



U i T

**THE ARCTIC
UNIVERSITY
OF NORWAY**

Faculty of Health Sciences

Department of Pharmacy

Molecular modelling of the androgen metabolising enzyme UDP-glucuronosyltransferase 2B17

Ingmar Trane

Master thesis in pharmacy – May 2018



Acknowledgements

This master thesis was written at the Department of Medical Biology, Faculty of Health Sciences, in collaboration with the Department of Pharmacy at UiT - The Arctic University of Norway from August 2017 to May 2018.

I would like to express my sincere gratitude to my main supervisor Prof. Aina Westrheim Ravna, for her guidance and support during this research. With her help I have deepened my knowledge and skills in the molecular modelling field. I would also like to thank my co-supervisor Prof. Georg Sager, for the support and constructive feedback, and for helping me in the writing process.

I also wish to express my gratitude to Andrew Orry, Senior Scientist at Molsoft LLC, for patiently answering all my questions about the computational commands and scripts. This master thesis could not have been completed without his help.

Finally, I would like to thank my girlfriend for her immense support and encouraging words through the process of writing this thesis.

Tromsø, May 2018

Ingmar Trane

Abstract

The enzyme UDP-glucuronosyltransferase 2B17 (UGT2B17) has a major role in androgen metabolism, being involved in the conjugation of both testosterone and its potent metabolite dihydrotestosterone. The enzyme catalyses the transfer of glucuronic acid from UDP-glucuronic acid to the lipophilic androgen substrate. As a consequence, the water solubility of the substrate is increased, and it is more easily excreted from the body. Testosterone levels are important for male fertility and vitality, and are at their highest during adolescence and early adulthood. As some men get older their testosterone levels gradually decline. Other factors that can affect the levels are nutrients, exercise, lifestyle factors, drugs and endocrine disruptors. Inhibitors of the UGT2B17 could help maintain normal testosterone levels in patients with declining levels caused by various factors.

Homology modelling is an *in silico* approach used to predict the 3D structure of an unknown protein structure, based on evolutionary related templates. An experimentally determined crystal structure of UGT2B17 had not been solved at the time of this study, consequently four homology models were constructed and refined using ICM. Molecular docking of inhibitors and decoys on the models was performed to gain insights in the interactions between ligand and binding site. Five residues in the binding pocket were proposed for future site-directed mutagenesis studies. The ability of the models to discriminate between inhibitors and decoys was evaluated using receiver operating characteristics curves, and the most accurate model was studied further with virtual ligand screening (VLS). Model_4AMG was identified as the most accurate, and VLS was performed on the model to screen structures from a chemical database for potential hit compounds. A hitlist of 25 compounds were identified as potential drug candidates, pending future *in vitro* testing to determine their binding affinity for UGT2B17.

Table of contents

Acknowledgements	3
Abstract	5
Table of contents	6
Abbreviations	8
1. Introduction	10
1.1 Endocrinology	10
1.1.1 Steroid hormones.....	10
1.1.2 Androgens.....	11
1.2 Metabolism	14
1.2.1 Functionalization and conjugation.....	14
1.2.2 Androgen metabolism.....	15
1.3 Pharmacodynamics	16
1.3.1 Drug targets.....	16
1.3.2 Drug binding interactions.....	17
1.3.3 Affinity.....	19
1.3.4 Drug-like properties.....	19
1.4 Proteins and protein structure	20
1.4.1 Glycosyltransferase.....	22
1.4.2 UDP-glucuronosyltransferase.....	23
1.5 Computational science and drug discovery	25
1.5.1 Molecular modelling.....	25
1.5.2 Homology modelling.....	26
1.5.3 Docking and scoring.....	29
1.5.4 Receiver operating characteristic (ROC) curves.....	30
1.5.5 Virtual ligand screening.....	31
2. Aim of the study	32
3. Methods	33
3.1 Software and databases	33
3.1.1 Molsoft Internal Coordinates Mechanics (Version 3.8.7).....	33
3.1.2 The Protein Data Bank.....	33
3.1.3 Universal Protein Resource Knowledgebase.....	33

3.1.4	Basic Logical Alignment Search Tool	34
3.1.5	Structural Analysis and Verification Server v5.0	34
3.1.6	PubChem	34
3.1.7	ChEMBL.....	35
3.1.8	DecoyFinder 2.0.....	35
3.1.9	eMolecules.....	35
3.2	Homology modelling.....	36
3.2.1	Template identification	36
3.2.2	Sequence Alignment	37
3.2.3	Model building.....	43
3.2.4	Model refinement	43
3.2.5	Model validation	43
3.3	Molecular Docking	44
3.3.1	Inhibitors and decoys	44
3.3.2	Ligand and model preparation.....	46
3.3.3	Identification of ligand binding pocket.....	46
3.3.4	Docking of inhibitors and decoys.....	46
3.3.5	Evaluation of docking	47
3.4	Virtual Ligand Screening	48
4.	Results and discussion	49
4.1	Homology modelling.....	49
4.1.1	Sequence alignment and model building.....	49
4.1.2	Model validation	52
4.2	Molecular docking.....	58
4.2.1	Identification of ligand binding pocket.....	58
4.2.2	Docking of inhibitors and decoys.....	59
4.2.3	Evaluation of docking	62
4.3	Virtual ligand screening.....	68
4.4	Future aspects.....	71
5.	Conclusion	72
6.	References.....	73

Abbreviations

2D	Two-dimensional
3D	Three-dimensional
3 β -HSD	3 β -hydroxysteroid dehydrogenase
17 β -HSD	17 β -hydroxysteroid dehydrogenase
AUC	Area under the curve
ADME	Absorption, distribution, metabolism and excretion
BLAST	Basic Local Alignment Search Tool
C $_{\alpha}$	Central carbon atom
CT	Carboxy terminal
CYP	Cytochrome P450
Da	Dalton
DHT	Dihydrotestosterone
DHEA	Dehydroepiandrosterone
DNA	Deoxyribonucleic acid
EC	Enzyme Commission number
E-value	Expectation value
E $_{\text{angle}}$	Angle bending energy
E $_{\text{bond}}$	Bond length energy
E $_{\text{tors}}$	Torsion energy
E $_{\text{tot}}$	Total steric energy
E $_{\text{vdw}}$	van der Waals energy
E $_{\text{elec}}$	Electrostatic energy
ER	Endoplasmic reticulum
FH	Follicle-stimulating hormone
FN	False negative
FP	False positive
FPR	False positive rate
GnRH	Gonadotropin-releasing hormone
GT	Glycosyltransferase
HBA	Hydrogen bond acceptor
HBD	Hydrogen bond donor
HSD	Hydroxysteroid dehydrogenase
IC $_{50}$	Half maximal inhibitory concentration
ICM	Internal Coordinate Mechanics
K $_{\text{D}}$	Dissociation constant

K _i	Binding inhibition constant
LBDD	Ligand-based drug design
LH	Luteinizing hormone
LogP	Octanol-water partition coefficient / hydrophobicity
MM	Molecular mechanics
MW	Molecular weight
NMR	Nuclear magnetic resonance
NT	Amino terminal
PAINS	Pan-assay interference compounds
PDB	Protein Data Bank
PDB id	Protein Data Bank identification code
PSA	Polar surface area
QM	Quantum mechanics
RB	Rotatable bonds
RMSD	Root-mean-square deviation
ROC	Receiver operating characteristics
SAVES	Structural analysis and verification server
SHBG	Sex hormone-binding globulin
SMILES	Simplified molecular-input line-entry system
SBDD	Structure-based drug design
TN	True negative
TP	True positive
TPR	True positive rate
UDP	Uridine diphosphate
UDPGA	Uridine diphosphate glucuronic acid
UDP-GlcNAc	Uridine diphosphate N-acetylglucosamine
UGT	UDP-glucuronosyltransferase
UGT2B7	UDP-glucuronosyltransferase isoform 2B7
UGT2B15	UDP-glucuronosyltransferase isoform 2B15
UGT2B17	UDP-glucuronosyltransferase isoform 2B17
UniProtKB	Universal Protein Knowledge Base
VLS	Virtual ligand screening
X-O-GA	Glucuronide product
X-OH	Aglycone
Å	Ångström

1. Introduction

1.1 Endocrinology

Endocrinology is the study of the medical aspects of endocrine glands and hormones. The endocrine system is one of two systems that regulate the communication and signalling between cells in the body, the other being the nerve system. The communication through these systems makes it possible for different cells to adjust their activity according to the needs of the body. The endocrine system consists of several endocrine glands in different parts of the body that synthesize, store and secrete hormones that act as messengers. The major endocrine glands in the body are hypothalamus, pituitary gland, pineal gland, thyroid gland, parathyroid gland, adrenal gland, pancreas and reproductive glands. Hormones are chemicals that are transported by the bloodstream from an endocrine gland to tissue or organs to regulate a wide range of physiological processes. The main processes regulated by hormones are; development and growth, regulation of metabolism and nutrient supply, reproduction and maintenance of internal environment. ^{1,2}

Hormones are grouped into three chemical classes based on their structure: (1) peptides/proteins, (2) amines, and (3) steroids. The most numerous are the peptides or protein hormones, which consists of chains of amino acids that vary in length. The pituitary gland, pancreas, parathyroid gland and the intestines synthesize the peptide hormones. Examples are growth hormone, insulin and prolactin. Amine hormones are derived from either tyrosine or from tryptophan. They are produced in the thyroid gland and the adrenal cortex, examples are thyroxin and catecholamines.

