptor with High Sensitivity to Chenodeoxycholate. *Journal of Biological Chemistry*, *277*(29), 25963–25969. https://doi.org/10.1074/jbc.M200824200

Cui, J., Huang, L., Zhao, A., Lew, J.-L., Yu, J., Sahoo, S., Meinke, P. T., Royo, I., Peláez, F., & Wright, S. D. (2003). Guggulsterone Is a Farnesoid X Receptor Antagonist in Coactivator Association Assays but Acts to Enhance Transcription of Bile Salt Export Pump. *Journal of Biological Chemistry*, *278*(12), 10214–10220. https://doi.org/10.1074/jbc.M209323200

D'Arrigo, G., Autiero, I., Gianquinto, E., Siragusa, L., Baroni, M., Cruciani, G., & Spyrakis, F. (2022). Exploring Ligand Binding Domain Dynamics in the NRs Superfamily. *International Journal of Molecular Sciences*, *23*(15), 8732. https://doi.org/10.3390/ijms23158732

Dai, S. Y., Burris, T. P., Dodge, J. A., Montrose-Rafizadeh, C., Wang, Y., Pascal, B. D., Chalmers, M. J., & Griffin, P. R. (2009). Unique Ligand Binding Patterns between Estrogen Receptor α and β Revealed by Hydrogen−Deuterium Exchange. *Biochemistry*, *48*(40), 9668–9676. https://doi.org/10.1021/bi901149t

Dash, A., Figler, R. A., Blackman, B. R., Marukian, S., Collado, M. S., Lawson, M. J., Hoang, S. A., Mackey, A. J., Manka, D., Cole, B. K., Feaver, R. E., Sanyal, A. J., & Wamhoff, B. R. (2017). Pharmacotoxicology of clinically-relevant concentrations of obeticholic acid in an organotypic human hepatocyte system. *Toxicology in Vitro*, *39*, 93–103. https://doi.org/10.1016/j.tiv.2016.11.014

Davit-Spraul, A., Gonzales, E., Baussan, C., & Jacquemin, E. (2009). Progressive familial intrahepatic cholestasis. *Orphanet Journal of Rare Diseases*, *4*(1), 1. https://doi.org/10.1186/1750-1172-4-1

Davit-Spraul, A., Gonzales, E., Baussan, C., & Jacquemin, E. (2010). The spectrum of liver diseases related to ABCB4 gene mutations: Pathophysiology and clinical aspects. In *Seminars in Liver Disease* (Vol. 30, Issue 2, pp. 134–146). https://doi.org/10.1055/s-0030-1253223

Dawson, P. A., Haywood, J., Craddock, A. L., Wilson, M., Tietjen, M., Kluckman, K., Maeda, N., & Parks, J. S. (2003). Targeted Deletion of the Ileal Bile Acid Transporter Eliminates Enterohepatic Cycling of Bile Acids in Mice. *Journal of Biological Chemistry*, *278*(36), 33920–33927. https://doi.org/10.1074/jbc.M306370200

Dayhoff, M. O. (1966). *Atlas of protein sequence and structure.* National Biomedical Research Foundation.

de Aguiar Vallim, T. Q., Tarling, E. J., & Edwards, P. A. (2013). Pleiotropic Roles of Bile Acids in Metabolism. *Cell Metabolism*, *17*(5), 657–669. https://doi.org/10.1016/j.cmet.2013.03.013

Deckmyn, B., Domenger, D., Blondel, C., Ducastel, S., Nicolas, E., Dorchies, E., Caron, E., Charton, J., Vallez, E., Deprez, B., Annicotte, J.-S., Lestavel, S., Tailleux, A., Magnan, C., Staels, B., & Bantubungi, K. (2022). Farnesoid X Receptor Activation in Brain Alters Brown Adipose Tissue Function via the Sympathetic System. *Frontiers in Molecular Neuroscience*, *14*. https://doi.org/10.3389/fnmol.2021.808603

Degiorgio, D., Colombo, C., Seia, M., Porcaro, L., Costantino, L., Zazzeron, L., Bordo, D., & Coviello, D. A. (2007). Molecular characterization and structural implications of 25 new ABCB4 mutations in progressive familial intrahepatic cholestasis type 3 (PFIC3). *European Journal of Human Genetics*, *15*(12), 1230–1238. https://doi.org/10.1038/sj.ejhg.5201908

Degiorgio, D., Corsetto, P. A., Rizzo, A. M., Colombo, C., Seia, M., Costantino, L., Montorfano, G., Tomaiuolo, R., Bordo, D., Sansanelli, S., Li, M., Tavian, D., Rastaldi, M. P., & Coviello, D. A. (2013). *Two ABCB4 point mutations of strategic NBD-motifs do not prevent protein targeting to the plasma membrane but promote MDR3 dysfunction*. *22*(5), 633–639. https://doi.org/10.1038/ejhg.2013.214

Degiorgio, D., Corsetto, P. A., Rizzo, A. M., Colombo, C., Seia, M., Costantino, L., Montorfano, G., Tomaiuolo, R., Bordo, D., Sansanelli, S., Li, M., Tavian, D., Rastaldi, M. P., & Coviello, D. A. (2014). Two ABCB4 point mutations of strategic NBD-motifs do not prevent protein targeting to the plasma membrane but promote MDR3 dysfunction. *European Journal of Human Genetics*, *22*(5), 633–639. https://doi.org/10.1038/ejhg.2013.214

Delaunay, J., Bruneau, A., Hoffmann, B., Durand-Schneider, A.-M., Barbu, V., Jacquemin, E., Maurice, M., Housset, C., Callebaut, I., & Aït-Slimane, T. (2017). Functional defect of variants in the adenosine triphosphate-binding sites of ABCB4 and their rescue by the cystic fibrosis transmembrane conductance regulator potentiator, ivacaftor (VX-770). *Hepatology*, *65*(2), 560–570. https://doi.org/10.1002/hep.28929

Delaunay, J., Durand-Schneider, A., Dossier, C., Falguières, T., Gautherot, J., Davit-Spraul, A., Aït-Slimane, T., Housset, C., Jacquemin, E., & Maurice, M. (2016). A functional classification of ABCB4 variations causing progressive familial intrahepatic cholestasis type 3. *Hepatology*, *63*(5), 1620–1631. https://doi.org/10.1002/hep.28300

Deleuze, J., Jacquemin, E., Dubuisson, C., Cresteil, D., Dumont, M., Erlinger, S., Bernard, O., & Hadchouel, M. (1996). Defect of multidrug-resistance 3 gene expression in a subtype of progressive familial intrahepatic cholestasis. *Hepatology*, *23*(4), 904–908. https://doi.org/10.1002/hep.510230435

Denson, L. A., Sturm, E., Echevarria, W., Zimmerman, T. L., Makishima, M., Mangelsdorf, D. J., & Karpen, S. J. (2001). The Orphan Nuclear Receptor, shp, Mediates Bile Acid-Induced Inhibition of the Rat Bile Acid Transporter, ntcp. *Gastroenterology*, *121*(1), 140–147. https://doi.org/10.1053/gast.2001.25503

Devarakonda, S., Harp, J. M., Kim, Y., Ozyhar, A., & Rastinejad, F. (2003). Structure of the heterodimeric ecdysone receptor DNA-binding complex. *The EMBO Journal*, *22*(21), 5827–5840. https://doi.org/10.1093/emboj/cdg569

Di Gregorio, M. C., Cautela, J., & Galantini, L. (2021). Physiology and physical chemistry of bile acids. *International Journal of Molecular Sciences*, *22*(4), 1–23. https://doi.org/10.3390/ijms22041780

Díaz-Holguín, A., Rashidian, A., Pijnenburg, D., Monteiro Ferreira, G., Stefela, A., Kaspar, M., Kudova, E., Poso, A., van Beuningen, R., Pavek, P., & Kronenberger, T. (2023). When Two Become One: Conformational Changes in FXR/RXR Heterodimers Bound to Steroidal Antagonists. *ChemMedChem*, *18*(4). https://doi.org/10.1002/cmdc.202200556

Dickson, C. J., Walker, R. C., & Gould, I. R. (2022). Lipid21: Complex Lipid Membrane Simulations with AMBER. *Journal of Chemical Theory and Computation*, *18*(3), 1726–1736. https://doi.org/10.1021/acs.jctc.1c01217

Dietterich, T. G. (2000). Ensemble methods in machine learning. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *1857 LNCS*, 1–15. https://doi.org/10.1007/3-540-45014-9_1

Dixon, P. H., Weerasekera, N., Linton, K. J., Donaldson, O., Chambers, J., Egginton, E., Weaver, J., Nelson-Piercy, C., De Swiet, M., Warnes, G., Elias, E., Higgins, C. F., Johnston, D. G., Mccarthy, M. I., & Williamson, C. (2000). Heterozygous MDR3 missense mutation associated with intrahepatic cholestasis of pregnancy: evidence for a defect in protein trafficking. In *Human Molecular Genetics* (Vol. 9, Issue 8).

Dobbin, K. K., & Simon, R. M. (2011). Optimally splitting cases for training and testing high dimensional classifiers. *BMC Medical Genomics*, *4*(1), 31. https://doi.org/10.1186/1755-8794-4-31

Dong, C., Condat, B., Picon-Coste, M., Chretien, Y., Potier, P., Noblinski, B., Arrivé, L., Hauuy, M.-P., Barbu, V., Maftouh, A., Gaouar, F., Ben Belkacem, K., Housset, C., Poupon, R., Zanditenas, D., Chazouilleres, O., & Corpechot, C. (2020). Low phospholipid-associated cholelithiasis syndrome: prevalence, clinical features, and comorbidities. *JHEP Reports*, 100201. https://doi.org/10.1016/j.jhepr.2020.100201

Dong, C., Condat, B., Picon-Coste, M., Chrétien, Y., Potier, P., Noblinski, B., Arrivé, L., Hauuy, M. P., Barbu, V.,

Maftouh, A., Gaouar, F., Ben Belkacem, K., Housset, C., Poupon, R., Zanditenas, D., Chazouillères, O., & Corpechot, C. (2021). Low-phospholipid-associated cholelithiasis syndrome: Prevalence, clinical features, and comorbidities. *JHEP Reports*, *3*(2). https://doi.org/10.1016/j.jhepr.2020.100201

Dong, X., Yu, Z., Cao, W., Shi, Y., & Ma, Q. (2020). A survey on ensemble learning. *Frontiers of Computer Science*, *14*(2), 241–258. https://doi.org/10.1007/s11704-019-8208-z

Dröge, C., Bonus, M., Baumann, U., Klindt, C., Lainka, E., Kathemann, S., Brinkert, F., Grabhorn, E., Pfister, E. D., Wenning, D., Fichtner, A., Gotthardt, D. N., Weiss, K. H., McKiernan, P., Puri, R. D., Verma, I. C., Kluge, S., Gohlke, H., Schmitt, L., … Keitel, V. (2017). Sequencing of FIC1, BSEP and MDR3 in a large cohort of patients with cholestasis revealed a high number of different genetic variants. *Journal of Hepatology*, *67*(6), 1253–1264. https://doi.org/10.1016/j.jhep.2017.07.004

Dröge, C., Götze, T., Behrendt, A., Gohlke, H., & Keitel, V. (2023). *Diagnostic workup of suspected hereditary cholestasis in adults : a case report Case report*. *3*, 34–43. https://doi.org/10.37349/edd.2023.00016

Dulac-Arnold, G., Levine, N., Mankowitz, D. J., Li, J., Paduraru, C., Gowal, S., & Hester, T. (2021). Challenges of real-world reinforcement learning: definitions, benchmarks and analysis. *Machine Learning*, *110*(9), 2419–2468. https://doi.org/10.1007/s10994-021-05961-4

Dupradeau, F.-Y., Pigache, A., Zaffran, T., Savineau, C., Lelong, R., Grivel, N., Lelong, D., Rosanski, W., & Cieplak, P. (2010). The R.E.D. tools: advances in RESP and ESP charge derivation and force field library building. *Physical Chemistry Chemical Physics*, *12*(28), 7821. https://doi.org/10.1039/c0cp00111b

Dzagania, T., Engelmann, G., Häussinger, D., Schmitt, L., Flechtenmacher, C., Rtskhiladze, I., & Kubitz, R. (2012). The histidin-loop is essential for transport activity of human MDR3. A novel mutation of MDR3 in a patient with progressive familial intrahepatic cholestasis type 3. *Gene*, *506*(1), 141–145. https://doi.org/10.1016/j.gene.2012.06.029

Eckhardt, E. R., Moschetta, A., Renooij, W., Goerdayal, S. S., van Berge-Henegouwen, G. P., & van Erpecum, K. J. (1999). Asymmetric distribution of phosphatidylcholine and sphingomyelin between micellar and vesicular phases. Potential implications for canalicular bile formation. *Journal of Lipid Research*, *40*(11), 2022–2033. http://www.ncbi.nlm.nih.gov/pubmed/10553006

Elbrecht, A., Chen, Y., Adams, A., Berger, J., Griffin, P., Klatt, T., Zhang, B., Menke, J., Zhou, G., Smith, R. G., & Moller, D. E. (1999). L-764406 Is a Partial Agonist of Human Peroxisome Proliferator-activated Receptor γ. *Journal of Biological Chemistry*, *274*(12), 7913–7922. https://doi.org/10.1074/jbc.274.12.7913

Elferink, R. P. J. O., Tytgat, G. N. J., & Groen, A. K. (1997). The role of mdr2 P-glycoprotein in hepatobiliary lipid transport. *The FASEB Journal*, *11*(1), 19–28. https://doi.org/10.1096/fasebj.11.1.9034162

Eppens, E. F., Van Mil, S. W. C., De Vree, J. M. L., Mok, K. S., Juijn, J. A., Oude Elferink, R. P. J., Berger, R., Houwen, R. H. J., & Klomp, L. W. J. (2001). FIC1, the protein affected in two forms of hereditary cholestasis, is localized in the cholangiocyte and the canalicular membrane of the hepatocyte. *Journal of Hepatology*, *35*(4), 436–443. https://doi.org/10.1016/S0168-8278(01)00158-1

Esposito, D., Weile, J., Shendure, J., Starita, L. M., Papenfuss, A. T., Roth, F. P., Fowler, D. M., & Rubin, A. F. (2019). MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome Biology*, *20*(1), 223. https://doi.org/10.1186/s13059-019-1845-6

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115–118. https://doi.org/10.1038/nature21056