Steroid hormones are hormones formed by stepwise transformation of cholesterol. Major sites of steroid production include the adrenal cortex, the gonads and placenta. Examples are testosterone, aldosterone and cortisol. ¹

The endocrine system is regulated by feedback mechanisms to ensure appropriate hormonal secretion. Often one hormone controls the action or secretion of another through negative or positive feedback loops. This controlled release of hormones helps maintaining homeostatic balance in the body. ³

1.1.1 Steroid hormones

Steroid hormones are lipophilic molecules that act on a wide range of tissues and influence many aspects of the normal physiology including sexual differentiation, metabolism, osmoregulation and reproduction. These hormones are synthesized and secreted from the adrenal cortex, testes, ovaries, and placenta. There are five major classes of steroid hormones: (1) glucocorticoids, (2) mineralocorticoids, (3) androgens, (4) oestrogens, and (5) progestogens, which contain 21,21,19,18

and 21 carbons, respectively. All steroid hormones have similar structures with a four-ringed carbon backbone, and are derived from a common cholesterol precursor with 27 carbons.³

Steroid hormones can be classified as either endocrine (distant target tissue), paracrine (neighbouring cell) cells, or autocrine (same cell), based on the distance of the target site from the site of synthesis and secretion. Steroid hormones are transported through the blood in a bound state because they are poorly soluble lipids. They are bound to specific water soluble carrier plasma proteins, examples are sex hormone-binding globulin (SHBG), corticosteroid-binding globulin or albumin. In the blood, about 90-99% of the steroid hormones are bound to transport proteins. A small amount of the hormones exists in their active state as free hormones, dissolved in plasma and not bound to carrier protein. The free hormone can leave the bloodstream and diffuse across the membrane to the target cells. Consequently, it's the free unbound concentration that triggers the biological effect of the hormones.^{2,3}

Steroid hormones can affect their target cells through many different mechanisms. These different pathways can be classified as having a genomic, or a non-genomic effect. Genomic effects are slow and result in altering gene transcription, these effects can take from hours to days.

Non-genomic effects results are much faster, and are involved in the rapid activation of a variety of cell-signalling molecules. These can occur within seconds to minutes after administration through non-genomic mechanisms.³

1.1.2 Androgens

Testosterone is one of the androgen hormones and is the primary male hormone. Other important male androgens are dihydrotestosterone (DHT), a more potent metabolite of testosterone, and androstenedione, a precursor for testosterone. Together these androgens play a key role in the male pubertal development of testes and prostate, as well as promoting masculine characteristics such as increased muscle and bone mass, height, and the growth of body hair. Through adolescence testosterone helps maintain the libido, sperm production, muscle and bone mass, and male hair pattern. In females, androgens are present but in lower levels, playing more subtle roles, affecting libido and sexual arousal. Androgens are also precursors for oestrogens in men and women.^{2,4}

The biosynthesis of the androgens is like other steroid hormones derived from a common cholesterol precursor. Dehydrogenases and cytochrome P450 (CYP) enzymes are involved in the multi-step synthesis, as seen in figure 1. The majority of testosterone, above 95%, is produced by the Leydig cell in the testes in men, while the adrenal cortex accounts for most of the remainder.³

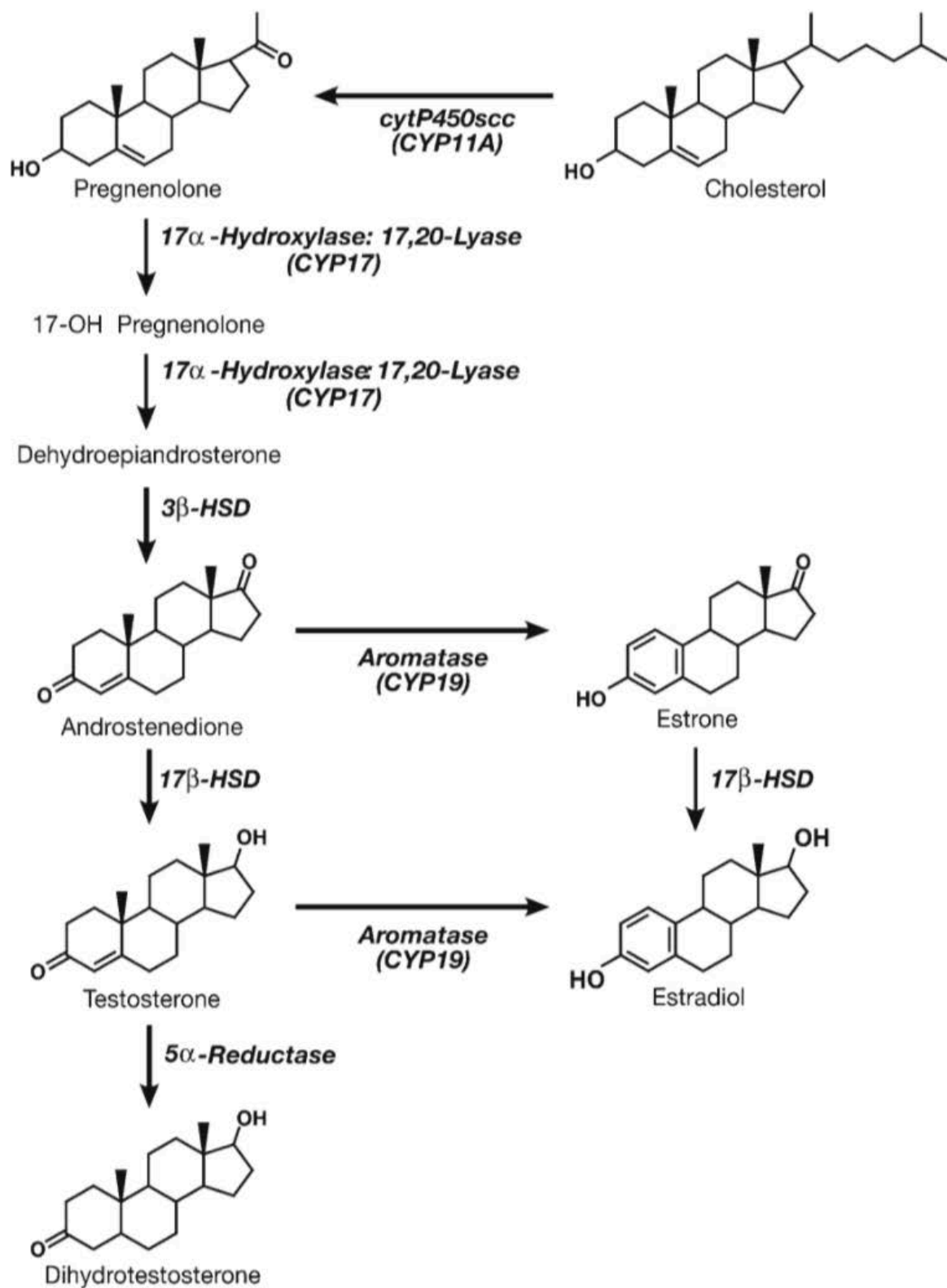


Figure 1 Biosynthesis of androgens. The first step involves the conversion of cholesterol to Pregnenolone by cytochrome P450-side-chain-cleavage, which is the rate limiting step of the synthesis, regulated by luteinizing hormone. In the next two steps two additional carbon atoms are removed by the CYP17 enzyme to yield 17 α -hydroxypregnenolone and dehydroepiandrosterone (DHEA). In the following step DHEA is converted to androstenedione by 3 β -hydroxysteroid dehydrogenase (3 β -HSD). In the final step androstenedione is converted to testosterone by 17 β -hydroxysteroid dehydrogenase (17 β -HSD). In some tissues testosterone can be converted to the metabolite DHT. Testosterone and androstenedione can also be converted into oestrogens by aromatase.³ (Reprinted from *Endocrinology: Basic and Clinical Principles* by Melmed et al. 2nd ed. Totowa: Humana Press; 2005)

The production of testosterone is regulated by luteinizing hormone (LH) and follicle-stimulating hormone (FSH) secreted by the anterior pituitary. When testosterone levels are low, gonadotropin-releasing hormone (GnRH) is released by the hypothalamus which stimulates the anterior pituitary to release LH and FSH. Testosterone is synthesized and secreted primarily by the Leydig cells of the testes. The number of Leydig cells is in turn regulated by LH. In case of elevated circulating levels of testosterone, negative feedback loops act on the hypothalamus and the anterior pituitary to inhibit the release of GnRH and FSH/LH, respectively, as shown in figure 2. ²

Testosterone levels are important for male fertility and vitality. The testosterone levels are at their highest during adolescence and early adulthood. As some men get older their testosterone levels gradually decline. Other factors that can affect the levels are nutrients, exercise, lifestyle factors, drugs and endocrine disruptors. ⁴⁻⁷

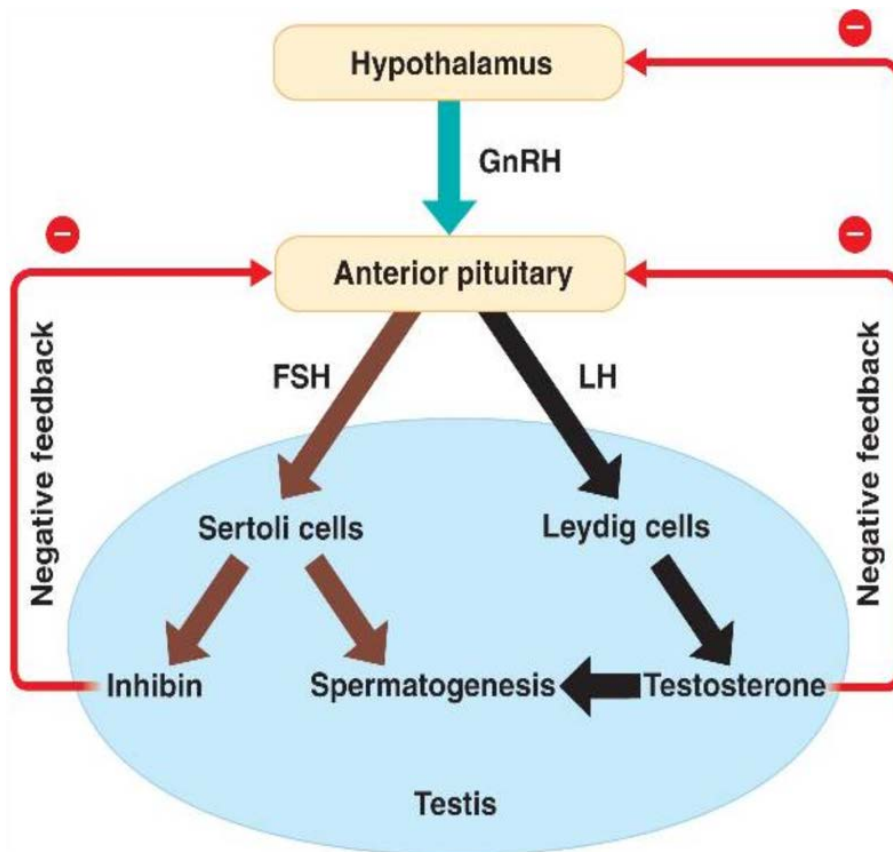


Figure 2 Hormonal control of testes (Reprinted from *Biology by Campbell et al. 8th ed. San Francisco: Pearson/Benjamin Cummings; 2008*)

1.2 Metabolism

Metabolism is the build-up or breakdown of chemical substances by enzymatic conversion within the body. The main purpose of metabolism is generally conversion of the foreign substance to energy or building blocks, or degradation or modification so that it can be more easily excreted. A substance that has undergone a metabolic reaction is called a metabolite. The main routes foreign substances and their metabolites can leave the body are the kidneys, the hepatic system and the lungs. ⁸

1.2.1 Functionalization and conjugation

Drug metabolism is divided into two kinds of reactions, known as phase I and phase II, or functionalization and conjugation, as shown in figure 3. The phase I reactions may occur by oxidation, reduction or hydrolysis, and are carried out by enzymes such as cytochrome P450. The phase I reactions involve the addition of polar or reactive functional groups to the foreign compound, making the compound more chemically reactive. This functional group then serves as the point of attack for the phase II conjugation reaction. The phase II reactions involve the addition of highly polar molecules to the functional group from phase I, the resulting conjugate is more easily excreted. Glucuronidation is the most common of these conjugation reactions, others are methylation, sulphation, acetylation, glutathione, and glycine conjugation. Glucuronidation and sulphation are both phase II reactions and they account for 50% of the metabolism of testosterone and DHT. Another 40% of testosterone is metabolized by the combined actions of 5 α -reductase, 5 β -reductase, 3 α -hydroxysteroid dehydrogenase and 17 β -HSD. ⁹

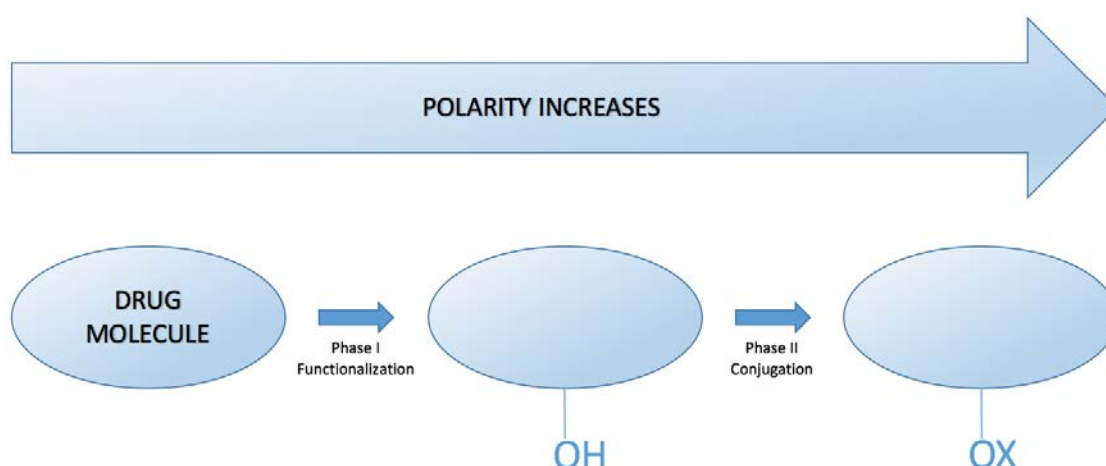


Figure 3 Schematic presentation of phase I and phase II reactions. Molecules can undergo both phase I and phase II reactions, or only one of the two. Oxygen (O) is shown in the figure, but it can also be nitrogen (N) or sulphur (S).

1.2.2 Androgen metabolism

The most important androgens, testosterone and DHT, are mainly metabolised in the liver. The main enzymes responsible for the glucuronidation reactions are part of the uridine-diphospho (UDP) glucuronosyltransferase family (UGT, EC 2.4.1.17). UGTs are the main phase II enzymes and they have an important role in the detoxification of endogenous and exogenous compounds in humans.^{3,10}

The role of the UGTs is to catalyse the transfer of a glucuronyl group to a lipophilic substrate following the phase I reaction, forming a more polar water soluble, less toxic and more rapidly excreted compound. Prior to the glucuronidation reaction, the substrates are referred to as aglycones. The glucuronyl group transferred is mainly a glucuronic acid moiety from the uridine-diphospho-glucuronic acid (UDPGA) co-substrate, as shown in figure 4. UGTs utilizes two substrates, aglycone and UDPGA, and forms two products, glucuronide and UDP. The enzyme mechanism is considered a compulsory ordered mechanism where UDPGA binds first.^{10,11}

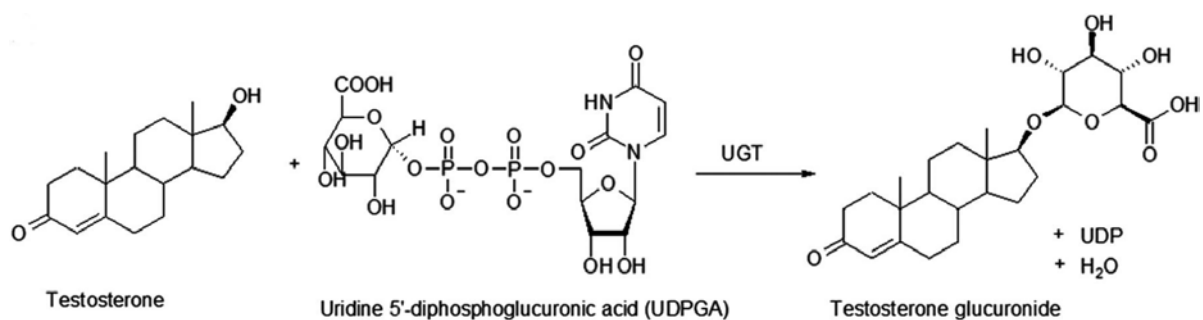


Figure 4 Schematic illustration of the testosterone glucuronidation reaction¹² (Reprinted from *Confounding factors and genetic polymorphism in the evaluation of individual steroid profiling* by Kuuranne et al. *British journal of sports medicine*. 2014)

1.3 Pharmacodynamics

Pharmacodynamics is the study of biological effects resulting from interactions between drugs and the biological system, with focus on how a drug affects the body. Most drugs assert their action by interacting with drug targets, thereby triggering an effect direct or indirect through a cascade of reactions. These are biochemical effects in cells and physiological effects in tissue and organs. ¹³

1.3.1 Drug targets

The main molecular targets for drugs are proteins, nucleic acids and lipids. These are macromolecules with molecular weight (MW) much larger than the typical drug. The interaction between a drug and macromolecule target involves a dynamic process called binding, which is both structure and stereospecific. This process takes place at a specific area of the macromolecule, called the binding site or active site as shown in figure 5. The binding site is usually a groove or a pocket on the surface of the macromolecule, allowing the drug to sink into the body of the larger molecule, forming a complex. A drug molecule that binds to a target macromolecule and forms a complex is called a ligand. The ligand-protein binding is often explained as “key in a lock”. ^{8,9,13}

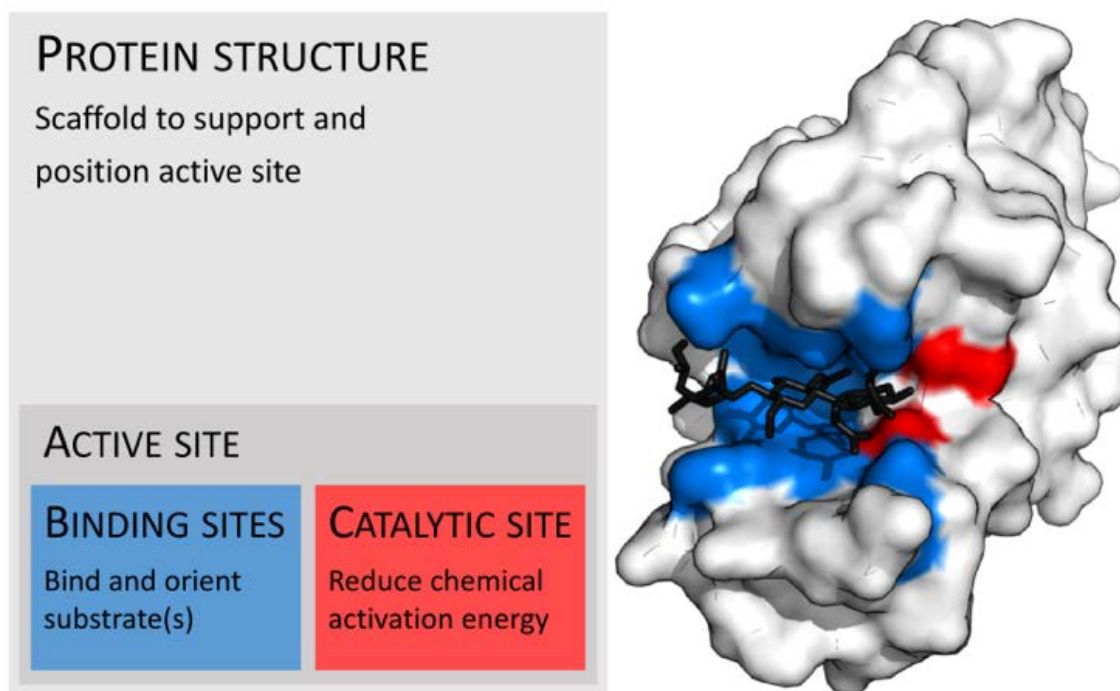


Figure 5 Active site of a protein (Retrieved from [wikipedia.org/wiki/active_site](https://en.wikipedia.org/wiki/active_site), Public Domain)