Fang, L. J., Wang, X. H., Knisely, A. S., Yu, H., Lu, Y., Liu, L. Y., & Wang, J. S. (2012). Chinese children with chronic intrahepatic cholestasis and high γ-glutamyl transpeptidase: Clinical features and association with ABCB4 mutations. *Journal of Pediatric Gastroenterology and Nutrition*, *55*(2), 150–156. https://doi.org/10.1097/MPG.0b013e31824ef36f

Fang, S., Suh, J. M., Reilly, S. M., Yu, E., Osborn, O., Lackey, D., Yoshihara, E., Perino, A., Jacinto, S., Lukasheva, Y., Atkins, A. R., Khvat, A., Schnabl, B., Yu, R. T., Brenner, D. A., Coulter, S., Liddle, C., Schoonjans, K., Olefsky, J. M., … Evans, R. M. (2015). Intestinal FXR agonism promotes adipose tissue browning and reduces obesity and insulin resistance. *Nature Medicine*, *21*(2), 159–165. https://doi.org/10.1038/nm.3760

Figueira, A. C. M., Saidemberg, D. M., Souza, P. C. T., Martínez, L., Scanlan, T. S., Baxter, J. D., Skaf, M. S., Palma, M. S., Webb, P., & Polikarpov, I. (2011). Analysis of Agonist and Antagonist Effects on Thyroid Hormone Receptor Conformation by Hydrogen/Deuterium Exchange. *Molecular Endocrinology*, *25*(1), 15–31. https://doi.org/10.1210/me.2010-0202

Finch, A., & Pillans, P. (2014). P-glycoprotein and its role in drug-drug interactions. *Australian Prescriber*, *37*(4), 137–139. https://doi.org/10.18773/austprescr.2014.050

Fiorucci, S., Biagioli, M., Sepe, V., Zampella, A., & Distrutti, E. (2020). Bile acid modulators for the treatment of nonalcoholic steatohepatitis (NASH). *Expert Opinion on Investigational Drugs*, *29*(6), 623–632. https://doi.org/10.1080/13543784.2020.1763302

Fiorucci, S., Biagioli, M., Zampella, A., & Distrutti, E. (2018). Bile Acids Activated Receptors Regulate Innate Immunity. *Frontiers in Immunology*, *9*. https://doi.org/10.3389/fimmu.2018.01853

Fiorucci, S., Zampella, A., Ricci, P., Distrutti, E., & Biagioli, M. (2022). Immunomodulatory functions of FXR. *Molecular and Cellular Endocrinology*, *551*, 111650. https://doi.org/10.1016/j.mce.2022.111650

Floreani, A., Carderi, I., Paternoster, D., Soardo, G., Azzaroli, F., Esposito, W., Montagnani, M., Marchesoni, D., Variola, A., Rosa Rizzotto, E., Braghin, C., & Mazzella, G. (2008). Hepatobiliary phospholipid transporter ABCB4, MDR3 gene variants in a large cohort of Italian women with intrahepatic cholestasis of pregnancy. *Digestive and Liver Disease*, *40*(5), 366–370. https://doi.org/10.1016/j.dld.2007.10.016

Floreani, A., Carderi, I., Paternoster, D., Soardo, G., Azzaroli, F., Esposito, W., Variola, A., Tommasi, A. M., Marchesoni, D., Braghin, C., & Mazzella, G. (2006). Intrahepatic cholestasis of pregnancy: Three novel MDR3 gene mutations. *Alimentary Pharmacology and Therapeutics*, *23*(11), 1649–1653. https://doi.org/10.1111/j.1365-2036.2006.02869.x

Folkertsma, S., Noort, P., Brandt, R., Bettler, E., Vriend, G., & Vlieg, J. (2005). The Nuclear Receptor Ligand-Binding Domain: A Family-Based Structure Analysis. *Current Medicinal Chemistry*, *12*(9), 1001–1016. https://doi.org/10.2174/0929867053764699

Folkman, L., Stantic, B., & Sattar, A. (2013). Sequence-only evolutionary and predicted structural features for the prediction of stability changes in protein mutants. *BMC Bioinformatics*, *14*(S2), S6. https://doi.org/10.1186/1471-2105-14-S2-S6

Forman, B. M., Goode, E., Chen, J., Oro, A. E., Bradley, D. J., Perlmann, T., Noonan, D. J., Burka, L. T., McMorris, T., Lamph, W. W., Evans, R. M., & Weinberger, C. (1995). Identification of a nuclear receptor that is activated by farnesol metabolites. *Cell*, *81*(5), 687–693. https://doi.org/10.1016/0092-8674(95)90530-8

Franzosa, E. A., & Xia, Y. (2009). Structural Determinants of Protein Evolution Are Context-Sensitive at the Residue Level. *Molecular Biology and Evolution*, *26*(10), 2387–2395. https://doi.org/10.1093/molbev/msp146

Frazer, J., Notin, P., Dias, M., Gomez, A., Min, J. K., Brock, K., Gal, Y., & Marks, D. S. (2021). Disease variant prediction with deep generative models of evolutionary data. *Nature*, *599*(7883), 91–95. https://doi.org/10.1038/s41586-021-04043-8

Freund, Y. (1995). Boosting a Weak Learning Algorithm by Majority. *Information and Computation*, *121*(2), 256–285. https://doi.org/10.1006/inco.1995.1136

Frider, B., Castillo, A., Gordo-Gilart, R., Bruno, A., Amante, M., Alvarez, L., & Mathet, V. (2015). Reversal of advanced fibrosis after long-term ursodeoxycholic acid therapy in a patient with residual expression of MDR3. *Annals of Hepatology*, *14*(5), 745–751. https://doi.org/10.1016/s1665-2681(19)30771-9

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5). https://doi.org/10.1214/aos/1013203451

Gadaleta, R. M., Oldenburg, B., Willemsen, E. C. L., Spit, M., Murzilli, S., Salvatore, L., Klomp, L. W. J., Siersema, P. D., van Erpecum, K. J., & van Mil, S. W. C. (2011). Activation of bile salt nuclear receptor FXR is repressed by pro-inflammatory cytokines activating NF-κB signaling in the intestine. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, *1812*(8), 851–858. https://doi.org/10.1016/j.bbadis.2011.04.005

Gaieb, Z., Liu, S., Gathiaka, S., Chiu, M., Yang, H., Shao, C., Feher, V. A., Walters, W. P., Kuhn, B., Rudolph, M. G., Burley, S. K., Gilson, M. K., & Amaro, R. E. (2018). D3R Grand Challenge 2: blind prediction of protein–ligand poses, affinity rankings, and relative binding free energies. *Journal of Computer-Aided Molecular Design*, *32*(1), 1–20. https://doi.org/10.1007/s10822-017-0088-4

Gampe, R. T., Montana, V. G., Lambert, M. H., Miller, A. B., Bledsoe, R. K., Milburn, M. V., Kliewer, S. A., Willson, T. M., & Xu, H. E. (2000). Asymmetry in the PPARγ/RXRα Crystal Structure Reveals the Molecular Basis of Heterodimerization among Nuclear Receptors. *Molecular Cell*, *5*(3), 545–555. https://doi.org/10.1016/S1097-2765(00)80448-7

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., Cerami, E., Sander, C., & Schultz, N. (2013). Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. *Science Signaling*, *6*(269). https://doi.org/10.1126/scisignal.2004088

Gautherot, J., Delautier, D., Maubert, M. A., Aït-Slimane, T., Bolbach, G., Delaunay, J. L., Durand-Schneider, A. M., Firrincieli, D., Barbu, V., Chignard, N., Housset, C., Maurice, M., & Falguières, T. (2014). Phosphorylation of ABCB4 impacts its function: Insights from disease-causing mutations. *Hepatology*, *60*(2), 610–621. https://doi.org/10.1002/hep.27170

Gauthier, J., Vincent, A. T., Charette, S. J., & Derome, N. (2019). A brief history of bioinformatics. *Briefings in Bioinformatics*, *20*(6), 1981–1996. https://doi.org/10.1093/bib/bby063

Genin, M. J., Bueno, A. B., Agejas Francisco, J., Manninen, P. R., Bocchinfuso, W. P., Montrose-Rafizadeh, C., Cannady, E. A., Jones, T. M., Stille, J. R., Raddad, E., Reidy, C., Cox, A., Michael, M. D., & Michael, L. F. (2015). Discovery of 6-(4-{[5-Cyclopropyl-3-(2,6-dichlorophenyl)isoxazol-4-yl]methoxy}piperidin-1-yl)-1-methyl-1 H -indole-3-carboxylic Acid: A Novel FXR Agonist for the Treatment of Dyslipidemia. *Journal of Medicinal*

Chemistry, *58*(24), 9768–9772. https://doi.org/10.1021/acs.jmedchem.5b01161

Gerloff, T., Geier, A., Roots, I., Meier, P. J., & Gartung, C. (2002). Functional analysis of the rat bile salt export pump gene promoter. *European Journal of Biochemistry*, *269*(14), 3495–3503. https://doi.org/10.1046/j.1432-1033.2002.03030.x

Gerloff, T., Stieger, B., Hagenbuch, B., Madon, J., Landmann, L., Roth, J., Hofmann, A. F., & Meier, P. J. (1998). The Sister of P-glycoprotein Represents the Canalicular Bile Salt Export Pump of Mammalian Liver. *Journal of Biological Chemistry*, *273*(16), 10046–10050. https://doi.org/10.1074/jbc.273.16.10046

Girisa, S., Henamayee, S., Parama, D., Rana, V., Dutta, U., & Kunnumakkara, A. B. (2021). Targeting Farnesoid X receptor (FXR) for developing novel therapeutics against cancer. *Molecular Biomedicine*, *2*(1), 21. https://doi.org/10.1186/s43556-021-00035-2

Goate, A. (2006). Segregation of a missense mutation in the amyloid β-protein precursor gene with familial Alzheimer's disease. *Journal of Alzheimer's Disease*, *9*(s3), 341–347. https://doi.org/10.3233/JAD-2006-9S338

Godlewska, U., Bulanda, E., & Wypych, T. P. (2022). Bile acids in immunity: Bidirectional mediators between the host and the microbiota. *Frontiers in Immunology*, *13*. https://doi.org/10.3389/fimmu.2022.949033

Gomez-Ospina, N., Potter, C. J., Xiao, R., Manickam, K., Kim, M.-S., Kim, K. H., Shneider, B. L., Picarsic, J. L., Jacobson, T. A., Zhang, J., He, W., Liu, P., Knisely, A. S., Finegold, M. J., Muzny, D. M., Boerwinkle, E., Lupski, J. R., Plon, S. E., Gibbs, R. A., … Moore, D. D. (2016). Mutations in the nuclear bile acid receptor FXR cause progressive familial intrahepatic cholestasis. *Nature Communications*, *7*(1), 10713. https://doi.org/10.1038/ncomms10713

Gonzales, E., Taylor, S. A., Davit-Spraul, A., Thébaut, A., Thomassin, N., Guettier, C., Whitington, P. F., & Jacquemin, E. (2017). MYO5B mutations cause cholestasis with normal serum gamma-glutamyl transferase activity in children without microvillous inclusion disease. *Hepatology*, *65*(1), 164–173. https://doi.org/10.1002/hep.28779

Goodwin, B., Jones, S. A., Price, R. R., Watson, M. A., McKee, D. D., Moore, L. B., Galardi, C., Wilson, J. G., Lewis, M. C., Roth, M. E., Maloney, P. R., Willson, T. M., & Kliewer, S. A. (2000). A Regulatory Cascade of the Nuclear Receptors FXR, SHP-1, and LRH-1 Represses Bile Acid Biosynthesis. *Molecular Cell*, *6*(3), 517–526. https://doi.org/10.1016/S1097-2765(00)00051-4

Gordo-Gilart, R., Andueza, S., Hierro, L., Martínez-Fernández, P., D'Agostino, D., Jara, P., & Alvarez, L. (2015). Functional analysis of ABCB4 mutations relates clinical outcomes of progressive familial intrahepatic cholestasis type 3 to the degree of MDR3 floppase activity. *Gut*, *64*(1), 147–155. https://doi.org/10.1136/gutjnl-2014-306896

Gordo-Gilart, R., Hierro, L., Andueza, S., Muñoz-Bartolo, G., López, C., Díaz, C., Jara, P., & Álvarez, L. (2016). Heterozygous ABCB4 mutations in children with cholestatic liver disease. *Liver International*, *36*(2), 258–267. https://doi.org/10.1111/liv.12910

Gotthardt, D., Runz, H., Keitel, V., Fischer, C., Flechtenmacher, C., Wirtenberger, M., Weiss, K. H., Imparato, S., Braun, A., Hemminki, K., Stremmel, W., Rüschendorf, F., Stiehl, A., Kubitz, R., Burwinkel, B., Schirmacher, P., Knisely, A. S., Zschocke, J., & Sauer, P. (2008). A mutation in the canalicular phospholipid transporter gene, ABCB4, is associated with cholestasis, ductopenia, and cirrhosis in adults. *Hepatology*, *48*(4), 1157–1166. https://doi.org/10.1002/hep.22485

Graf, G. A., Yu, L., Li, W.-P., Gerard, R., Tuma, P. L., Cohen, J. C., & Hobbs, H. H. (2003). ABCG5 and ABCG8 Are Obligate Heterodimers for Protein Trafficking and Biliary Cholesterol Excretion. *Journal of Biological Chemistry*, *278*(48), 48275–48282. https://doi.org/10.1074/jbc.M310223200

Greener, J. G., Kandathil, S. M., Moffat, L., & Jones, D. T. (2022). A guide to machine learning for biologists. *Nature Reviews Molecular Cell Biology*, *23*(1), 40–55. https://doi.org/10.1038/s41580-021-00407-0

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? *NeurIPS 2022 Datasets and Benchmarks*.