There are four main types of drug targets available for ligands to bind: (1) enzymes, (2) receptors, (3) transport proteins and (4) ion channels, all of these are proteins. ⁸

Enzymes are proteins that control chemical reactions in the body by acting as catalysts. They accelerate the chemical reactions without being consumed themselves. Most of the metabolic processes in the cells need enzyme catalysis in order to occur at rates fast enough to sustain life. Enzymes must bind to their substrates before they can catalyse a chemical reaction. The substrate specificity is determined by the binding site. The catalytic site is responsible for interacting with the substrate to lower the activation energy of the reaction. Enzymes can undergo conformational changes upon substrate binding, and in doing so close around the substrate to initiate catalysis. The catalysis takes place in the closed form, and the enzyme opens up again to release the product. Many enzymes require additional non-protein substances called co-factors for the reaction to take place. Enzyme activity can be affected by drugs in different ways, either by increasing or decreasing the activity.^{8,9}

Receptors are protein molecules that bind to signalling molecules from outside the cell. The binding causes a conformational change in the receptor, which triggers a cascade of cellular responses. Receptors are the most important drug targets and show a great variety. Based on molecular structure and transduction mechanism, receptors can be divided into four types: (1) Ligand-gated ion channels, (2) G-protein coupled receptors, (3) Kinase-linked and related receptors and (4) Nuclear receptors.^{8,13}

Transport proteins are responsible for the movement of ions and small organic molecules across cell membranes. Na⁺, Ca⁺, amino acids, neurotransmitters and catechol-amines are examples of molecules being transported. Drugs acting on transport proteins block the transport mechanism by either inhibiting the transport or acting as false substrates.⁸

Ion channels are gateways through cell membranes that selectively allow particular ions to pass between the inside and outside of the cell. They are induced to open or close through a variety of mechanisms, and are vital to many biological functions including membrane potentials. Drugs can affect ion channel function by either binding to the channel itself or by indirect interaction involving G-proteins or other intermediaries.⁸

Drugs can bind to drug targets in different ways, they can be agonists or antagonists. A ligand that binds to a target and triggers a strong biological response is called a full agonist. On the other hand, a ligand that binds to a target and inhibits an effect by blocking it, is called an antagonist or inhibitor. Another type of ligand is a partial agonist, which due to low efficacy only triggers a partial response.⁸

1.3.2 Drug binding interactions

The interactions that bind drugs to the active sites of drug targets are strong covalent interactions or weak intermolecular interactions. Covalent bonds are irreversible and occur when two atoms share a pair of electrons. These interactions are very strong, with a bond strength of 200-400 kJ mol⁻¹.

Intermolecular interactions are the most common between drug and target. These interactions are reversible and much weaker ($0.5\text{-}70\text{ kJ mol}^{-1}$), since the atoms are not directly bound to each other. The bonds can be formed, then broken again. This means there is an equilibrium between the drug being bound and unbound to target. The binding forces are strong enough to hold the substrate for a certain period, but weak enough to allow it to depart once it has done its job. The number of interactions between drug and target influence the length of time the drug remains bound. The relative strength of the different intermolecular forces in place is also an important factor. Intermolecular interactions include ionic bonds, hydrogen bonds, dipole-dipole, and ion-dipole interactions, as well as van der Waals interactions. ⁹

Ionic bonds are electrostatic interactions between groups of opposite charges. The strength of the interaction increases with the charge of the groups, and with shorter distance between the groups. The environment also affects the strength, being stronger in hydrophobic environments than in polar environments. These are the strongest of the intermolecular interactions, with a binding energy of $20\text{-}40\text{ kJ mol}^{-1}$.

Dipole-dipole interactions occur when a positive region of one molecule (dipole) attracts the negative region of a second molecule. The different charges are a result of different electronegativity of the atoms and functional groups present in the molecule. This is a weak type of interaction; it can have a binding energy of $0.5\text{-}3\text{ kJ mol}^{-1}$.

Ion-dipole interactions occur when a charged or ionic group in one molecule interacts with a dipole in a second molecule. The binding energy of this type of interaction is $3\text{-}10\text{ kJ mol}^{-1}$.

Hydrogen bond interactions takes place between a hydrogen atom covalently bound to an electronegative atom (O, N, F, or Cl), and another adjacent atom with a lone pair of electrons. As the electronegative atom has greater attraction for electrons, the electron distribution in the covalent bond is weighted against the electronegative atom, giving the hydrogen a slight positive charge. Such a hydrogen can act as a hydrogen bond donor (HBD). The electron rich adjacent atom that receives the hydrogen bond is called hydrogen bond acceptor (HBA). Hydrogen bonds are angle dependent (180°) and can be influenced by water. The binding energy of hydrogen bonds are moderate in strength, varying from $16\text{-}60\text{ kJ mol}^{-1}$.

Van der Waals interactions are very weak interactions between hydrophobic regions in different molecules, such as aliphatic substituents or the overall carbon skeleton. These interactions are independent of direction, but are distance dependent. This type of interaction has a binding energy of $2\text{-}4\text{ kJ mol}^{-1}$. Although these interactions are individually weak, there may be many such interactions between a drug and its target, so the overall contribution is often crucial to binding. ⁹

1.3.3 Affinity

The correlation between ligand concentration at the binding site, and the resulting effect is a central aspect of pharmacodynamics. Affinity is defined as the extent of binding of a ligand to a receptor. Higher affinity equals stronger binding, and consequently more effect. The affinity between ligand and receptor is described by the equilibrium dissociation constant (K_D), as shown in the equation:

$$K_D = \frac{k_{off}}{k_{on}} = \frac{[L] \times [R]}{[L-R]}$$

k_{off} = Rate constant for dissociation
 k_{on} = Rate constant for association
[L] = concentration of free ligand
[R] = concentration of free receptor
[L-R] = concentration of ligand-receptor complex

The equation shows that K_D equals the ligand concentration needed to occupy 50% of the receptors. The smaller the K_D value, the greater the binding affinity of the ligand for its target. The larger the K_D value, the weaker the target and ligand are attracted to and bind to one another. Drugs on the market usually have affinities in nanomolar range, usually around $10^{-8}M$ (10nM).⁹

1.3.4 Drug-like properties

Drug-like properties are physiochemical properties that are essential for the bioavailability of a drug intended for oral administration. These drug-like properties includes molecule size, number of HBA, number of HBD, hydrophobicity, polar surface area (PSA), and number of rotatable bonds (RB).

The MW should be less than 500Da, which equals about 36 heavy atoms (C, N, O, S). The optimal size is about 25-30 heavy atoms for good affinity, more or less heavy atoms would affect the affinity. The number of HBA, expressed as the sum of N and O in the molecule, should be no more than ten. The number of HBD, expressed as the sum of OH and NH in the molecule, should be no more than five. The hydrophobicity of a molecule is measured by logP, and should be no more than five, giving solubility in both fat and water. The PSA is defined as the surface of all polar atoms, primarily N and O, including their attached hydrogens, and should be less than 140\AA^2 . The number of rotatable bonds describes the molecular flexibility, and should be no more than ten, since too many rotatable bonds would give a vast number of conformations. Poor passive absorption or permeability of a drug is more likely if the drug violate two or more of these rules.

These rules describe the physiochemical properties needed for a drug's pharmacokinetics in the human body, including their absorption, distribution, metabolism, and excretion (ADME). However, druglikeness does not predict if a compound is pharmacologically active. The druglikeness value calculated by the ICM software is a prediction based on drug-like properties, and a value less than zero indicates that the compound may have some non-drug-like properties.^{14,15}

1.4 Proteins and protein structure

Proteins are macromolecules consisting one or many chains of amino acid residues. About half the mass of the human body is built up of proteins. They perform a vast array of functions within an organism, including catalysing chemical reactions, DNA replication, transporting molecules, providing structure and support for cells, and responding to stimuli. All proteins are built up by amino acid residues.

There are twenty natural amino acids, all have in common a central carbon atom (C_α) to which are attached a hydrogen atom (H), an amino group (NH_2), and a carboxyl group ($COOH$). What distinguishes one amino acid from another is the sidechain (R) attached to the C_α . Amino acids are joined together by peptide bonds during protein synthesis when the carboxyl group of one amino acid condenses with the amino group of the next to eliminate water, this process is repeated forming a polypeptide or protein. This succession of residues linked by peptide bonds is called a backbone or main-chain, and from this backbone the various sidechains project, as seen in figure 6. The conformation of the whole backbone and the folding of a protein is determined by two conformational angles, phi (ϕ) and psi (ψ), for each residue. Because of steric hindrance between backbone and sidechains, only certain combinations of these angles are allowed. ¹⁶

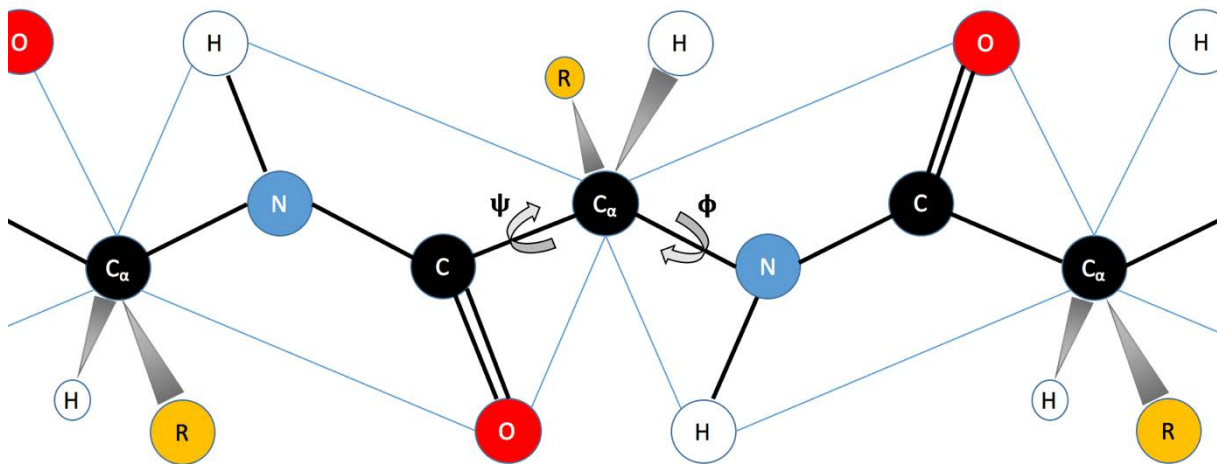


Figure 6 Backbone of amino acid residues joined together by peptide bonds. Conformational angles ψ and ϕ determine the planarity of the backbone. The blue boxes represent the planar nature of the peptide bonds. Sidechains are shown as R.

Amino acids are often divided into four different classes defined by the chemical properties of the side chain: (1) hydrophobic, (2) polar, (3) aromatic and (4) charged. Their names are abbreviated by three-letter and one-letter codes, given in table 1.

Table 1 Classification of amino acids

Hydrophobic		Glycine (Gly/G), Alanine (Ala/A), Valine (Val/V) Leucine (Leu/L), Methionine (Met/M), Isoleucine (Iso/I)
Polar		Serine (Ser/S), Threonine (Thr/T), Cysteine (Cys/C) Proline (Pro/P), Asparagine (Asn/N), Glutamine (Gln/Q)
Aromatic		Phenylalanine (Phe/F), Tyrosine (Tyr/Y), Tryptophan (Trp/W)
Charged	Positively	Lysine (Lys/K), Histidine (His/H), Arginine (Arg/R)
	Negatively	Glutamine (Glu/E), Aspartate (Asp/D)

Proteins differ from each other primarily in their amino acid sequences, which results in the protein folding into a specific three dimensional (3D) structure that determines its function. Proteins have four levels of structure: (1) primary, (2) secondary, (3) tertiary, and (4) quaternary, as shown in figure 7. The primary structure is the amino acid sequence of a protein's polypeptide chain. The secondary structure consists of regions of ordered structure elements called α -helix and β -sheet. The tertiary structure is the overall 3D shape of the protein and is formed by folding secondary structure elements, into compact globular units called domains, or in an ordered shape. The quaternary structure is several polypeptide chains (subunits) arranged into the functional protein. Secondary, tertiary, and quaternary structures are formed to maximize favourable intermolecular and intramolecular bonds and to minimize unfavourable interactions, thus stabilizing the protein.^{9,16}

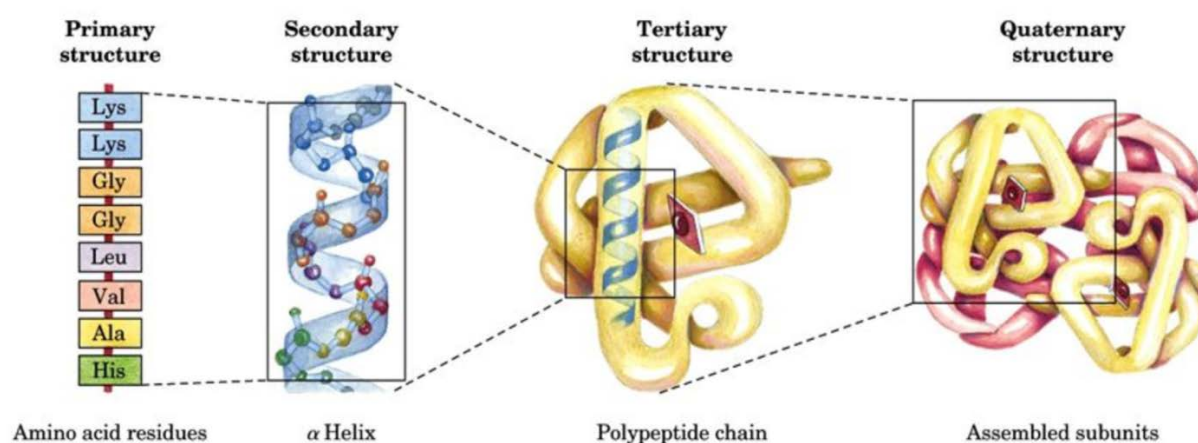


Figure 7 Different levels of protein structure (Reprinted from Introduction to protein structure by Brändén & Tooze, 2nd ed. New York, Garland Pub, 1999)

1.4.1 Glycosyltransferase

Based on amino acid sequence and predicted structure, human UGTs belong to glycosyltransferase (GT) superfamily. (EC 2.4) GTs are enzymes that transfer sugars to other molecules from an activated nucleotide sugar donor, mainly UDP-glucose. Two structural folds have been identified for the enzymes, GT-A and GT-B, as shown in figure 8.

The fold of GT-A proteins consists of a $\alpha/\beta/\alpha$ sandwich resembling a Rossmann-like domain, and also contains a divalent metal binding motif that is important to ligand binding. The GT-B folds have structures that are built up of two separate Rossmann-like domains that associate to form a catalytic site in the cleft between the domains. The two domains are connected through a flexible hydrophobic linker region. The amino-terminal (NT) domain binds the substrate, and the carboxy-terminal (CT) domain binds the nucleotide-sugar donor. In contrast to GT-A fold, the activities of GT-B are not dependent on metals. The structural conservation between homologous members of the GT-B family is excellent, particularly the CT domain.¹⁷

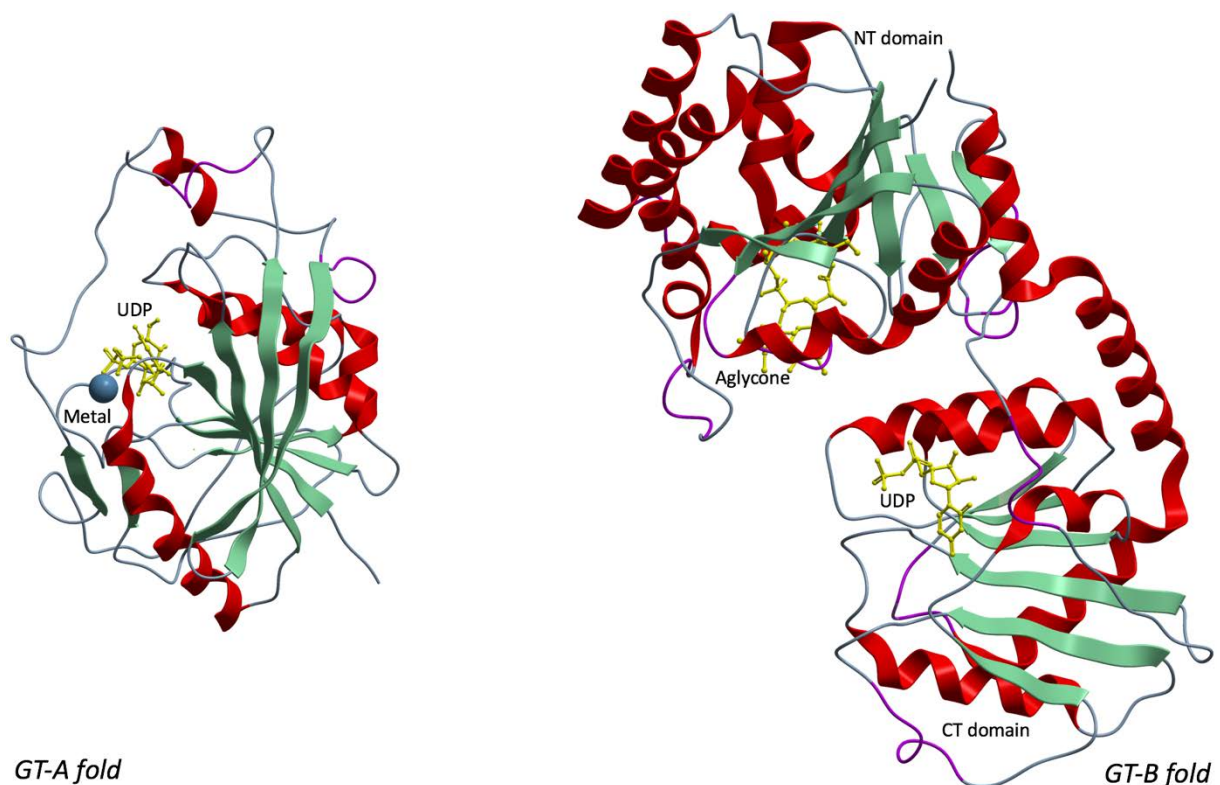


Figure 8 Cartoon representation of glycosyltransferases with GT-A and GT-B folds. α -helices shown as red, and β -sheets shown as green, while substrates are shown as yellow.

1.4.2 UDP-glucuronosyltransferase

UGTs are responsible for the transfer of a glucuronic acid moiety from UDPGA to a wide range of structurally unrelated substances possessing hydroxyl-, amino-, carboxyl-, or sulfhydryl groups, converting them to water soluble molecules. UGTs belong to the GT family and are thought to adopt a GT-B fold. The UGTs are membrane bound enzymes of approximately 530 amino acid residues. The majority of UGTs are localized in the endoplasmic reticulum (ER), as shown in figure 9, and some are found in nuclear membranes.¹⁸

There are currently 21 different human UGTs known based on sequence and gene organisation, these are divided into families UGT1, UGT2, and UGT3. Substrate specificity varies between the isoforms of UGTs, some are relatively strict, while others accept a wide variety of structurally unrelated substrates, in addition one substrate is usually glucuronidated by several isoforms. UGTs from gene family UGT1 and UGT2 all utilize UDPGA as cofactor, while the enzymes of UGT3 family prefer UDP-N-acetylglucosamine (UDP-GlcNAc), UDP-glucose or UDP-xylose. Androgens can be metabolized by three different isoforms of the UGT2 family, UGT2B7, UGT2B15 and UGT2B17, with the latter being the most efficient. UGT2B17 also have the ability to conjugate DHT, making it the most important androgen conjugating enzyme and the focus of this study.^{10,18,19}

The enzymes are composed of two functional domains, a highly variable NT domain (residues 1-265) and a highly conserved CT domain (residues 266-530), with a catalytic site in the cleft between. The NT domain contains a signal peptide that mediates the integration into the ER-lumen, the aglycone binding site, and a membrane interacting region. The CT domain contains most of the UDPGA co-factor binding site and a transmembrane helix near the carboxy-terminus with a cytosolic tail. The enzyme is predicted to form dimers in endoplasmic reticulum membranes, this may have an effect on function and acceptor ligand specificity.^{10,18,20}

An important region of GT-B fold enzymes is the conserved diphosphate nucleotide binding site formed by the CT domain. The structural similarity between GTs in this area is remarkably high, with a highly conserved 44 residues long region (residues 357-400) making up most of the binding pocket. Most mammalian UGTs binds the co-factor UDPGA, while GTs of plants and bacteria utilize other nucleotide-sugars as co-factor, mainly UDP-glucose.