Grober, J., Zaghini, I., Fujii, H., Jones, S. A., Kliewer, S. A., Willson, T. M., Ono, T., & Besnard, P. (1999). Identification of a Bile Acid-responsive Element in the Human Ileal Bile Acid-binding Protein Gene. *Journal of Biological Chemistry*, *274*(42), 29749–29754. https://doi.org/10.1074/jbc.274.42.29749

Gudbjartsson, D. F., Helgason, H., Gudjonsson, S. A., Zink, F., Oddson, A., Gylfason, A., Besenbacher, S., Magnusson, G., Halldorsson, B. V, Hjartarson, E., Sigurdsson, G. T., Stacey, S. N., Frigge, M. L., Holm, H., Saemundsdottir, J., Helgadottir, H. T., Johannsdottir, H., Sigfusson, G., Thorgeirsson, G., … Stefansson, K. (2015). Large-scale whole-genome sequencing of the Icelandic population. *Nature Genetics*, *47*(5), 435–444. https://doi.org/10.1038/ng.3247

Gunning, A. C., Fryer, V., Fasham, J., Crosby, A. H., Ellard, S., Baple, E. L., & Wright, C. F. (2021). Assessing performance of pathogenicity predictors using clinically relevant variant datasets. *Journal of Medical Genetics*, *58*(8), 547–555. https://doi.org/10.1136/jmedgenet-2020-107003

Guo, C., LaCerte, C., Edwards, J. E., Brouwer, K. R., & Brouwer, K. L. R. (2018). Farnesoid X Receptor Agonists Obeticholic Acid and Chenodeoxycholic Acid Increase Bile Acid Efflux in Sandwich-Cultured Human Hepatocytes: Functional Evidence and Mechanisms. *Journal of Pharmacology and Experimental Therapeutics*, *365*(2), 413–421. https://doi.org/10.1124/jpet.117.246033

Gustafsson, B. E., Gustafsson, J.-Å., Sjövall, J., Bowie, J. H., Williams, D. H., Bunnenberg, E., Djerassi, C., & Records, R. (1966). Intestinal and Fecal Sterols in Germfree and Conventional Rats. Bile Acids and Steroids 172. *Acta Chemica Scandinavica*, *20*, 1827–1835. https://doi.org/10.3891/acta.chem.scand.20-1827

Guzior, D. V., & Quinn, R. A. (2021). Review: microbial transformations of human bile acids. *Microbiome*, *9*(1), 1–13. https://doi.org/10.1186/s40168-021-01101-1

Halder, K., Dölker, N., Van, Q., Gregor, I., Dickmanns, A., Baade, I., Kehlenbach, R. H., Ficner, R., Enderlein, J., Grubmüller, H., & Neumann, H. (2015). MD Simulations and FRET Reveal an Environment-Sensitive Conformational Plasticity of Importin-β. *Biophysical Journal*, *109*(2), 277–286. https://doi.org/10.1016/j.bpj.2015.06.014

Halilbasic, E., Claudel, T., & Trauner, M. (2013). Bile acid transporters and regulatory nuclear receptors in the liver and beyond. *Journal of Hepatology*, *58*(1), 155–168. https://doi.org/10.1016/j.jhep.2012.08.002

Hamelryck, T. (2005). An amino acid has two sides: A new 2D measure provides a different view of solvent exposure. *Proteins: Structure, Function and Genetics*, *59*(1), 38–48. https://doi.org/10.1002/prot.20379

Han, C. (2018). Update on FXR Biology: Promising Therapeutic Target? *International Journal of Molecular Sciences*, *19*(7), 2069. https://doi.org/10.3390/ijms19072069

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques.* Elsevier.

Hariharan, P. C., & Pople, J. A. (1972). The effect of d-functions on molecular orbital energies for hydrocarbons. *Chemical Physics Letters*, *16*(2), 217–219. https://doi.org/10.1016/0009-2614(72)80259-8

Hasle, N., Matreyek, K. A., & Fowler, D. M. (2019). The Impact of Genetic Variants on PTEN Molecular Functions and Cellular Phenotypes. *Cold Spring Harbor Perspectives in Medicine*, *9*(11), a036228. https://doi.org/10.1101/cshperspect.a036228

Heery, D. M., Kalkhoven, E., Hoare, S., & Parker, M. G. (1997). A signature motif in transcriptional co-activators mediates binding to nuclear receptors. *Nature*, *387*(6634), 733–736. https://doi.org/10.1038/42750

Heidari, Z., Chrisman, I. M., Nemetchek, M. D., Novick, S. J., Blayo, A.-L., Patton, T., Mendes, D. E., Diaz, P., Kamenecka, T. M., Griffin, P. R., & Hughes, T. S. (2019). Definition of functionally and structurally distinct repressive states in the nuclear receptor PPARγ. *Nature Communications*, *10*(1), 5825. https://doi.org/10.1038/s41467-019-13768-0

Heil, B. J., Hoffman, M. M., Markowetz, F., Lee, S.-I., Greene, C. S., & Hicks, S. C. (2021). Reproducibility standards for machine learning in the life sciences. *Nature Methods*, *18*(10), 1132–1135. https://doi.org/10.1038/s41592-021-01256-7

Heyman, R. A., Mangelsdorf, D. J., Dyck, J. A., Stein, R. B., Eichele, G., Evans, R. M., & Thaller, C. (1992). 9-cis retinoic acid is a high affinity ligand for the retinoid X receptor. *Cell*, *68*(2), 397–406. https://doi.org/10.1016/0092-8674(92)90479-V

Hillman, E. T., Lu, H., Yao, T., & Nakatsu, C. H. (2017). Microbial Ecology along the Gastrointestinal Tract. *Microbes and Environments*, *32*(4), 300–313. https://doi.org/10.1264/jsme2.ME17017

Hoeke, M. O., Plass, J. R. M., Heegsma, J., Geuken, M., van Rijsbergen, D., Baller, J. F. W., Kuipers, F., Moshage, H., Jansen, P. L. M., & Faber, K. N. (2009). Low retinol levels differentially modulate bile salt-induced expression of human and mouse hepatic bile salt transporters. *Hepatology*, *49*(1), 151–159. https://doi.org/10.1002/hep.22661

Hofmann, Alan, F. (2009). The enterohepatic circulation of bile acids in mammals: form and functions. *Frontiers in Bioscience*, *Volume*(14), 2584. https://doi.org/10.2741/3399

Hofmann, A. F. (1976). The enterohepatic circulation of bile acids in man. *Advances in Internal Medicine*, *21*, 501–534. http://www.ncbi.nlm.nih.gov/pubmed/766594

Hollingsworth, S. A., & Dror, R. O. (2018). Molecular Dynamics Simulation for All. *Neuron*, *99*(6), 1129–1143. https://doi.org/10.1016/j.neuron.2018.08.011

Holt, J. A., Luo, G., Billin, A. N., Bisi, J., McNeill, Y. Y., Kozarsky, K. F., Donahee, M., Wang, D. Y., Mansfield, T. A., Kliewer, S. A., Goodwin, B., & Jones, S. A. (2003). Definition of a novel growth factor-dependent signal cascade for the suppression of bile acid biosynthesis. *Genes & Development*, *17*(13), 1581–1591. https://doi.org/10.1101/gad.1083503

Hopf, C., Beuers, U., Bikker, H., Denk, G. U., & Rust, C. (2011). 44-jährige Patientin mit unklarer Leberwerterhöhung und familiär gehäuftem Gallensteinleiden. *Internist*, *52*(10), 1234–1237. https://doi.org/10.1007/s00108-010-2775-2

Hopf, T. A., Ingraham, J. B., Poelwijk, F. J., Schärfe, C. P. I., Springer, M., Sander, C., & Marks, D. S. (2017). Mutation effects predicted from sequence co-variation. *Nature Biotechnology*, *35*(2), 128–135.

https://doi.org/10.1038/nbt.3769

Horn, A. H. C. (2003). Essentials of Computational Chemistry, Theories and Models By Christopher J. Cramer. Wiley: Chichester, England. 2002. 562 pp. ISBN 0-471-48551-9 (hardcover). $110. ISBN 0-471-48552-7 (paperback). $45. In *Journal of Chemical Information and Computer Sciences* (Vol. 43, Issue 5). https://doi.org/10.1021/ci010445m

Hornbeck, P. V., Zhang, B., Murray, B., Kornhauser, J. M., Latham, V., & Skrzypek, E. (2015). PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Research*, *43*(D1), D512–D520. https://doi.org/10.1093/nar/gku1267

Hsu, C.-W., Hsieh, J.-H., Huang, R., Pijnenburg, D., Khuc, T., Hamm, J., Zhao, J., Lynch, C., van Beuningen, R., Chang, X., Houtman, R., & Xia, M. (2016). Differential modulation of FXR activity by chlorophacinone and ivermectin analogs. *Toxicology and Applied Pharmacology*, *313*, 138–148. https://doi.org/10.1016/j.taap.2016.10.017

Huang, C., Wang, J., Hu, W., Wang, C., Lu, X., Tong, L., Wu, F., & Zhang, W. (2016). Identification of functional farnesoid X receptors in brain neurons. *FEBS Letters*, *590*(18), 3233–3242. https://doi.org/10.1002/1873-3468.12373

Huang, L., Zhao, A., Lew, J.-L., Zhang, T., Hrywna, Y., Thompson, J. R., de Pedro, N., Royo, I., Blevins, R. A., Peláez, F., Wright, S. D., & Cui, J. (2003). Farnesoid X Receptor Activates Transcription of the Phospholipid Pump MDR3. *Journal of Biological Chemistry*, *278*(51), 51085–51090. https://doi.org/10.1074/jbc.M308321200

Huang, W., Ma, K., Zhang, J., Qatanani, M., Cuvillier, J., Liu, J., Dong, B., Huang, X., & Moore, D. D. (2006). Nuclear Receptor-Dependent Bile Acid Signaling Is Required for Normal Liver Regeneration. *Science*, *312*(5771), 233–236. https://doi.org/10.1126/science.1121435

Huber, R. M., Murphy, K., Miao, B., Link, J. R., Cunningham, M. R., Rupar, M. J., Gunyuzlu, P. L., Haws, T. F., Kassam, A., Powell, F., Hollis, G. F., Young, P. R., Mukherjee, R., & Burn, T. C. (2002). Generation of multiple farnesoid-X-receptor isoforms through the use of alternative promoters. *Gene*, *290*(1–2), 35–43. https://doi.org/10.1016/S0378-1119(02)00557-7

Hughes, T. S., Chalmers, M. J., Novick, S., Kuruvilla, D. S., Chang, M. R., Kamenecka, T. M., Rance, M., Johnson, B. A., Burris, T. P., Griffin, P. R., & Kojetin, D. J. (2012). Ligand and Receptor Dynamics Contribute to the Mechanism of Graded PPARγ Agonism. *Structure*, *20*(1), 139–150. https://doi.org/10.1016/j.str.2011.10.018

Ijssennagger, N., Janssen, A. W. F., Milona, A., Ramos Pittol, J. M., Hollman, D. A. A., Mokry, M., Betzel, B., Berends, F. J., Janssen, I. M., van Mil, S. W. C., & Kersten, S. (2016). Gene expression profiling in human precision cut liver slices in response to the FXR agonist obeticholic acid. *Journal of Hepatology*, *64*(5), 1158–1166. https://doi.org/10.1016/j.jhep.2016.01.016

Ikeda, Y., Morita, S., & Terada, T. (2017). Cholesterol attenuates cytoprotective effects of phosphatidylcholine against bile salts. *Scientific Reports*, *7*(1), 306. https://doi.org/10.1038/s41598-017-00476-2

Inagaki, T., Choi, M., Moschetta, A., Peng, L., Cummins, C. L., McDonald, J. G., Luo, G., Jones, S. A., Goodwin, B., Richardson, J. A., Gerard, R. D., Repa, J. J., Mangelsdorf, D. J., & Kliewer, S. A. (2005). Fibroblast growth factor 15 functions as an enterohepatic signal to regulate bile acid homeostasis. *Cell Metabolism*, *2*(4), 217–225. https://doi.org/10.1016/j.cmet.2005.09.001

Ioannidis, N. M., Rothstein, J. H., Pejaver, V., Middha, S., McDonnell, S. K., Baheti, S., Musolf, A., Li, Q., Holzinger, E., Karyadi, D., Cannon-Albright, L. A., Teerlink, C. C., Stanford, J. L., Isaacs, W. B., Xu, J., Cooney, K. A., Lange, E. M., Schleutker, J., Carpten, J. D., … Sieh, W. (2016). REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *The American Journal of Human Genetics*, *99*(4), 877–885. https://doi.org/10.1016/j.ajhg.2016.08.016

Iqbal, M. J., Javed, Z., Sadia, H., Qureshi, I. A., Irshad, A., Ahmed, R., Malik, K., Raza, S., Abbas, A., Pezzani, R., & Sharifi-Rad, J. (2021). Clinical applications of artificial intelligence and machine learning in cancer diagnosis: looking into the future. *Cancer Cell International*, *21*(1), 270. https://doi.org/10.1186/s12935-021-01981-1

Ittisoponpisan, S., Islam, S. A., Khanna, T., Alhuzimi, E., David, A., & Sternberg, M. J. E. (2019). Can Predicted Protein 3D Structures Provide Reliable Insights into whether Missense Variants Are Disease Associated? *Journal of Molecular Biology*, *431*(11), 2197–2212. https://doi.org/10.1016/j.jmb.2019.04.009

Jacquemin, E., DeVree, J. M. L., Cresteil, D., Sokal, E. M., Sturm, E., Dumont, M., Scheffer, G. L., Paul, M., Burdelski, M., Bosma, P. J., Bernard, O., Hadchouel, M., & Oude Elferink, R. P. J. (2001). The wide spectrum of multidrug resistance 3 deficiency: From neonatal cholestasis to cirrhosis of adulthood. *Gastroenterology*, *120*(6), 1448–1458. https://doi.org/10.1053/gast.2001.23984

Jagadeesh, K. A., Wenger, A. M., Berger, M. J., Guturu, H., Stenson, P. D., Cooper, D. N., Bernstein, J. A., & Bejerano, G. (2016). M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nature Genetics*, *48*(12), 1581–1586. https://doi.org/10.1038/ng.3703

Jedlitschky, G, Leier, I., Buchholz, U., Hummel-Eisenbeiss, J., Burchell, B., & Keppler, D. (1997). ATP-dependent transport of bilirubin glucuronides by the multidrug resistance protein MRP1 and its hepatocyte canalicular isoform MRP2. *The Biochemical Journal*, *327 (Pt 1*(Pt 1), 305–310. https://doi.org/10.1042/bj3270305

Jedlitschky, Gabriele, Hoffmann, U., & Kroemer, H. K. (2006). Structure and function of the MRP2 (ABCC2) protein and its role in drug disposition. *Expert Opinion on Drug Metabolism & Toxicology*, *2*(3), 351–366. https://doi.org/10.1517/17425255.2.3.351

Jia, W., Sun, M., Lian, J., & Hou, S. (2022). Feature dimensionality reduction: a review. *Complex & Intelligent Systems*, *8*(3), 2663–2693. https://doi.org/10.1007/s40747-021-00637-x

Jiang, L., Zhang, H., Xiao, D., Wei, H., & Chen, Y. (2021). Farnesoid X receptor (FXR): Structures and ligands. *Computational and Structural Biotechnology Journal*, *19*, 2148–2159. https://doi.org/10.1016/j.csbj.2021.04.029

Jiao, Y., Lu, Y., & Li, X. (2015). Farnesoid X receptor: a master regulator of hepatic triglyceride and glucose homeostasis. *Acta Pharmacologica Sinica*, *36*(1), 44–50. https://doi.org/10.1038/aps.2014.116

Jin, L., Feng, X., Rong, H., Pan, Z., Inaba, Y., Qiu, L., Zheng, W., Lin, S., Wang, R., Wang, Z., Wang, S., Liu, H., Li, S., Xie, W., & Li, Y. (2013). The antiparasitic drug ivermectin is a novel FXR ligand that regulates metabolism. *Nature Communications*, *4*(1), 1937. https://doi.org/10.1038/ncomms2924

Jin, L., Wang, R., Zhu, Y., Zheng, W., Han, Y., Guo, F., Ye, F. Bin, & Li, Y. (2015). Selective targeting of nuclear receptor FXR by avermectin analogues with therapeutic effects on nonalcoholic fatty liver disease. *Scientific Reports*, *5*(1), 17288. https://doi.org/10.1038/srep17288

John, G. H., & Langley, P. (1995). Estimating continuous distributions in bayesian classifiers. In *Proccedings of the Eleventh conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc.