The highly variable NT domain binds the aglycone and is responsible for substrate specificity, important substrates for UGT2B17 are testosterone and DHT among others. The aglycone binding site is located in the core of the NT domain, together with residues forming the catalytic site. Because of a lack of crystal structures of human enzymes of this domain, the specific residues responsible for aglycone binding is uncertain. The available GT templates binds other aglycones, and consequently have different binding pockets.

The catalytic site is built up by two coordinated residues responsible for initiating the glucuronidation mechanism. UGTs utilize a serine hydrolase like mechanism for catalysis, where residues H35 and D152 functions as an acid base pair. H35 functions as a base, deprotonating the aglycone and increases its nucleophilicity, thereby facilitating a nucleophilic attack from the aglycone on the glucuronic acid moiety of UDPGA. The role of D152 is to stabilize the deprotonated H35 and to ensure its favourable position relative to the aglycone. The result of the catalysis is the transfer of glucuronic acid moiety over to the aglycone.

The NT domain also contains a membrane attached region, proposed to be involved in helping lipophilic substances reach the active site. This membrane interacting region may be the cause of the lack of crystal structures of the NT domain of human UGTs, since crystallizing membrane proteins is an extremely difficult process.^{16,18,20-22}

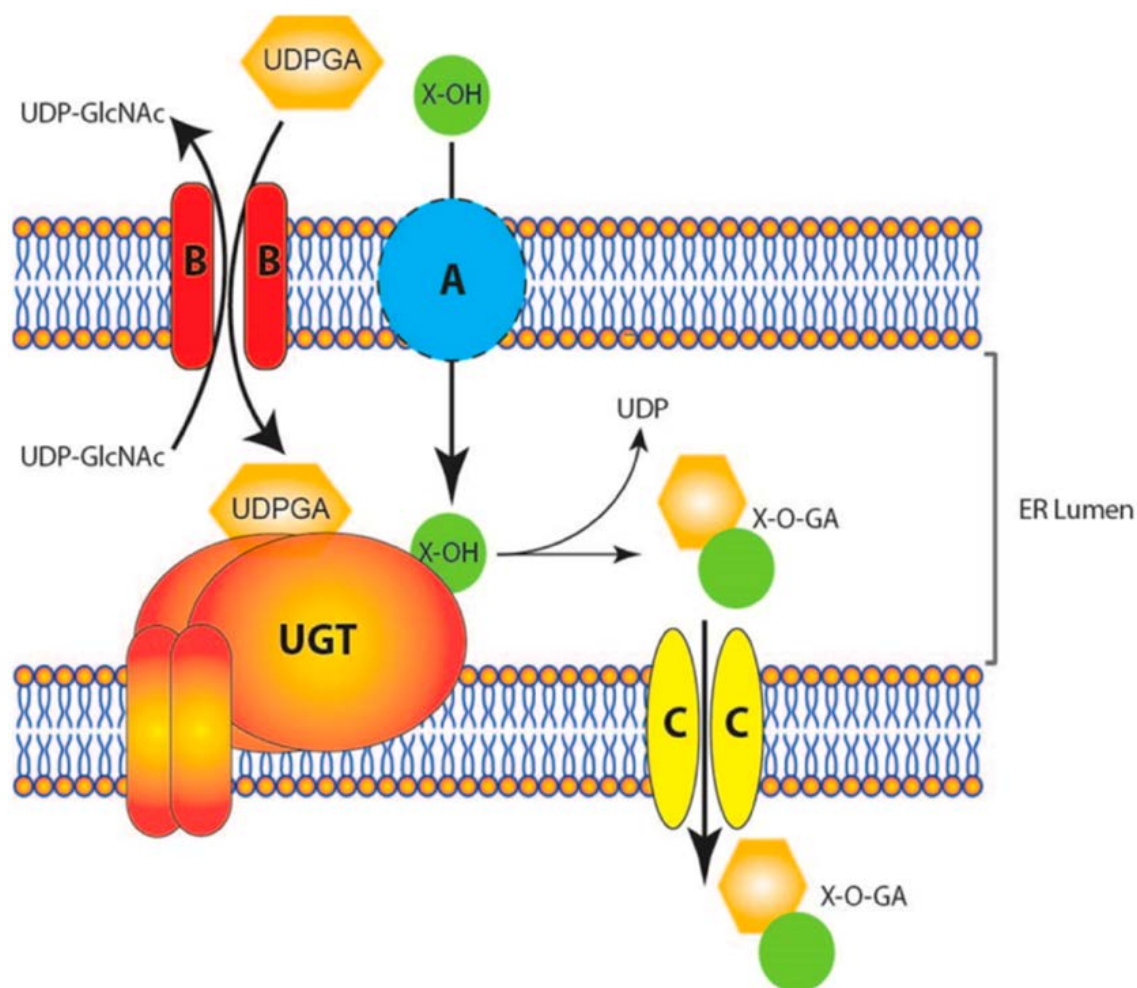


Figure 9 The glucuronidation system in the endoplasmic reticulum lumen. UDPGA is transported into the lumen by an antiporter (B) and aglycones (X-OH) enter by diffusion (A). The aglycones are conjugated by UGT, and the glucuronide products (X-O-GA) are removed from the lumen by transporters (C) (Reprinted from: *Revisiting the Latency of Uridine Diphosphate-Glucuronosyltransferases (UGTs)—How Does the Endoplasmic Reticulum Membrane Influence Their Function?* by Liu & Coughtrie. *Pharmaceutics*. 2017)

1.5 Computational science and drug discovery

Computational science has a major role in modern medicinal chemistry and are important in both drug discovery and drug development. Rapid advances of both software and hardware has meant that many of the operations that once was exclusive for experts with supercomputers now are available on ordinary laboratory computers for a larger group of scientists with little experience in quantum mechanics. Computer based methods in drug discovery allows rapid screening of large compound databases and determination of potential binders through modelling, simulation and visualization techniques.²³

1.5.1 Molecular modelling

Molecular modelling is a collective term for different computational techniques used for analysing, visualizing and manipulating 3D structures of molecular systems, ranging from small chemical systems to large biological macromolecules. The operations carried out in molecular modelling involve the use of programs or algorithms that calculate the structure and property for the molecule of interest. Two computational methods are used to calculate structure and property data, molecular mechanics (MM) and quantum mechanics (QM).

The MM method is based upon calculation of molecular conformational geometries and energies using a combination of empirical force fields. The molecule is treated as a series of spheres (atoms) connected by springs (bonds). Using equations derived from classical mechanics, the total steric energy (E_{tot}) of the molecule is calculated as the sum of energies from bond stretching (E_{bond}), angle bending (E_{angle}), torsion energies (E_{tors}), and non-bonded interactions (E_{vdw} , E_{elec}), as shown in the following equation:

$$E_{tot} = (E_{bond} + E_{angle} + E_{tors}) + (E_{vdw} + E_{elec})$$

These calculations require parameters or data such as ideal bond lengths, angles and torsions etc, which are stored in tables within the software used. All aberrations from ideal values will give the molecule increased energy, which is disadvantageous. MM is fast and less intensive than QM, enabling the use of the method on large molecules. The MM method is suitable for calculating energy minimizations, identifying stable conformations, generating different conformations, energy calculations for specific conformations and studying molecular motion.

The QM method uses quantum physics to calculate the properties of a molecule by considering the interactions between electrons and nucleus of the molecule. The computational calculations are substantial and time consuming, thereby restricting the QM method to smaller molecules. The QM method is suitable for calculating molecular orbital energies and coefficients, partial atomic charges,

transition state geometries and energies, heat of formation for specific conformations, dipole moments, bond dissociation energies and electrostatic potentials.⁹

1.5.2 Homology modelling

The functional properties of a protein are dependent on its 3D structure, which in turn is determined by its amino acid sequence. Information about the 3D structures are decisive for understanding the protein function mechanisms, identification of bindings sites, understanding the origin of dynamics and stability properties, and may also contribute to modern drug design.¹⁶

The experimental techniques for determination of 3D structure of biological macromolecules have significantly progressed recently, with x-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy being the main methods. A vast amount of 3D structures has been experimentally determined and submitted into the Protein Data Bank (PDB), but because it is difficult and time-consuming, there are still many proteins with unknown 3D structure. Consequently, it is of major interest to use *in silico* approaches such as homology modelling to study some of these and direct further experimental work.²⁴

Homology modelling is based on the observation that proteins with similar amino acid sequences will have similar 3D structures. The method is used to predict an atomic resolution model of a target protein from its amino acid sequence, based on a template. The template is a known 3D structure of a related homologous protein, determined experimentally by x-ray crystallography or NMR spectroscopy. Homologous proteins have evolved from a common ancestor, and within these the structural conformation is better conserved than the amino acid sequence. Thus, proteins sharing a significant sequence similarity can be expected to share common structural properties, particularly the overall protein fold. Using an experimentally determined 3D structure from a similar protein as a template, a homologous model can be predicted.^{24,25}

Homology models are less reliable than an experimentally obtained structure, however the model is often sufficient for use in structure-based drug design strategies. Another advantage is that a homology model can be developed within a very short time frame compared to experimental structure determination. Homology modelling is a multi-steps process, summarized in the following way; (1) template identification, (2) sequence alignment, (3) model building, (4) model refinement and (5) model validation, as seen in figure 10.