Joosten, R. P., te Beek, T. A. H., Krieger, E., Hekkelman, M. L., Hooft, R. W. W., Schneider, R., Sander, C., & Vriend, G. (2011). A series of PDB related databases for everyday needs. *Nucleic Acids Research*, *39*(Database), D411–D419. https://doi.org/10.1093/nar/gkq1105

Joseph, V. R. (2022). Optimal ratio for data splitting. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *15*(4), 531–538. https://doi.org/10.1002/sam.11583

Jovel, J., & Greiner, R. (2021). An Introduction to Machine Learning Approaches for Biomedical Research. *Frontiers in Medicine*, *8*. https://doi.org/10.3389/fmed.2021.771607

Jung, D., Elferink, M. G. L., Stellaard, F., & Groothuis, G. M. M. (2007). Analysis of bile acid-induced regulation of FXR target genes in human liver slices. *Liver International*, *27*(1). https://doi.org/10.1111/j.1478-3231.2006.01393.x

Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, *22*(12), 2577–2637. https://doi.org/10.1002/bip.360221211

Kainuma, M., Takada, I., Makishima, M., & Sano, K. (2018). Farnesoid X Receptor Activation Enhances Transforming Growth Factor β-Induced Epithelial-Mesenchymal Transition in Hepatocellular Carcinoma Cells. *International Journal of Molecular Sciences*, *19*(7), 1898. https://doi.org/10.3390/ijms19071898

Kallenberger, B. C., Love, J. D., Chatterjee, V. K. K., & Schwabe, J. W. R. (2003). A dynamic mechanism of nuclear receptor activation and its perturbation in a human disease. *Nature Structural Biology*, *10*(2), 136–140. https://doi.org/10.1038/nsb892

Kamatani, T., Fukunaga, K., Miyata, K., Shirasaki, Y., Tanaka, J., Baba, R., Matsusaka, M., Kamatani, N., Moro, K., Betsuyaku, T., & Uemura, S. (2017). Construction of a system using a deep learning algorithm to count cell numbers in nanoliter wells for viable single-cell experiments. *Scientific Reports*, *7*(1), 16831. https://doi.org/10.1038/s41598-017-17012-x

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., … MacArthur, D. G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, *581*(7809), 434–443. https://doi.org/10.1038/s41586-020-2308-7

Karplus, M., & McCammon, J. A. (2002). Molecular dynamics simulations of biomolecules. *Nature Structural Biology*, *9*(9), 646–652. https://doi.org/10.1038/nsb0902-646

Katafuchi, T., & Makishima, M. (2022). Molecular Basis of Bile Acid-FXR-FGF15/19 Signaling Axis. *International Journal of Molecular Sciences*, *23*(11), 6046. https://doi.org/10.3390/ijms23116046

Kawamata, Y., Fujii, R., Hosoya, M., Harada, M., Yoshida, H., Miwa, M., Fukusumi, S., Habata, Y., Itoh, T., Shintani, Y., Hinuma, S., Fujisawa, Y., & Fujino, M. (2003). A G Protein-coupled Receptor Responsive to Bile Acids. *Journal of Biological Chemistry*, *278*(11), 9435–9440. https://doi.org/10.1074/jbc.M209706200

Keerthi, S. S., Shevade, S. K., Bhattacharyya, C., & Murthy, K. R. K. (2001). Improvements to Platt's SMO Algorithm for SVM Classifier Design. *Neural Computation*, *13*(3), 637–649.

https://doi.org/10.1162/089976601300014493

Keitel, V., Burdelski, M., Warskulat, U., Kühlkamp, T., Keppler, D., Häussinger, D., & Kubitz, R. (2005). Expression and localization of hepatobiliary transport proteins in progressive familial intrahepatic cholestasis. *Hepatology*, *41*(5), 1160–1172. https://doi.org/10.1002/hep.20682

Keitel, V., Dröge, C., Stepanow, S., Fehm, T., Mayatepek, E., Köhrer, K., & Häussinger, D. (2016). Intrahepatic cholestasis of pregnancy (ICP): case report and review of the literature. *Zeitschrift Fur Gastroenterologie*, *54*(12), 1327–1333. https://doi.org/10.1055/s-0042-118388

Keitel, V., Vogt, C., Häussinger, D., & Kubitz, R. (2006). Combined Mutations of Canalicular Transporter Proteins Cause Severe Intrahepatic Cholestasis of Pregnancy. *Gastroenterology*, *131*(2), 624–629. https://doi.org/10.1053/j.gastro.2006.05.003

Khabou, B., Durand-Schneider, A. M., Delaunay, J. L., Aït-Slimane, T., Barbu, V., Fakhfakh, F., Housset, C., & Maurice, M. (2017). Comparison of in silico prediction and experimental assessment of ABCB4 variants identified in patients with biliary diseases. *International Journal of Biochemistry and Cell Biology*, *89*(May), 101–109. https://doi.org/10.1016/j.biocel.2017.05.028

Khan, S. H., Braet, S. M., Koehler, S. J., Elacqua, E., Anand, G. S., & Okafor, C. D. (2022). Ligand-induced shifts in conformational ensembles that describe transcriptional activation. *ELife*, *11*. https://doi.org/10.7554/eLife.80140

Kim, I., Ahn, S.-H., Inagaki, T., Choi, M., Ito, S., Guo, G. L., Kliewer, S. A., & Gonzalez, F. J. (2007). Differential regulation of bile acid homeostasis by the farnesoid X receptor in liver and intestine. *Journal of Lipid Research*, *48*(12), 2664–2672. https://doi.org/10.1194/jlr.M700330-JLR200

Kipp, H., & Arias, I. M. (2000). Newly Synthesized Canalicular ABC Transporters Are Directly Targeted from the Golgi to the Hepatocyte Apical Domain in Rat Liver. *Journal of Biological Chemistry*, *275*(21), 15917–15925. https://doi.org/10.1074/jbc.M909875199

Kirschner, K. N., Yongye, A. B., Tschampel, S. M., González-Outeiriño, J., Daniels, C. R., Foley, B. L., & Woods, R. J. (2008). GLYCAM06: A generalizable biomolecular force field. Carbohydrates. *Journal of Computational Chemistry*, *29*(4), 622–655. https://doi.org/10.1002/jcc.20820

Klar, T. A., Jakobs, S., Dyba, M., Egner, A., & Hell, S. W. (2000). Fluorescence microscopy with diffraction resolution barrier broken by stimulated emission. *Proceedings of the National Academy of Sciences*, *97*(15), 8206–8210. https://doi.org/10.1073/pnas.97.15.8206

Kluth, M., Stindt, J., Dröge, C., Linnemann, D., Kubitz, R., & Schmitt, L. (2015). A mutation within the extended X loop abolished substrateinduced ATPase activity of the human liver ATP-binding cassette (ABC) transporter MDR3. *Journal of Biological Chemistry*, *290*(8), 4896–4907. https://doi.org/10.1074/jbc.M114.588566

Knell, A. J. (1980). Liver function and failure: the evolution of liver physiology. *Journal of the Royal College of Physicians of London*, *14*(3), 205–208. http://www.ncbi.nlm.nih.gov/pubmed/7009850

Koch, A., Bonus, M., Gohlke, H., & Klöcker, N. (2019). Isoform-specific Inhibition of N-methyl-D-aspartate Receptors by Bile Salts. *Scientific Reports*, *9*(1), 1–17. https://doi.org/10.1038/s41598-019-46496-y

Koes, D. R., & Vries, J. K. (2017). Evaluating amber force fields using computed NMR chemical shifts. *Proteins: Structure, Function, and Bioinformatics*, *85*(10), 1944–1956. https://doi.org/10.1002/prot.25350

Kopanos, C., Tsiolkas, V., Kouris, A., Chapple, C. E., Albarca Aguilera, M., Meyer, R., & Massouras, A. (2019). VarSome: the human genomic variant search engine. *Bioinformatics*, *35*(11), 1978–1980. https://doi.org/10.1093/bioinformatics/bty897

Kroker, A. J., & Bruning, J. B. (2015). Review of the Structural and Dynamic Mechanisms of PPAR γ Partial Agonism. *PPAR Research*, *2015*, 1–15. https://doi.org/10.1155/2015/816856

Kubitz, R., Bode, J., Erhardt, A., Graf, D., Kircheis, G., Müller-Stöver, I., Reinehr, R., Reuter, S., Richter, J., Sagir, A., Schmitt, M., & Donner, M. (2011). Cholestatic liver diseases from child to adult: The diversity of MDR3 disease. *Zeitschrift Fur Gastroenterologie*, *49*(6), 728–736. https://doi.org/10.1055/s-0031-1273427

Kumar, M., Gouw, M., Michael, S., Sámano-Sánchez, H., Pancsa, R., Glavina, J., Diakogianni, A., Valverde, J. A., Bukirova, D., Čalyševa, J., Palopoli, N., Davey, N. E., Chemes, L. B., & Gibson, T. J. (2019). ELM—the eukaryotic linear motif resource in 2020. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkz1030

Kumar, R., & Thompson, E. B. (1999). The structure of the nuclear hormone receptors. *Steroids*, *64*(5), 310–319. https://doi.org/10.1016/S0039-128X(99)00014-8

Kumari, A., Mittal, L., Srivastava, M., Pathak, D. P., & Asthana, S. (2021). Conformational Characterization of the Co-Activator Binding Site Revealed the Mechanism to Achieve the Bioactive State of FXR. *Frontiers in Molecular Biosciences*, *8*(August), 1–21. https://doi.org/10.3389/fmolb.2021.658312

Kumari, A., Mittal, L., Srivastava, M., Pathak, D. P., & Asthana, S. (2023). Deciphering the Structural Determinants Critical in Attaining the FXR Partial Agonism. *Journal of Physical Chemistry B*, *127*(2), 465–485. https://doi.org/10.1021/acs.jpcb.2c06325

L. Mercer, S., & Coop, A. (2011). Opioid Analgesics and P-Glycoprotein Efflux Transporters: A Potential Systems-

Level Contribution to Analgesic Tolerance. *Current Topics in Medicinal Chemistry*, *11*(9), 1157–1164. https://doi.org/10.2174/156802611795371288

Laffitte, B. A., Kast, H. R., Nguyen, C. M., Zavacki, A. M., Moore, D. D., & Edwards, P. A. (2000). Identification of the DNA Binding Specificity and Potential Target Genes for the Farnesoid X-activated Receptor. *Journal of Biological Chemistry*, *275*(14), 10638–10647. https://doi.org/10.1074/jbc.275.14.10638

Laimer, J., Hofer, H., Fritz, M., Wegenkittl, S., & Lackner, P. (2015). MAESTRO - multi agent stability prediction upon point mutations. *BMC Bioinformatics*, *16*(1). https://doi.org/10.1186/s12859-015-0548-6

Lammert, F., Wang, D. Q.-H., Hillebrandt, S., Geier, A., Fickert, P., Trauner, M., Matern, S., Paigen, B., & Carey, M. C. (2004). Spontaneous cholecysto- and hepatolithiasis in Mdr2-/- mice: A model for low phospholipid-associated cholelithiasis. *Hepatology*, *39*(1), 117–128. https://doi.org/10.1002/hep.20022

Landrier, J.-F., Eloranta, J. J., Vavricka, S. R., & Kullak-Ublick, G. A. (2006). The nuclear receptor for bile acids, FXR, transactivates human organic solute transporter-α and -β genes. *American Journal of Physiology-Gastrointestinal and Liver Physiology*, *290*(3), G476–G485. https://doi.org/10.1152/ajpgi.00430.2005

Lang, C., Meier, Y., Stieger, B., Beuers, U., Lang, T., Kerb, R., Kullak-Ublick, G. A., Meier, P. J., & Pauli-Magnus, C. (2007). Mutations and polymorphisms in the bile salt export pump and the multidrug resistance protein 3 associated with drug-induced liver injury. In *Pharmacogenetics and Genomics* (Vol. 17). Lippincott Williams & Wilkins.

Latorraca, N. R., Fastman, N. M., Venkatakrishnan, A. J., Frommer, W. B., Dror, R. O., & Feng, L. (2017). Mechanism of Substrate Translocation in an Alternating Access Transporter. *Cell*, *169*(1), 96-107.e12. https://doi.org/10.1016/j.cell.2017.03.010

Leboeuf, J. S., LeBlanc, F., & Marchand, M. (2020). Decision trees as partitioning machines to characterize their generalization properties. *Advances in Neural Information Processing Systems*, *2020-Decem*(NeurIPS 2020).

Lee, F. Y., Lee, H., Hubbert, M. L., Edwards, P. A., & Zhang, Y. (2006). FXR, a multipurpose nuclear receptor. *Trends in Biochemical Sciences*, *31*(10), 572–580. https://doi.org/10.1016/j.tibs.2006.08.002

Levitt, M., & Warshel, A. (1975). Computer simulation of protein folding. *Nature*, *253*(5494), 694–698. https://doi.org/10.1038/253694a0

Li, B., Krishnan, V. G., Mort, M. E., Xin, F., Kamati, K. K., Cooper, D. N., Mooney, S. D., & Radivojac, P. (2009). Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics*, *25*(21), 2744–2750. https://doi.org/10.1093/bioinformatics/btp528

Li, M., Cai, S.-Y., & Boyer, J. L. (2017). Mechanisms of bile acid mediated inflammation in the liver. *Molecular Aspects of Medicine*, *56*, 45–53. https://doi.org/10.1016/j.mam.2017.06.001

Li, W., Fu, J., Cheng, F., Zheng, M., Zhang, J., Liu, G., & Tang, Y. (2012). Unbinding Pathways of GW4064 from Human Farnesoid X Receptor As Revealed by Molecular Dynamics Simulations. *Journal of Chemical Information and Modeling*, *52*(11), 3043–3052. https://doi.org/10.1021/ci300459k

Lieu, T., Jayaweera, G., Zhao, P., Poole, D. P., Jensen, D., Grace, M., McIntyre, P., Bron, R., Wilson, Y. M., Krappitz, M., Haerteis, S., Korbmacher, C., Steinhoff, M. S., Nassini, R., Materazzi, S., Geppetti, P., Corvera, C. U., & Bunnett, N. W. (2014). The bile acid receptor TGR5 activates the TRPA1 channel to induce itch in mice. *Gastroenterology*, *147*(6), 1417–1428. https://doi.org/10.1053/j.gastro.2014.08.042

Liu, Y., Chen, P.-H. C., Krause, J., & Peng, L. (2019). How to Read Articles That Use Machine Learning. *JAMA*, *322*(18), 1806. https://doi.org/10.1001/jama.2019.16489

Livesey, B. J., & Marsh, J. A. (2023). Updated benchmarking of variant effect predictors using deep mutational scanning. *Molecular Systems Biology*, *19*(8). https://doi.org/10.15252/msb.202211474

Lo Vercio, L., Amador, K., Bannister, J. J., Crites, S., Gutierrez, A., MacDonald, M. E., Moore, J., Mouches, P., Rajashekar, D., Schimert, S., Subbanna, N., Tuladhar, A., Wang, N., Wilms, M., Winder, A., & Forkert, N. D. (2020). Supervised machine learning tools: a tutorial for clinicians. *Journal of Neural Engineering*, *17*(6), 062001. https://doi.org/10.1088/1741-2552/abbff2

Love, O., Galindo-Murillo, R., Zgarbová, M., Šponer, J., Jurečka, P., & Cheatham, T. E. (2023). Assessing the Current State of Amber Force Field Modifications for DNA–2023 Edition. *Journal of Chemical Theory and Computation*, *19*(13), 4299–4307. https://doi.org/10.1021/acs.jctc.3c00233

Lu, T. T., Makishima, M., Repa, J. J., Schoonjans, K., Kerr, T. A., Auwerx, J., & Mangelsdorf, D. J. (2000). Molecular Basis for Feedback Regulation of Bile Acid Synthesis by Nuclear Receptors. *Molecular Cell*, *6*(3), 507–515. https://doi.org/10.1016/S1097-2765(00)00050-2

Lucena, J. F., Herrero, J. I., Quiroga, J., Sangro, B., Garcia-Foncillas, J., Zabalegui, N., Sola, J., Herraiz, M., Medina, J. F., & Prieto, J. (2003). A multidrug resistance 3 gene mutation causing cholelithiasis, cholestasis of pregnancy, and adulthood biliary cirrhosis. *Gastroenterology*, *124*(4), 1037–1042. https://doi.org/10.1053/gast.2003.50144

Ma, K., Saha, P. K., Chan, L., & Moore, D. D. (2006). Farnesoid X receptor is essential for normal glucose

homeostasis. *The Journal of Clinical Investigation*, *116*(4), 1102–1109. https://doi.org/10.1172/JCI25604

Maddirevula, S., Alhebbi, H., Alqahtani, A., Algoufi, T., Alsaif, H. S., Ibrahim, N., Abdulwahab, F., Barr, M., Alzaidan, H., Almehaideb, A., AlSasi, O., Alhashem, A., Hussaini, H. A.-, Wali, S., & Alkuraya, F. S. (2019). Identification of novel loci for pediatric cholestatic liver disease defined by KIF12, PPM1F, USP53, LSR, and WDR83OS pathogenic variants. *Genetics in Medicine*, *21*(5), 1164–1172. https://doi.org/10.1038/s41436-018-0288-x

Makishima, M., Okamoto, A. Y., Repa, J. J., Tu, H., Learned, R. M., Luk, A., Hull, M. V., Lustig, K. D., Mangelsdorf, D. J., & Shan, B. (1999). Identification of a Nuclear Receptor for Bile Acids. *Science*, *284*(5418), 1362–1365. https://doi.org/10.1126/science.284.5418.1362

Malakhov, M. P., Mattern, M. R., Malakhova, O. A., Drinker, M., Weeks, S. D., & Butt, T. R. (2004). SUMO fusions and SUMO-specific protease for efficient expression and purification of proteins. *Journal of Structural and Functional Genomics*, *5*(1/2), 75–86. https://doi.org/10.1023/B:JSFG.0000029237.70316.52

Maloney, P. R., Parks, D. J., Haffner, C. D., Fivush, A. M., Chandra, G., Plunket, K. D., Creech, K. L., Moore, L. B., Wilson, J. G., Lewis, M. C., Jones, S. A., & Willson, T. M. (2000). Identification of a Chemical Tool for the Orphan Nuclear Receptor FXR. *Journal of Medicinal Chemistry*, *43*(16), 2971–2974. https://doi.org/10.1021/jm0002127

Markham, A., & Keam, S. J. (2016). Obeticholic Acid: First Global Approval. *Drugs*, *76*(12), 1221–1226. https://doi.org/10.1007/s40265-016-0616-x

Marshall, J. C. (1998). The gut as a potential trigger of exercise-induced inflammatory responses. *Canadian Journal of Physiology and Pharmacology*, *76*(5), 479–484. https://doi.org/10.1139/cjpp-76-5-479

Massafra, V., Pellicciari, R., Gioiello, A., & van Mil, S. W. C. (2018). Progress and challenges of selective Farnesoid X Receptor modulation. *Pharmacology and Therapeutics*, *191*, 162–177. https://doi.org/10.1016/j.pharmthera.2018.06.009

Mehta, S., Kumar, K., Bhardwaj, R., Malhotra, S., Goyal, N., & Sibal, A. (2022). Progressive Familial Intrahepatic Cholestasis: A Study in Children From a Liver Transplant Center in India. *Journal of Clinical and Experimental Hepatology*, *12*(2), 454–460. https://doi.org/10.1016/j.jceh.2021.06.006

Merk, D., Sreeramulu, S., Kudlinzki, D., Saxena, K., Linhard, V., Gande, S. L., Hiller, F., Lamers, C., Nilsson, E., Aagaard, A., Wissler, L., Dekker, N., Bamberg, K., Schubert-Zsilavecz, M., & Schwalbe, H. (2019). Molecular tuning of farnesoid X receptor partial agonism. *Nature Communications*, *10*(1), 1–14. https://doi.org/10.1038/s41467-019-10853-2

Meyer zu Schwabedissen, H. E., Böttcher, K., Chaudhry, A., Kroemer, H. K., Schuetz, E. G., & Kim, R. B. (2010). Liver X receptor α and farnesoid X receptor are major transcriptional regulators of OATP1B1. *Hepatology*, *52*(5), 1797–1807. https://doi.org/10.1002/hep.23876

Mi, L.-Z., Devarakonda, S., Harp, J. M., Han, Q., Pellicciari, R., Willson, T. M., Khorasanizadeh, S., & Rastinejad, F. (2003). Structural basis for bile acid binding and activation of the nuclear receptor FXR. *Molecular Cell*, *11*(4), 1093–1100. https://doi.org/10.1016/s1097-2765(03)00112-6

Michelucci, U. (2018). Applied deep learning: A case-based approach to understanding deep neural networks. In *APRESS Media, LLC*. https://doi.org/10.1007/978-1-4842-3790-8

Miller, M. P., & Kumar, S. (2001). Understanding human disease mutations through the use of interspecific genetic variation. *Human Molecular Genetics*, *10*(21), 2319–2328. https://doi.org/10.1093/hmg/10.21.2319

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533. https://doi.org/10.1038/nature14236

Mooney, S. D., Krishnan, V. G., & Evani, U. S. (2010). Bioinformatic Tools for Identifying Disease Gene and SNP Candidates. In *Genetic Variation.* (pp. 307–319). https://doi.org/10.1007/978-1-60327-367-1_17

Morales, E. F., & Escalante, H. J. (2022). A brief introduction to supervised, unsupervised, and reinforcement learning. In *Biosignal Processing and Classification Using Computational Learning and Intelligence* (pp. 111–129). Elsevier. https://doi.org/10.1016/B978-0-12-820125-1.00017-8

Müller, T., Hess, M. W., Schiefermeier, N., Pfaller, K., Ebner, H. L., Heinz-Erian, P., Ponstingl, H., Partsch, J., Röllinghoff, B., Köhler, H., Berger, T., Lenhartz, H., Schlenck, B., Houwen, R. J., Taylor, C. J., Zoller, H., Lechner, S., Goulet, O., Utermann, G., … Janecke, A. R. (2008). MYO5B mutations cause microvillus inclusion disease and disrupt epithelial cell polarity. *Nature Genetics*, *40*(10), 1163–1165. https://doi.org/10.1038/ng.225

Mustonen, E.-K., Lee, S. M. L., Nieß, H., Schwab, M., Pantsar, T., & Burk, O. (2021). Identification and characterization of novel splice variants of human farnesoid X receptor. *Archives of Biochemistry and Biophysics*, *705*, 108893. https://doi.org/10.1016/j.abb.2021.108893

Nagy, L., Kao, H.-Y., Love, J. D., Li, C., Banayo, E., Gooch, J. T., Krishna, V., Chatterjee, K., Evans, R. M., & Schwabe,

J. W. R. (1999). Mechanism of corepressor binding and release from nuclear hormone receptors. *Genes & Development*, *13*(24), 3209–3216. https://doi.org/10.1101/gad.13.24.3209

Nakahara, M., Furuya, N., Takagaki, K., Sugaya, T., Hirota, K., Fukamizu, A., Kanda, T., Fujii, H., & Sato, R. (2005). Ileal Bile Acid-binding Protein, Functionally Associated with the Farnesoid X Receptor or the Ileal Bile Acid Transporter, Regulates Bile Acid Activity in the Small Intestine. *Journal of Biological Chemistry*, *280*(51), 42283–42289. https://doi.org/10.1074/jbc.M507454200

Neimark, E., Chen, F., Li, X., & Shneider, B. L. (2004). Bile acid-induced negative feedback regulation of the human ileal bile acid transporter. *Hepatology*, *40*(1), 149–156. https://doi.org/10.1002/hep.20295

Neuschwander-Tetri, B. A., Loomba, R., Sanyal, A. J., Lavine, J. E., Van Natta, M. L., Abdelmalek, M. F., Chalasani, N., Dasarathy, S., Diehl, A. M., Hameed, B., Kowdley, K. V, McCullough, A., Terrault, N., Clark, J. M., Tonascia, J., Brunt, E. M., Kleiner, D. E., & Doo, E. (2015). Farnesoid X nuclear receptor ligand obeticholic acid for non-cirrhotic, non-alcoholic steatohepatitis (FLINT): a multicentre, randomised, placebo-controlled trial. *The Lancet*, *385*(9972), 956–965. https://doi.org/10.1016/S0140-6736(14)61933-4

Niroula, A., Urolagin, S., & Vihinen, M. (2015). PON-P2: Prediction method for fast and reliable identification of harmful variants. *PLoS ONE*, *10*(2). https://doi.org/10.1371/journal.pone.0117380

Niroula, A., & Vihinen, M. (2015). Classification of Amino Acid Substitutions in Mismatch Repair Proteins Using PON-MMR2. *Human Mutation*, *36*(12), 1128–1134. https://doi.org/10.1002/humu.22900

Nishimura, T., Utsunomiya, Y., Hoshikawa, M., Ohuchi, H., & Itoh, N. (1999). Structure and expression of a novel human FGF, FGF-19, expressed in the fetal brain. *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression*, *1444*(1), 148–151. https://doi.org/10.1016/S0167-4781(98)00255-3

Nitta, M., Ku, S., Brown, C., Okamoto, A. Y., & Shan, B. (1999). CPF: An orphan nuclear receptor that regulates liver-specific expression of the human cholesterol 7α-hydroxylase gene. *Proceedings of the National Academy of Sciences*, *96*(12), 6660–6665. https://doi.org/10.1073/pnas.96.12.6660

Niu, B., Scott, A. D., Sengupta, S., Bailey, M. H., Batra, P., Ning, J., Wyczalkowski, M. A., Liang, W.-W., Zhang, Q., McLellan, M. D., Sun, S. Q., Tripathi, P., Lou, C., Ye, K., Mashl, R. J., Wallis, J., Wendl, M. C., Chen, F., & Ding, L. (2016). Protein-structure-guided discovery of functional mutations across 19 cancer types. *Nature Genetics*, *48*(8), 827–837. https://doi.org/10.1038/ng.3586

Nosol, K., Bang-Sørensen, R., Irobalieva, R. N., Erramilli, S. K., Stieger, B., Kossiakoff, A. A., & Locher, K. P. (2021). Structures of ABCB4 provide insight into phosphatidylcholine translocation. *Proceedings of the National Academy of Sciences*, *118*(33). https://doi.org/10.1073/pnas.2106702118

Oh, S., Jo, Y., Jung, S., Yoon, S., & Yoo, K. H. (2020). From genome sequencing to the discovery of potential biomarkers in liver disease. *BMB Reports*, *53*(6), 299–310. https://doi.org/10.5483/BMBRep.2020.53.6.074

Olsen, J. A., Alam, A., Kowal, J., Stieger, B., & Locher, K. P. (2020). Structure of the human lipid exporter ABCB4 in a lipid environment. *Nature Structural and Molecular Biology*, *27*(1), 62–70. https://doi.org/10.1038/s41594-019-0354-3

Orellana, L. (2019). Large-Scale Conformational Changes and Protein Function: Breaking the in silico Barrier. *Frontiers in Molecular Biosciences*, *6*. https://doi.org/10.3389/fmolb.2019.00117

Otte, K., Kranz, H., Kober, I., Thompson, P., Hoefer, M., Haubold, B., Remmel, B., Voss, H., Kaiser, C., Albers, M., Cheruvallath, Z., Jackson, D., Casari, G., Koegl, M., Pääbo, S., Mous, J., Kremoser, C., & Deuschle, U. (2003). Identification of Farnesoid X Receptor β as a Novel Mammalian Nuclear Receptor Sensing Lanosterol. *Molecular and Cellular Biology*, *23*(3), 864–872. https://doi.org/10.1128/MCB.23.3.864-872.2003