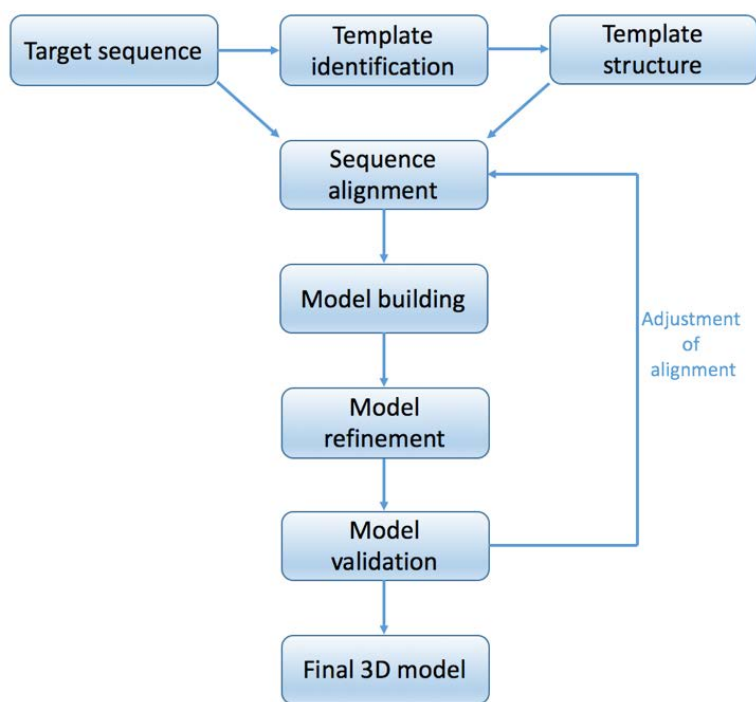


Figure 10 Steps in homology modelling

The first step of homology modelling is finding and choosing the most suitable crystal structure to be used as a template. The template will act as a pattern to build a new model of the target protein of interest. Using the target sequence as a query for Basic Local Alignment Search Tool (BLAST), suitable templates with an acceptable sequence identity can be identified. Templates can be retrieved from the Protein Data Bank (PDB). The quality of the model is directly linked to the template selected, and sequence identity between target and template. Sequence identity over 50% give highly accurate models that can be used for drug discovery experiments, sequence identity between 30 and 50% can contribute to mutagenesis experiments, the quality of a model sharply decreases below 30%. Another aspect contributing to model quality is the resolution of the protein used as a template. A high resolution close to 1Å indicates good quality of the data collected during the crystallization process, and that it is easy to see every atom in the electron density map. Structures with resolution below 3Å are considered reliable. In general, the most suitable templates have a high sequence identity and a high resolution. ²⁴

The next step of the modelling procedure is constructing an optimal target-template sequence alignment. The aim is to create correspondence between amino acid residues of target and template by superimposing the two structures. Unless target and template are closely related, there will be regions of considerable structural difference between the two. These structurally dissimilar regions are most often a consequence of insertions, deletions, or extensive changes in the amino acid sequence. Assignment of residue correspondence in such regions can be difficult and also meaningless. An

accurate alignment should include the structurally and evolutionary residue pairs, and also leaving out structurally different regions. It can be useful to align the target with multiple templates to improve the alignment. This may be an option when an accurate alignment cannot be achieved with the same template, instead different regions in the target sequence is aligned with different templates. This gives the opportunity for model improvement but also introduces additional complexity into the modelling procedure. The target-template alignment procedure can be divided into 3 tasks: (1) generating initial sequence-structure alignment, (2) finding alignment regions needing adjustment, and (3) improving the alignment.^{24,26}

Model building involves the construction of a 3D structure of the target protein based on the target-template alignment. The model building procedure is built up in three main steps: (1) core modelling, where the backbone is constructed (2) loop modelling based on structures in the protein database, and (3) optimization of sidechains and backbone.²⁴ In this study, the ICM software with its homology macro module was used for all steps in model building.

The model refinement is an important process where structural errors in the newly made model are eliminated, this will increase the quality and optimize the energy of the model. The most uncertain parts of the model are refined first, the process being dependent of the quality of the model made. Energy functions are used to enforce the correct covalent geometry, avoid steric clashes between residues, and atomic overlap. This is done using energy minimizations, Monte Carlo simulations, or molecular dynamics calculations. The refinement process will construct a structure with as low free energy as possible, this is done on the basis that the native structure of a protein is uniquely determined by its amino acid sequence and the conformation with the lowest free energy.^{27,28}

Model validation is done to ensure the quality and reliability of the built model. Bond angles, bond length and torsion angles are checked to make sure they are within the accepted normal ranges, and the correctness of residue chirality has to be proved. Validation of the model can be done by online tools such as Structural Analysis and Verification Server (SAVES), by site-directed mutagenesis studies, or by docking known binders and non-binding molecules (decoys) to target protein. The energetic stability of the model can also be assessed by running molecular dynamics simulations.²⁴

1.5.3 Docking and scoring

Docking is a process in molecular modelling which predicts the preferred orientation and conformation of a ligand within a target binding site of a protein. Accurate structural modelling and correct prediction of activity and binding affinity are the aims of docking studies. The process of docking relies on computer sampling algorithms to generate ligand binding modes by placing the ligand within the binding site, as shown in figure 11. These algorithms are complemented by scoring functions that predicts binding affinity through the evaluation of interactions between compounds and target, ranking the results.

The aim of the scoring function is to identify the most reliable binding pose, and to distinguish true ligands from decoys. These results are not exact measures of affinity, but rather an estimation. Three types of scoring functions are used in general: (1) force field-based, (2) knowledge-based and (3) empirical. Force-field based scoring quantify the sum of two energies, the ligand-target interaction energy and the internal ligand energy. These are derived from electrostatic interactions, van der Waals interactions, bond stretching, angle bending, and torsional forces. Knowledge based scoring function use statistical energy potentials of ligand-target complexes, derived from experimentally determined structures. Empirical scoring function calculate the binding affinity based on a set of weighted energy terms, such as electrostatic interactions, van der Waals interactions, hydrogen binding, hydrophobicity, entropy and desolvation. In addition, there is a fourth scoring function called consensus score, which combines the three main functions in order to balance errors, adjust any imperfections, and improve the probability of identifying true ligands. Generally, a score below -32 is regarded as a good docking score in the ICM software, but it is dependent of the system docked into.^{25,29}

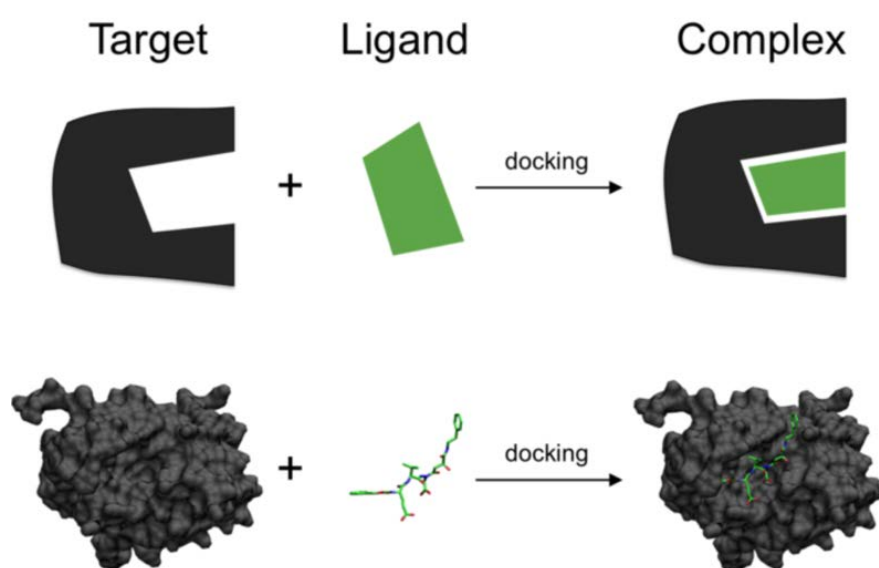


Figure 11 Schematic illustration of ligand docking to protein target, forming a protein-ligand complex (Retrieved from [wikipedia.org/wiki/docking_\(molecular\)](https://wikipedia.org/wiki/docking_(molecular)), Public Domain)

There are three different approaches for docking procedures, (1) rigid, (2) flexible, and (3) semi-flexible. Rigid docking is the simplest approach, it treats both ligand and receptor as rigid bodies. This is an acceptable option if the active conformation of the ligand is known. Flexible docking is the most complex approach, where both ligand and receptor are considered flexible. This is the ideal approach, since it reflects the natural structural flexibility of proteins and ligands. Unfortunately, this option is very challenging and extremely expensive in terms on computational time, limiting its use for induced-fit docking. The most common approach is semi-flexible docking, which treats the receptor as rigid, but the ligand as fully flexible, allowing it to adopt different conformations. This option is a trade-off between computational time and accuracy.^{9,25}

1.5.4 Receiver operating characteristic (ROC) curves

The ROC-curves are graphical plots used in statistics to illustrate true positive rate (TPR) against false positive rate (FPR) for different possible cutpoints in a diagnostic test. The TPR is the sensitivity of the test, while the FPR is the fallout of the test, the ROC curve is thus the sensitivity as a function of fallout. ROC curves can be used to evaluate the overall predictability of homology models, thereby indicating which of the built models are best suited for further work with virtual ligand screening. This is done by docking known binders and decoys to a drug target, and scoring their binding affinity. Decoys are molecules with similar physiochemical properties as active compounds, but with different chemical structures, assumed to be non-binders. The ROC curve is created by plotting the TPR against the FPR, giving a graphical representation from which the area under the curve (AUC) can be calculated as a measure of accuracy of the models, as shown in figure 12.