Oude Elferink, R. P. J., & Paulusma, C. C. (2007). Function and pathophysiological importance of ABCB4 (MDR3 P-glycoprotein). In *Pflugers Archiv European Journal of Physiology* (Vol. 453, Issue 5, pp. 601–610). https://doi.org/10.1007/s00424-006-0062-9

Park, H. J., Kim, T. H., Kim, S. W., Noh, S. H., Cho, K. J., Choi, C., Kwon, E. Y., Choi, Y. J., Gee, H. Y., & Choi, J. H. (2016). Functional characterization of ABCB4 mutations found in progressive familial intrahepatic cholestasis type 3. *Scientific Reports*, *6*(May), 1–9. https://doi.org/10.1038/srep26872

Parks, D. J., Blanchard, S. G., Bledsoe, R. K., Chandra, G., Consler, T. G., Kliewer, S. A., Stimmel, J. B., Willson, T. M., Zavacki, A. M., Moore, D. D., & Lehmann, J. M. (1999). Bile Acids: Natural Ligands for an Orphan Nuclear Receptor. *Science*, *284*(5418), 1365–1368. https://doi.org/10.1126/science.284.5418.1365

Pauli-Magnus, C., Lang, T., Meier, Y., Zodan-Marin, T., Jung, D., Breymann, C., Zimmermann, R., Kenngott, S., Beuers, U., Reichel, C., Kerb, R., Penger, A., Meier, P. J., & Kullak-Ublick, G. A. (2004). Sequence analysis of bile salt export pump (ABCB11) and multidrug resistance p-glycoprotein 3 (ABCB4, MDR3) in patients with intrahepatic cholestasis of pregnancy. *Lippincott Williams & Wilkins Pharmacogenetics*, *14*, 91–102. https://doi.org/10.1097/01.fpc.0000054155.92680.24

Paulusma, C. C., Groen, A., Kunne, C., Ho-Mok, K. S., Spijkerboer, A. L., Rudi de Waart, D., Hoek, F. J., Vreeling, H., Hoeben, K. A., van Marle, J., Pawlikowska, L., Bull, L. N., Hofmann, A. F., Knisely, A. S., & Oude Elferink, R. P. J. (2006). Atp8b1 deficiency in mice reduces resistance of the canalicular membrane to hydrophobic bile

salts and impairs bile salt transport. *Hepatology*, *44*(1), 195–204. https://doi.org/10.1002/hep.21212

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: machine learning in python. *The Journal of Machine Learning Research*, *12*, 2825–2830. https://doi.org/10.1289/EHP4713

Pejaver, V., Mooney, S. D., & Radivojac, P. (2017). Missense variant pathogenicity predictors generalize well across a range of function-specific prediction challenges. *Human Mutation*, *38*(9), 1092–1108. https://doi.org/10.1002/humu.23258

Pejaver, V., Urresti, J., Lugo-Martinez, J., Pagel, K. A., Lin, G. N., Nam, H. J., Mort, M., Cooper, D. N., Sebat, J., Iakoucheva, L. M., Mooney, S. D., & Radivojac, P. (2020). Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-19669-x

Pellicciari, R., Fiorucci, S., Camaioni, E., Clerici, C., Costantino, G., Maloney, P. R., Morelli, A., Parks, D. J., & Willson, T. M. (2002). 6α-Ethyl-Chenodeoxycholic Acid (6-ECDCA), a Potent and Selective FXR Agonist Endowed with Anticholestatic Activity. *Journal of Medicinal Chemistry*, *45*(17), 3569–3572. https://doi.org/10.1021/jm025529g

Perino, A., Demagny, H., Velazquez-Villegas, L., & Schoonjans, K. (2021). Molecular physiology of bile acid signaling in health, disease, and aging. *Physiological Reviews*, *101*(2), 683–731. https://doi.org/10.1152/physrev.00049.2019

Pfister, E., Dröge, C., Liebe, R., Stalke, A., Buhl, N., Ballauff, A., Cantz, T., Bueltmann, E., Stindt, J., Luedde, T., Baumann, U., & Keitel, V. (2022). Extrahepatic manifestations of progressive familial intrahepatic cholestasis syndromes: Presentation of a case series and literature review. *Liver International*, *42*(5), 1084–1096. https://doi.org/10.1111/liv.15200

Pietrucci, F. (2017). Strategies for the exploration of free energy landscapes: Unity in diversity and challenges ahead. *Reviews in Physics*, *2*, 32–45. https://doi.org/10.1016/j.revip.2017.05.001

Plass, J. R. M., Mol, O., Heegsma, J., Geuken, M., Faber, K. N., Jansen, P. L. M., & Müller, M. (2002). Farnesoid X receptor and bile salts are involved in transcriptional regulation of the gene encoding the human bile salt export pump. *Hepatology*, *35*(3), 589–596. https://doi.org/10.1053/jhep.2002.31724

Pohjolainen, E., Chen, X., Malola, S., Groenhof, G., & Häkkinen, H. (2016). A Unified AMBER-Compatible Molecular Mechanics Force Field for Thiolate-Protected Gold Nanoclusters. *Journal of Chemical Theory and Computation*, *12*(3), 1342–1350. https://doi.org/10.1021/acs.jctc.5b01053

Poupon, R., Barbu, V., Chamouard, P., Wendum, D., Rosmorduc, O., & Housset, C. (2010). Combined features of low phospholipid-associated cholelithiasis and progressive familial intrahepatic cholestasis 3. *Liver International*, *30*(2), 327–331. https://doi.org/10.1111/j.1478-3231.2009.02148.x

Poupon, R., Rosmorduc, O., Boëlle, P. Y., Chrétien, Y., Corpechot, C., Chazouillères, O., Housset, C., & Barbu, V. (2013). Genotype-phenotype relationships in the low-phospholipid-associated cholelithiasis syndrome: A study of 156 consecutive patients. *Hepatology*, *58*(3), 1105–1110. https://doi.org/10.1002/hep.26424

Prescher, M., Bonus, M., Stindt, J., Keitel-Anselmino, V., Smits, S. H. J., Gohlke, H., & Schmitt, L. (2021). Evidence for a credit-card-swipe mechanism in the human PC floppase ABCB4. *Structure*, *29*(10), 1144-1155.e5. https://doi.org/10.1016/j.str.2021.05.013

Prescher, M., Kroll, T., & Schmitt, L. (2019). ABCB4 / MDR3 in health and disease - At the crossroads of biochemistry and medicine. *Biological Chemistry*, *400*(10), 1245–1259. https://doi.org/10.1515/hsz-2018-0441

Qiu, Y. L., Gong, J. Y., Feng, J. Y., Wang, R. X., Han, J., Liu, T., Lu, Y., Li, L. T., Zhang, M. H., Sheps, J. A., Wang, N. L., Yan, Y. Y., Li, J. Q., Chen, L., Borchers, C. H., Sipos, B., Knisely, A. S., Ling, V., Xing, Q. H., & Wang, J. S. (2017). Defects in myosin VB are associated with a spectrum of previously undiagnosed low γ-glutamyltransferase cholestasis. *Hepatology*, *65*(5), 1655–1669. https://doi.org/10.1002/hep.29020

Quinlan, J. R. (1993). *Programs for Machine Learning.* Morgan Kaufmann Publishers Inc.

Quinn, R. A., Melnik, A. V., Vrbanac, A., Fu, T., Patras, K. A., Christy, M. P., Bodai, Z., Belda-Ferre, P., Tripathi, A., Chung, L. K., Downes, M., Welch, R. D., Quinn, M., Humphrey, G., Panitchpakdi, M., Weldon, K. C., Aksenov, A., da Silva, R., Avila-Pacheco, J., … Dorrestein, P. C. (2020). Global chemical effects of the microbiome include new bile-acid conjugations. *Nature*, *579*(7797), 123–129. https://doi.org/10.1038/s41586-020-2047-9

Radun, R., & Trauner, M. (2021). Role of FXR in Bile Acid and Metabolic Homeostasis in NASH: Pathogenetic Concepts and Therapeutic Opportunities. *Seminars in Liver Disease*, *41*(04), 461–475. https://doi.org/10.1055/s-0041-1731707

Rai, V., Gaur, M., Shukla, S., Shukla, S., Ambudkar, S. V., Komath, S. S., & Prasad, R. (2006). Conserved Asp327 of Walker B Motif in the N-Terminal Nucleotide Binding Domain (NBD-1) of Cdr1p of Candida albicans Has Acquired a New Role in ATP Hydrolysis. *Biochemistry*, *45*(49), 14726–14739.

https://doi.org/10.1021/bi061535t

Ramos Pittol, J. M., Milona, A., Morris, I., Willemsen, E. C. L., van der Veen, S. W., Kalkhoven, E., & van Mil, S. W. C. (2020). FXR Isoforms Control Different Metabolic Functions in Liver Cells via Binding to Specific DNA Motifs. *Gastroenterology*, *159*(5), 1853-1865.e10. https://doi.org/10.1053/j.gastro.2020.07.036

Ramsey, D. C., Scherrer, M. P., Zhou, T., & Wilke, C. O. (2011). The Relationship Between Relative Solvent Accessibility and Evolutionary Rate in Protein Evolution. *Genetics*, *188*(2), 479–488. https://doi.org/10.1534/genetics.111.128025

Rastinejad, F., Wagner, T., Zhao, Q., & Khorasanizadeh, S. (2000). Structure of the RXR–RAR DNA-binding complex on the retinoic acid response element DR1. *The EMBO Journal*, *19*(5), 1045–1054. https://doi.org/10.1093/emboj/19.5.1045

Reel, P. S., Reel, S., Pearson, E., Trucco, E., & Jefferson, E. (2021). Using machine learning approaches for multi-omics data analysis: A review. *Biotechnology Advances*, *49*, 107739. https://doi.org/10.1016/j.biotechadv.2021.107739

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-Validation. In *Encyclopedia of Database Systems* (pp. 532–538). Springer US. https://doi.org/10.1007/978-0-387-39940-9_565

Renaud, J. P., Rochel, N., Ruff, M., Vivat, V., Chambon, P., Gronemeyer, H., & Moras, D. (1995). Crystal structure of the RAR-gamma ligand-binding domain bound to all-trans retinoic acid. *Nature*, *378*(6558), 681–689. https://doi.org/10.1038/378681a0

Renga, B., Mencarelli, A., Vavassori, P., Brancaleone, V., & Fiorucci, S. (2010). The bile acid sensor FXR regulates insulin transcription and secretion. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, *1802*(3), 363–372. https://doi.org/10.1016/j.bbadis.2010.01.002

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K., & Rehm, H. L. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine*, *17*(5), 405–424. https://doi.org/10.1038/gim.2015.30

Riera, C., Padilla, N., & de la Cruz, X. (2016). The Complementarity Between Protein-Specific and General Pathogenicity Predictors for Amino Acid Substitutions. *Human Mutation*, *37*(10), 1013–1024. https://doi.org/10.1002/humu.23048

Rivera-Lopez, R., Canul-Reich, J., Mezura-Montes, E., & Cruz-Chávez, M. A. (2022). Induction of decision trees as classification models through metaheuristics. *Swarm and Evolutionary Computation*, *69*, 101006. https://doi.org/10.1016/j.swevo.2021.101006

Rochel, N., Ciesielski, F., Godet, J., Moman, E., Roessle, M., Peluso-Iltis, C., Moulin, M., Haertlein, M., Callow, P., Mély, Y., Svergun, D. I., & Moras, D. (2011). Common architecture of nuclear receptor heterodimers on DNA direct repeat elements with different spacings. *Nature Structural & Molecular Biology*, *18*(5), 564–570. https://doi.org/10.1038/nsmb.2054

Rodriguez, J. D., Perez, A., & Lozano, J. A. (2010). Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(3), 569–575. https://doi.org/10.1109/TPAMI.2009.187

Rodríguez, J. D., Pérez, A., & Lozano, J. A. (2013). A general framework for the statistical analysis of the sources of variance for classification error estimators. *Pattern Recognition*, *46*(3), 855–864. https://doi.org/10.1016/j.patcog.2012.09.007

Rokach, L., & Maimon, O. (2005). Top-Down Induction of Decision Trees Classifiers—A Survey. *IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews)*, *35*(4), 476–487. https://doi.org/10.1109/TSMCC.2004.843247

Ropponen, A., Sund, R., Riikonen, S., Ylikorkala, O., & Aittomäki, K. (2006). Intrahepatic cholestasis of pregnancy as an indicator of liver and biliary diseases: A population-based study. *Hepatology*, *43*(4), 723–728. https://doi.org/10.1002/hep.21111

Rosen, D. R., Siddique, T., Patterson, D., Figlewicz, D. A., Sapp, P., Hentati, A., Donaldson, D., Goto, J., O'Regan, J. P., Deng, H.-X., Rahmani, Z., Krizus, A., McKenna-Yasek, D., Cayabyab, A., Gaston, S. M., Berger, R., Tanzi, R. E., Halperin, J. J., Herzfeldt, B., … Brown, R. H. (1993). Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. *Nature*, *362*(6415), 59–62. https://doi.org/10.1038/362059a0

Rosmorduc, O., Hermelin, B., Boelle, P. Y., Parc, R., Taboury, J., & Poupon, R. (2003). ABCB4 gene mutation-associated cholelithiasis in adults. *Gastroenterology*, *125*(2), 452–459. https://doi.org/10.1016/S0016-5085(03)00898-9

Rosmorduc, O., Hermelin, B., & Poupon, R. (2001). *MDR3 Gene Defect in Adults With Symptomatic Intrahepatic and Gallbladder Cholesterol Cholelithiasis*. 1459–1467. https://doi.org/10.1053/gast.2001.23947

Russell, D. W. (2003). The Enzymes, Regulation, and Genetics of Bile Acid Synthesis. *Annual Review of Biochemistry*, *72*(1), 137–174. https://doi.org/10.1146/annurev.biochem.72.121801.161712

Saen-Oon, S., Lozoya, E., Segarra, V., Guallar, V., & Soliva, R. (2019). Atomistic simulations shed new light on the activation mechanisms of RORγ and classify it as Type III nuclear hormone receptor regarding ligand-binding paths. *Scientific Reports*, *9*(1), 17249. https://doi.org/10.1038/s41598-019-52319-x

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *WIREs Data Mining and Knowledge Discovery*, *8*(4). https://doi.org/10.1002/widm.1249

Saleem, K., Cui, Q., Zaib, T., Zhu, S., Qin, Q., Wang, Y., Dam, J., Ji, W., Liu, P., Jia, X., Wu, J., Bai, J., Fu, S., & Sun, W. (2020). Evaluation of a Novel Missense Mutation in ABCB4 Gene Causing Progressive Familial Intrahepatic Cholestasis Type 3. *Disease Markers*, *2020*. https://doi.org/10.1155/2020/6292818

Salem, N., & Hussein, S. (2019). Data dimensional reduction and principal components analysis. *Procedia Computer Science*, *163*, 292–299. https://doi.org/10.1016/j.procs.2019.12.111

Salomon-Ferrer, R., Case, D. A., & Walker, R. C. (2013). An overview of the Amber biomolecular simulation package. *WIREs Computational Molecular Science*, *3*(2), 198–210. https://doi.org/10.1002/wcms.1121

Sambrotta, M., Strautnieks, S., Papouli, E., Rushton, P., Clark, B. E., Parry, D. A., Logan, C. V, Newbury, L. J., Kamath, B. M., Ling, S., Grammatikopoulos, T., Wagner, B. E., Magee, J. C., Sokol, R. J., Mieli-Vergani, G., Smith, J. D., Johnson, C. A., McClean, P., Simpson, M. A., … Thompson, R. J. (2014). Mutations in TJP2 cause progressive cholestatic liver disease. *Nature Genetics*, *46*(4), 326–328. https://doi.org/10.1038/ng.2918

Sambrotta, M., & Thompson, R. J. (2015). Mutations in TJP2, encoding zona occludens 2, and liver disease. *Tissue Barriers*, *3*(3), 1–5. https://doi.org/10.1080/21688370.2015.1026537

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(3), 160. https://doi.org/10.1007/s42979-021-00592-x

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, *5*(2), 197–227. https://doi.org/10.1007/BF00116037

Schepers, B., & Gohlke, H. (2020). AMBER-DYES in AMBER: Implementation of fluorophore and linker parameters into AmberTools. *The Journal of Chemical Physics*, *152*(22). https://doi.org/10.1063/5.0007630

Schmitt, L., & Tampé, R. (2002). Structure and mechanism of ABC transporters. *Current Opinion in Structural Biology*, *12*(6), 754–760. https://doi.org/10.1016/s0959-440x(02)00399-8

Schote, A. B., Turner, J. D., Schiltz, J., & Muller, C. P. (2007). Nuclear receptors in human immune cells: Expression and correlations. *Molecular Immunology*, *44*(6), 1436–1445. https://doi.org/10.1016/j.molimm.2006.04.021

Shalon, D., Culver, R. N., Grembi, J. A., Folz, J., Treit, P. V., Shi, H., Rosenberger, F. A., Dethlefsen, L., Meng, X., Yaffe, E., Aranda-Díaz, A., Geyer, P. E., Mueller-Reif, J. B., Spencer, S., Patterson, A. D., Triadafilopoulos, G., Holmes, S. P., Mann, M., Fiehn, O., … Huang, K. C. (2023). Profiling the human intestinal environment under physiological conditions. In *Nature* (Vol. 617, Issue 7961). Springer US. https://doi.org/10.1038/s41586-023-05989-7

Sharada, K., Alghamdi, W., Karthika, K., Alawadi, A. H., Nozima, G., & Vijayan, V. (2023). Deep Learning Techniques for Image Recognition and Object Detection. *E3S Web of Conferences*, *399*, 1–11. https://doi.org/10.1051/e3sconf/202339904032

Shen, H., Zhang, Y., Ding, H., Wang, X., Chen, L., Jiang, H., & Shen, X. (2008). Farnesoid X Receptor Induces GLUT4 Expression Through FXR Response Element in the GLUT4 Promoter. *Cellular Physiology and Biochemistry*, *22*(1–4), 001–014. https://doi.org/10.1159/000149779

Shwartz-ziv, R., & Armon, A. (2021). Tabular Data: Deep Learning Is Not All You Need. *ICML 2021 Workshop AutoML Blind Submission*, 1–11.

Simons, S. S., Edwards, D. P., & Kumar, R. (2014). Minireview: Dynamic Structures of Nuclear Hormone Receptors: New Promises and Challenges. *Molecular Endocrinology*, *28*(2), 173–182. https://doi.org/10.1210/me.2013-1334

Sinal, C. J., Tohkin, M., Miyata, M., Ward, J. M., Lambert, G., & Gonzalez, F. J. (2000). Targeted Disruption of the Nuclear Receptor FXR/BAR Impairs Bile Acid and Lipid Homeostasis. *Cell*, *102*(6), 731–744. https://doi.org/10.1016/S0092-8674(00)00062-3

Singh, N., Yadav, M., Singh, A. K., Kumar, H., Dwivedi, S. K. D., Mishra, J. S., Gurjar, A., Manhas, A., Chandra, S., Yadav, P. N., Jagavelu, K., Siddiqi, M. I., Trivedi, A. K., Chattopadhyay, N., & Sanyal, S. (2014). Synthetic FXR Agonist GW4064 Is a Modulator of Multiple G Protein–Coupled Receptors. *Molecular Endocrinology*, *28*(5), 659–673. https://doi.org/10.1210/me.2013-1353

Singh, T., Poterba, T., Curtis, D., Akil, H., Al Eissa, M., Barchas, J. D., Bass, N., Bigdeli, T. B., Breen, G., Bromet, E. J., Buckley, P. F., Bunney, W. E., Bybjerg-Grauholm, J., Byerley, W. F., Chapman, S. B., Chen, W. J., Churchhouse, C., Craddock, N., Cusick, C. M., … Daly, M. J. (2022). Rare coding variants in ten genes confer substantial risk for schizophrenia. *Nature*, *604*(7906), 509–516. https://doi.org/10.1038/s41586-022-

04556-w

Skjaerven, L., Grant, B., Muga, A., Teigen, K., McCammon, J. A., Reuter, N., & Martinez, A. (2011). Conformational Sampling and Nucleotide-Dependent Transitions of the GroEL Subunit Probed by Unbiased Molecular Dynamics Simulations. *PLoS Computational Biology*, *7*(3), e1002004. https://doi.org/10.1371/journal.pcbi.1002004

Smith, A. J., Timmermans-Hereijgers, J. L. P. M., Roelofsen, B., Wirtz, K. W. A., van Blitterswijk, W. J., Smit, J. J. M., Schinkel, A. H., & Borst, P. (1994). The human MDR3 P-glycoprotein promotes translocation of phosphatidylcholine through the plasma membrane of fibroblasts from transgenic mice. *FEBS Letters*, *354*(3), 263–266. https://doi.org/10.1016/0014-5793(94)01135-4

Smith, A. J., van Helvoort, A., van Meer, G., Szabó, K., Welker, E., Szakács, G., Váradi, A., Sarkadi, B., & Borst, P. (2000). MDR3 P-glycoprotein, a Phosphatidylcholine Translocase, Transports Several Cytotoxic Drugs and Directly Interacts with Drugs as Judged by Interference with Nucleotide Trapping. *Journal of Biological Chemistry*, *275*(31), 23530–23539. https://doi.org/10.1074/jbc.M909002199

Smith, L. M., Sanders, J. Z., Kaiser, R. J., Hughes, P., Dodd, C., Connell, C. R., Heiner, C., Kent, S. B. H., & Hood, L. E. (1986). Fluorescence detection in automated DNA sequence analysis. *Nature*, *321*(6071), 674–679. https://doi.org/10.1038/321674a0

Sohail, M. I., Dönmez-Cakil, Y., Szöllősi, D., Stockner, T., & Chiba, P. (2021). The Bile Salt Export Pump: Molecular Structure, Study Models and Small-Molecule Drugs for the Treatment of Inherited BSEP Deficiencies. *International Journal of Molecular Sciences*, *22*(2), 784. https://doi.org/10.3390/ijms22020784

Song, K.-H., Li, T., Owsley, E., Strom, S., & Chiang, J. Y. L. (2009). Bile acids activate fibroblast growth factor 19 signaling in human hepatocytes to inhibit cholesterol 7α-hydroxylase gene expression. *Hepatology*, *49*(1), 297–305. https://doi.org/10.1002/hep.22627

Srivastava, A. (2014). Progressive familial intrahepatic cholestasis. *Journal of Clinical and Experimental Hepatology*, *4*(1), 25–36. https://doi.org/10.1016/j.jceh.2013.10.005

Stalke, A., Behrendt, A., Hennig, F., Gohlke, H., Buhl, N., Reinkens, T., Baumann, U., Schlegelberger, B., Illig, T., Pfister, E. D., & Skawran, B. (2023). Functional characterization of novel or yet uncharacterized ATP7B missense variants detected in patients with clinical Wilson's disease. *Clinical Genetics*, *104*(2), 174–185. https://doi.org/10.1111/cge.14352

Stalke, A., Sgodda, M., Cantz, T., Skawran, B., Lainka, E., Hartleben, B., Baumann, U., & Pfister, E.-D. (2022). KIF12 Variants and Disturbed Hepatocyte Polarity in Children with a Phenotypic Spectrum of Cholestatic Liver Disease. *The Journal of Pediatrics*, *240*, 284-291.e9. https://doi.org/10.1016/j.jpeds.2021.09.019

Starita, L. M., Ahituv, N., Dunham, M. J., Kitzman, J. O., Roth, F. P., Seelig, G., Shendure, J., & Fowler, D. M. (2017). Variant Interpretation: Functional Assays to the Rescue. *The American Journal of Human Genetics*, *101*(3), 315–325. https://doi.org/10.1016/j.ajhg.2017.07.014

Stättermayer, A. F., Halilbasic, E., Wrba, F., Ferenci, P., & Trauner, M. (2020). Variants in ABCB4 (MDR3) across the spectrum of cholestatic liver diseases in adults. *Journal of Hepatology*, *73*(3), 651–663. https://doi.org/10.1016/j.jhep.2020.04.036

Stenson, P. D., Ball, E. V., Mort, M., Phillips, A. D., Shiel, J. A., Thomas, N. S. T., Abeysinghe, S., Krawczak, M., & Cooper, D. N. (2003). Human Gene Mutation Database (HGMD ®): 2003 update. *Human Mutation*, *21*(6), 577–581. https://doi.org/10.1002/humu.10212

Sticova, E., & Jirsa, M. (2020). ABCB4 disease: Many faces of one gene deficiency. *Annals of Hepatology*, *19*(2), 126–133. https://doi.org/10.1016/j.aohep.2019.09.010

Stoppelman, J. P., Ng, T. T., Nerenberg, P. S., & Wang, L.-P. (2021). Development and Validation of AMBER-FB15-Compatible Force Field Parameters for Phosphorylated Amino Acids. *The Journal of Physical Chemistry B*, *125*(43), 11927–11942. https://doi.org/10.1021/acs.jpcb.1c07547

Stormo, G. D., Schneider, T. D., Gold, L., & Ehrenfeucht, A. (1982). Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. *Nucleic Acids Research*, *10*(9), 2997–3011. https://doi.org/10.1093/nar/10.9.2997

Strautnieks, S. S., Bull, L. N., Knisely, A. S., Kocoshis, S. A., Dahl, N., Arnell, H., Sokal, E., Dahan, K., Childs, S., Ling, V., Tanner, M. S., Kagalwalla, A. F., Németh, A., Pawlowska, J., Baker, A., Mieli-Vergani, G., Freimer, N. B., Gardiner, R. M., & Thompson, R. J. (1998). A gene encoding a liver-specific ABC transporter is mutated in progressive familial intrahepatic cholestasis. *Nature Genetics*, *20*(3), 233–238. https://doi.org/10.1038/3034

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction.* (MIT press). MIT press.

Svensson, S., Ostberg, T., Jacobsson, M., Norström, C., Stefansson, K., Hallén, D., Johansson, I. C., Zachrisson, K., Ogg, D., & Jendeberg, L. (2003). Crystal structure of the heterodimeric complex of LXRalpha and RXRbeta ligand-binding domains in a fully agonistic conformation. *The EMBO Journal*, *22*(18), 4625–4633. https://doi.org/10.1093/emboj/cdg456

Thusberg, J., & Vihinen, M. (2009). Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods. *Human Mutation*, *30*(5), 703–714. https://doi.org/10.1002/humu.20938

Tian, C., Kasavajhala, K., Belfon, K. A. A., Raguette, L., Huang, H., Migues, A. N., Bickel, J., Wang, Y., Pincay, J., Wu, Q., & Simmerling, C. (2020). ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. *Journal of Chemical Theory and Computation*, *16*(1), 528–552. https://doi.org/10.1021/acs.jctc.9b00591

Tien, M. Z., Meyer, A. G., Sydykova, D. K., Spielman, S. J., & Wilke, C. O. (2013). Maximum allowed solvent accessibilites of residues in proteins. *PLoS ONE*, *8*(11). https://doi.org/10.1371/journal.pone.0080635

Tilg, H., Adolph, T. E., & Trauner, M. (2022). Gut-liver axis: Pathophysiological concepts and clinical implications. *Cell Metabolism*, *34*(11), 1700–1718. https://doi.org/10.1016/j.cmet.2022.09.017

Tóth-Petróczy, Á., & Tawfik, D. S. (2014). The robustness and innovability of protein folds. *Current Opinion in Structural Biology*, *26*, 131–138. https://doi.org/10.1016/j.sbi.2014.06.007

Tougeron, D., Fotsing, G., Barbu, V., & Beauchant, M. (2012). ABCB4/MDR3 gene mutations and cholangiocarcinomas. In *Journal of Hepatology* (Vol. 57, Issue 2, pp. 467–468). https://doi.org/10.1016/j.jhep.2012.01.025

Trauner, M., & Boyer, J. L. (2003). Bile Salt Transporters: Molecular Characterization, Function, and Regulation. *Physiological Reviews*, *83*(2), 633–671. https://doi.org/10.1152/physrev.00027.2002

Trefts, E., Gannon, M., & Wasserman, D. H. (2017). The liver. *Current Biology : CB*, *27*(21), R1147–R1151. https://doi.org/10.1016/j.cub.2017.09.019

Trubetskoy, V., Pardiñas, A. F., Qi, T., Panagiotaropoulou, G., Awasthi, S., Bigdeli, T. B., Bryois, J., Chen, C.-Y., Dennison, C. A., Hall, L. S., Lam, M., Watanabe, K., Frei, O., Ge, T., Harwood, J. C., Koopmans, F., Magnusson, S., Richards, A. L., Sidorenko, J., … van Os, J. (2022). Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature*, *604*(7906), 502–508. https://doi.org/10.1038/s41586-022-04434-5

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., Olsson, I., Edlund, K., Lundberg, E., Navani, S., Szigyarto, C. A.-K., Odeberg, J., Djureinovic, D., Takanen, J. O., Hober, S., … Pontén, F. (2015). Tissue-based map of the human proteome. *Science*, *347*(6220). https://doi.org/10.1126/science.1260419

Unwin, P. N. T., & Henderson, R. (1975). Molecular structure determination by electron microscopy of unstained crystalline specimens. *Journal of Molecular Biology*, *94*(3), 425–440. https://doi.org/10.1016/0022-2836(75)90212-0

Urbatsch, I. L., Julien, M., Carrier, I., Rousseau, M.-E., Cayrol, R., & Gros, P. (2000). Mutational Analysis of Conserved Carboxylate Residues in the Nucleotide Binding Sites of P-Glycoprotein. *Biochemistry*, *39*(46), 14138–14149. https://doi.org/10.1021/bi001128w

Urizar, N. L., Liverman, A. B., Dodds, D. T., Silva, F. V., Ordentlich, P., Yan, Y., Gonzalez, F. J., Heyman, R. A., Mangelsdorf, D. J., & Moore, D. D. (2002). A Natural Product That Lowers Cholesterol As an Antagonist Ligand for FXR. *Science*, *296*(5573), 1703–1706. https://doi.org/10.1126/science.1072891

Van der Bliek, A. M., Baas, F., Ten Houte de Lange, T., Kooiman, P. M., Van der Velde-Koerts, T., & Borst, P. (1987). The human mdr3 gene encodes a novel P-glycoprotein homologue and gives rise to alternatively spliced mRNAs in liver. *The EMBO Journal*, *6*(11), 3325–3331. https://doi.org/10.1002/j.1460-2075.1987.tb02653.x

van der Bliek, A. M., Kooiman, P. M., Schneider, C., & Borst, P. (1988). Sequence of mdr3 cDNA encoding a human P-glycoprotein. *Gene*, *71*(2), 401–411. https://doi.org/10.1016/0378-1119(88)90057-1

van Helvoort, A., Smith, A. J., Sprong, H., Fritzsche, I., Schinkel, A. H., Borst, P., & van Meer, G. (1996). MDR1 P-Glycoprotein Is a Lipid Translocase of Broad Specificity, While MDR3 P-Glycoprotein Specifically Translocates Phosphatidylcholine. *Cell*, *87*(3), 507–517. https://doi.org/10.1016/S0092-8674(00)81370-7

van Mil, S. W. C., Milona, A., Dixon, P. H., Mullenbach, R., Geenes, V. L., Chambers, J., Shevchuk, V., Moore, G. E., Lammert, F., Glantz, A. G., Mattsson, L., Whittaker, J., Parker, M. G., White, R., & Williamson, C. (2007). Functional Variants of the Central Bile Acid Sensor FXR Identified in Intrahepatic Cholestasis of Pregnancy. *Gastroenterology*, *133*(2), 507–516. https://doi.org/10.1053/j.gastro.2007.05.015

Vaquero, J., Monte, M. J., Dominguez, M., Muntané, J., & Marin, J. J. G. (2013). Differential activation of the human farnesoid X receptor depends on the pattern of expressed isoforms and the bile acid pool composition. *Biochemical Pharmacology*, *86*(7), 926–939. https://doi.org/10.1016/j.bcp.2013.07.022

Vihinen, M. (2012). How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. *BMC Genomics*, *13*(Suppl 4), S2. https://doi.org/10.1186/1471-2164-13-S4-S2

Vlahcevic, Z. R., Heuman, D. M., & Hylemon, P. B. (1991). Regulation of bile acid synthesis. *Hepatology*, *13*(3), 590–600. https://doi.org/10.1002/hep.1840130331

Wagner, M., Zollner, G., & Trauner, M. (2008). Nuclear bile acid receptor farnesoid X receptor meets nuclear

factor-κB: New insights into hepatic inflammation. *Hepatology*, *48*(5), 1383–1386. https://doi.org/10.1002/hep.22668

Walker, J. E., Saraste, M., Runswick, M. J., & Gay, N. J. (1982). Distantly related sequences in the alpha- and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. *The EMBO Journal*, *1*(8), 945–951. https://doi.org/10.1002/j.1460-2075.1982.tb01276.x

Wang, H., Chen, J., Hollister, K., Sowers, L. C., & Forman, B. M. (1999). Endogenous Bile Acids Are Ligands for the Nuclear Receptor FXR/BAR. *Molecular Cell*, *3*(5), 543–553. https://doi.org/10.1016/S1097-2765(00)80348-2

Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A., & Case, D. A. (2004). Development and testing of a general amber force field. *Journal of Computational Chemistry*, *25*(9), 1157–1174. https://doi.org/10.1002/jcc.20035

Wang, X., Ren, H., Ren, J., Song, W., Qiao, Y., Ren, Z., Zhao, Y., Linghu, L., Cui, Y., Zhao, Z., Chen, L., & Qiu, L. (2023). Machine learning-enabled risk prediction of chronic obstructive pulmonary disease with unbalanced data. *Computer Methods and Programs in Biomedicine*, *230*, 107340. https://doi.org/10.1016/j.cmpb.2023.107340

Wang, Y.-D., Chen, W.-D., Moore, D. D., & Huang, W. (2008). FXR: a metabolic regulator and cell protector. *Cell Research*, *18*(11), 1087–1095. https://doi.org/10.1038/cr.2008.289

Wang, Y.-D., Chen, W.-D., Wang, M., Yu, D., Forman, B. M., & Huang, W. (2008). Farnesoid X receptor antagonizes nuclear factor κB in hepatic inflammatory response. *Hepatology*, *48*(5), 1632–1643. https://doi.org/10.1002/hep.22519

Wang, Y., Crittenden, D. B., Eng, C., Zhang, Q., Guo, P., Chung, D., Fenaux, M., Klucher, K., Jones, C., Jin, F., Quirk, E., & Charlton, M. R. (2021). Safety, Pharmacokinetics, Pharmacodynamics, and Formulation of Liver-Distributed Farnesoid X-Receptor Agonist TERN-101 in Healthy Volunteers. *Clinical Pharmacology in Drug Development*, *10*(10), 1198–1208. https://doi.org/10.1002/cpdd.960

Wei, Q., & Dunbrack, R. L. (2013). The Role of Balanced Training and Testing Data Sets for Binary Classifiers in Bioinformatics. *PLoS ONE*, *8*(7), e67863. https://doi.org/10.1371/journal.pone.0067863

Weikum, E. R., Liu, X., & Ortlund, E. A. (2018). The nuclear receptor superfamily: A structural perspective. *Protein Science*, *27*(11), 1876–1892. https://doi.org/10.1002/pro.3496

Weile, J., & Roth, F. P. (2018). Multiplexed assays of variant effects contribute to a growing genotype–phenotype atlas. *Human Genetics*, *137*(9), 665–678. https://doi.org/10.1007/s00439-018-1916-x

Wen, A. E. C., & Campbell, C. B. (1977). Bile Salt Metabolism:I. The Physiology of Bile Salts. *Australian and New Zealand Journal of Medicine*, *7*(6), 579–586. https://doi.org/10.1111/j.1445-5994.1977.tb02312.x

Wen, J., Lord, H., Knutson, N., & Wikström, M. (2020). Nano differential scanning fluorimetry for comparability studies of therapeutic proteins. *Analytical Biochemistry*, *593*, 113581. https://doi.org/10.1016/j.ab.2020.113581

Wendum, D., Barbu, V., Rosmorduc, O., Arrivé, L., Fléjou, J. F., & Poupon, R. (2012). Aspects of liver pathology in adult patients with MDR3/ABCB4 gene mutations. *Virchows Archiv*, *460*(3), 291–298. https://doi.org/10.1007/s00428-012-1202-6

Wu, Y., Liu, H., Li, R., Sun, S., Weile, J., & Roth, F. P. (2021). Improved pathogenicity prediction for rare human missense variants. *The American Journal of Human Genetics*, *108*(10), 1891–1906. https://doi.org/10.1016/j.ajhg.2021.08.012

Wurtz, J.-M., Bourguet, W., Renaud, J.-P., Vivat, V., Chambon, P., Moras, D., & Gronemeyer, H. (1996). A canonical structure for the ligand-binding domain of nuclear receptors. *Nature Structural Biology*, *3*(1), 87–94. https://doi.org/10.1038/nsb0196-87

Xie, M.-H., Holcomb, I., Deuel, B., Dowd, P., Huang, A., Vagts, A., Foster, J., Liang, J., Brush, J., Gu, Q., Hillan, K., Goddard, A., & Gurney, A. L. (1999). FGF-19, a novel fibroblast growth factor with unique specificity for FGFR4. *Cytokine*, *11*(10), 729–735. https://doi.org/10.1006/cyto.1999.0485

Xu, H. E., Stanley, T. B., Montana, V. G., Lambert, M. H., Shearer, B. G., Cobb, J. E., McKee, D. D., Galardi, C. M., Plunket, K. D., Nolte, R. T., Parks, D. J., Moore, J. T., Kliewer, S. A., Willson, T. M., & Stimmel, J. B. (2002). Structural basis for antagonist-mediated recruitment of nuclear co-repressors by PPARα. *Nature*, *415*(6873), 813–817. https://doi.org/10.1038/415813a

Yan, M., Man, S., Sun, B., Ma, L., Guo, L., Huang, L., & Gao, W. (2023). Gut liver brain axis in diseases: the implications for therapeutic interventions. *Signal Transduction and Targeted Therapy*, *8*(1). https://doi.org/10.1038/s41392-023-01673-4

Yan, X., Broderick, D., Leid, M. E., Schimerlik, M. I., & Deinzer, M. L. (2004). Dynamics and Ligand-Induced Solvent Accessibility Changes in Human Retinoid X Receptor Homodimer Determined by Hydrogen Deuterium Exchange and Mass Spectrometry. *Biochemistry*, *43*(4), 909–917. https://doi.org/10.1021/bi030183c

Yang, F., Huang, X., Yi, T., Yen, Y., Moore, D. D., & Huang, W. (2007). Spontaneous Development of Liver Tumors

in the Absence of the Bile Acid Receptor Farnesoid X Receptor. *Cancer Research*, *67*(3), 863–867. https://doi.org/10.1158/0008-5472.CAN-06-1078

Yang, Y. I., Shao, Q., Zhang, J., Yang, L., & Gao, Y. Q. (2019). Enhanced sampling in molecular dynamics. *The Journal of Chemical Physics*, *151*(7). https://doi.org/10.1063/1.5109531

Yip, S. S. F., Parmar, C., Kim, J., Huynh, E., Mak, R. H., & Aerts, H. J. W. L. (2017). Impact of experimental design on PET radiomics in predicting somatic mutation status. *European Journal of Radiology*, *97*, 8–15. https://doi.org/10.1016/j.ejrad.2017.10.009

Yoo, J., Winogradoff, D., & Aksimentiev, A. (2020). Molecular dynamics simulations of DNA–DNA and DNA–protein interactions. *Current Opinion in Structural Biology*, *64*, 88–96. https://doi.org/10.1016/j.sbi.2020.06.007

You, W., Chen, B., Liu, X., Xue, S., Qin, H., & Jiang, H. (2017). Farnesoid X receptor, a novel proto-oncogene in non-small cell lung cancer, promotes tumor growth via directly transactivating CCND1. *Scientific Reports*, *7*(1), 591. https://doi.org/10.1038/s41598-017-00698-4

You, W., Li, L., Sun, D., Liu, X., Xia, Z., Xue, S., Chen, B., Qin, H., Ai, J., & Jiang, H. (2019). Farnesoid X Receptor Constructs an Immunosuppressive Microenvironment and Sensitizes FXRhighPD-L1low NSCLC to Anti–PD-1 Immunotherapy. *Cancer Immunology Research*, *7*(6), 990–1000. https://doi.org/10.1158/2326-6066.CIR-17-0672

Yu, L., Liu, Y., Wang, S., Zhang, Q., Zhao, J., Zhang, H., Narbad, A., Tian, F., Zhai, Q., & Chen, W. (2023). Cholestasis: exploring the triangular relationship of gut microbiota-bile acid-cholestasis and the potential probiotic strategies. *Gut Microbes*, *15*(1). https://doi.org/10.1080/19490976.2023.2181930

Zaitseva, J., Jenewein, S., Jumpertz, T., Holland, I. B., & Schmitt, L. (2005). H662 is the linchpin of ATP hydrolysis in the nucleotide-binding domain of the ABC transporter HlyB. *The EMBO Journal*, *24*(11), 1901–1910. https://doi.org/10.1038/sj.emboj.7600657

Zgarbová, M., Šponer, J., & Jurečka, P. (2021). Z-DNA as a Touchstone for Additive Empirical Force Fields and a Refinement of the Alpha/Gamma DNA Torsions for AMBER. *Journal of Chemical Theory and Computation*, *17*(10), 6292–6301. https://doi.org/10.1021/acs.jctc.1c00697

Zhang, Y., Jackson, J. P., St. Claire, R. L., Freeman, K., Brouwer, K. R., & Edwards, J. E. (2017). Obeticholic acid, a selective farnesoid X receptor agonist, regulates bile acid homeostasis in sandwich-cultured human hepatocytes. *Pharmacology Research & Perspectives*, *5*(4), e00329. https://doi.org/10.1002/prp2.329

Zhang, Yanqiao, Kast-Woelbern, H. R., & Edwards, P. A. (2003). Natural Structural Variants of the Nuclear Receptor Farnesoid X Receptor Affect Transcriptional Activation. *Journal of Biological Chemistry*, *278*(1), 104–110. https://doi.org/10.1074/jbc.M209505200

Zhang, Z., Miteva, M. A., Wang, L., & Alexov, E. (2012). Analyzing Effects of Naturally Occurring Missense Mutations. *Computational and Mathematical Methods in Medicine*, *2012*, 1–15. https://doi.org/10.1155/2012/805827

Zheng, W., Lu, Y., Tian, S., Ma, F., Wei, Y., Xu, S., & Li, Y. (2018). Structural insights into the heterodimeric complex of the nuclear receptors FXR and RXR. *Journal of Biological Chemistry*, *293*(32), 12535–12541. https://doi.org/10.1074/jbc.RA118.004188

Ziol, M., Barbu, V., Rosmorduc, O., Frassati-Biaggi, A., Barget, N., Hermelin, B., Scheffer, G. L., Bennouna, S., Trinchet, J. C., Beaugrand, M., & Ganne-Carrié, N. (2008). ABCB4 Heterozygous Gene Mutations Associated With Fibrosing Cholestatic Liver Disease in Adults. *Gastroenterology*, *135*(1), 131–141. https://doi.org/10.1053/j.gastro.2008.03.044