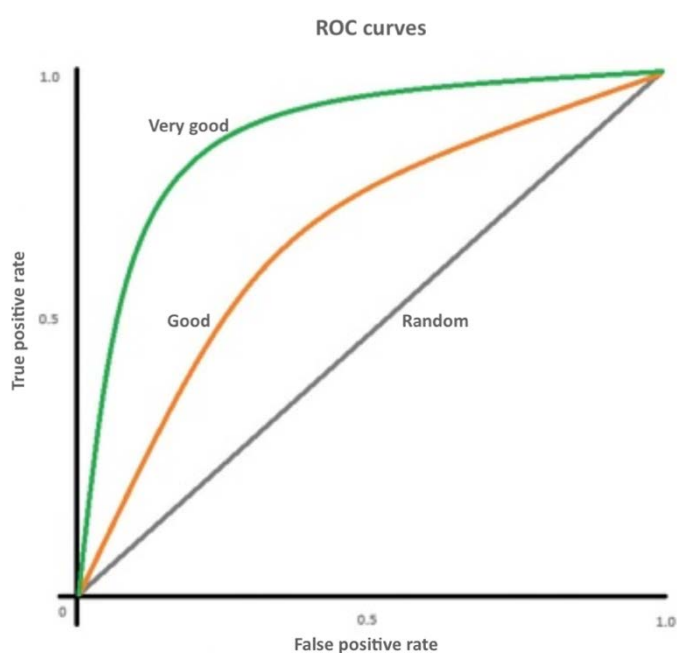


Figure 12 Graphical representation of a ROC curve.

The AUC summarizes the entire location of the ROC curve rather than depending on a specific point. When using ROC curves to evaluate homology models, the outcomes are labelled as the following: (1) True positive (TP) is true ligand binding classified as positive, (2) False negative (FN) is true ligand binding classified as negative, (3) True negative (TN) is decoys classified as negative, and (4) False positive (FP) is decoys classified as positives.

$$\text{True positive rate (TPR)} = \frac{TP}{TP + FN}$$

$$\text{False positive rate (FPR)} = \frac{FP}{TN + FP}$$

A diagonal ROC curve represents a model which is a random classifier that is no better than chance, and not able to discriminate between TP and FP. The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the model. An accurate model is able to classify true ligand binders as TP, and decoys as FP. A calculated AUC value of 0.9-1.0 represents an excellent classifier, and an AUC value of 0.5 represents a random worthless classifier.³⁰

1.5.5 Virtual ligand screening

Virtual ligand screening (VLS) is a computer based method in drug discovery used to search huge compound databases containing millions of molecules for active ligands, and predicting their binding affinity to a target receptor. Experimental limitations such as solubility and aggregate formation do not need to be considered, but an important prerequisite is knowledge of the spatial and energetic criteria responsible for ligand binding. A 3D structure of target, or a rigid reference ligand with known active conformation in the putative binding site must be available. Since the testing is done using computer programs, the compounds doesn't consume valuable substance material and hence costs are less than for regular high-throughput screening methods.³¹

The methods used in VLS are broadly classified as either structure-based drug design (SBDD) or ligand-based drug design (LBDD) strategies. LBDD uses information about the ligand for predicting activity, depending on the ligands similarity or dissimilarity to previous known active or inactive ligands. This helps deducing the properties of the complementary binding site as the 3D structure of the target is unknown. SBDD is a strategy based on knowledge the detailed 3D structure of the drug target, including the binding site. The properties of the target structure and binding site are used to identify possible drug candidates through docking protocols. The choice of method depends on the amount and quality of data available, if both the ligand and the structure of target is known, a combination of the two strategies can be used.²³

2. Aim of the study

The aim of this study was to improve the understanding of the binding properties of UGT2B17, thereby making it possible to develop selective inhibitors of the enzyme. Inhibitors of UGT2B17 could help maintain normal testosterone levels in patients with declining levels caused by various factors.

The 3D crystal structure of UGT2B17 was not experimentally determined at the time of this study. Consequently, a homology modelling procedure was used to generate models of the UGT2B17 enzyme based on templates with known crystal structure. Molecular docking of inhibitors on the models was performed to gain further insights in the interactions between ligand and binding site, and to determine which of the models had the best accuracy. The best model was selected for further studies, using virtual ligand screening to find novel drug candidates.

3. Methods

3.1 Software and databases

3.1.1 Molsoft Internal Coordinates Mechanics (Version 3.8.7)

Molsoft molecular modelling technology is based on the coordinate method and optimization procedures implemented in the software. The use of Internal Coordinate Mechanics (ICM) gives a general modelling and structure prediction framework for many different tasks of structural biology and rational drug design. The ICM method has been extensively validated in bioinformatics and drug discovery projects^{32,33} In this thesis, the ICM software was used to build homology models of the enzymes, docking of ligands and decoys, and for virtual ligand screening. The Molsoft ICM software is available at <http://www.molsoft.com>

3.1.2 The Protein Data Bank

The Protein Data Bank (PDB) is an archive of information about experimentally determined 3D structures of biological macromolecules, such as proteins and nucleic acids found in all organisms. Structures are solved through X-ray crystallography, electron microscope and NMR spectroscopy, with the first being most common. At the moment there are 140000 structures deposited in the database, with the number increasing continuously.³⁴ The PDB database provided the protein crystal structures used as templates for homology modelling. The database is available at <https://www.rcsb.org>

3.1.3 Universal Protein Resource Knowledgebase

The Universal Protein Resource Knowledgebase (UniProtKB) is a comprehensive resource for protein sequence and functional information with detailed annotations. The database consists of two sections, Swiss-Prot and TrEMBL. Swiss-Prot contains manually annotated records with information extracted from literature and evaluated computational analysis, reviewed by curators. TrEMBL contains computationally analysed records that await manual annotation and reviewing. All sequences in the knowledgebase are given a unique accession number.³⁵ The database was used to find amino acid sequences for target and template proteins. The database is available at <http://www.uniprot.org>

3.1.4 Basic Logical Alignment Search Tool

The Basic Logical Alignment Search Tool (Blast) is a search tool that finds regions of similarity between biological sequences, from the National Center for Biotechnology Information sequence database. The program compares nucleotide or protein sequences to sequence databases and through calculations finds statistical significance of matches. Blast can be used to investigate functional and evolutionary relationships between sequences as well as identify members of gene families. Different algorithms are available for a standard protein Blast. Protein-protein Blast compares a protein query to a protein database. Delta-Blast constructs a position-specific scoring matrix using the results of a conserved domain database search and searches a sequence database. Psi-Blast allows the user to build a position-specific scoring matrix using the results of the first run.³⁶ In this thesis, the search tool was used to find potential templates with sequence homology of known 3D structures. The search tool is available at <https://blast.ncbi.nlm.nih.gov/Blast.cgi>

3.1.5 Structural Analysis and Verification Server v5.0

The Structural Analysis and Verification Server (SAVES) metaserver is a part of web services provided by the Molecular Biology Institute at the University of California, Los Angeles. The metaserver have several different programs used to analyse and validate protein structures before and after model refinement. Over 4000 verification jobs are run every day on the server. The metaserver is available at <https://servicesn.mbi.ucla.edu/SAVES/>

3.1.6 PubChem

PubChem is a public domain database containing information about chemical molecules and their activities against biological assays. The database is maintained by the National Center for Biotechnology Information, and consists of three interlinked sections: (1) Compounds (2) Substances, and (3) Bioassays. At this time the database contains over 94 million compounds, 242 million substance descriptions and over 1.25 million bioassays.³⁷ PubChem was used to search for active and inactive compounds to the target protein. The database is available at the following webpage <https://pubchem.ncbi.nlm.nih.gov>

3.1.7 ChEMBL

ChEMBL is a manually curated chemical database of bioactive compounds with drug-like properties. The database is maintained by the European Bioinformatics Institute of the European Molecular Biology Laboratory, and contains information about binding, functional properties and ADME Toxicity for a vast number of compounds. There are at the moment 14 million bioactivity measurements for over 2 million compounds and 11000 protein targets in the database.³⁸ In this study the database was used to search for compounds, targets and assays. The database is available at <https://www.ebi.ac.uk/chembl/>

3.1.8 DecoyFinder 2.0

DecoyFinder is a graphical tool designed to help molecular docking programs by providing challenging decoys for a given group of active ligands. The DecoyFinder software finds molecules which have similar number of rotational bonds, HBA, HBD, logP value, and molecular weight, but are chemically different from the active ligands used as input. The software acquires the decoys directly from the ZINC compounds database.³⁹ In this study, DecoyFinder was used to retrieve decoys with similar physiochemical properties assumed to be inactive for the UGT2B17 enzyme. 15 decoys were found for each ligand, giving a total of 145 decoys in the output chemical table after duplicates had been deleted. The DecoyFinder software is freely available at <http://urvnutrigenomica-ctns.github.io/DecoyFinder/>

3.1.9 eMolecules

eMolecules is a public domain database of diverse chemical building blocks, screening compounds and antibodies. The database is owned by a private company and has its headquarter in San Diego. Currently there are over 1.5 million building blocks, 7 million screening compounds and over 600000 antibodies available at the database. The search engine allows substructure, similarity or exact searches when searching for chemicals. When performing a sequence similarity search it's possible to enter the desired percentage of similarity. For this study, eMolecules was used to find compounds for VLS. The database is available at <https://www.emolecules.com>

3.2 Homology modelling

The 3D structure of the androgen metabolising enzyme UGT2B17 was not experimentally determined at the time of this study. Consequently, the ICM software with its homology modelling module was used to generate models of the enzyme.

3.2.1 Template identification

The amino acid sequence of human UGT2B17 was retrieved from the UniProtKB database, accession number O75795. This amino acid sequence will act as the target of this study, and will be used to find homologous proteins with known 3D structures that can be used as templates.

Close homologues were found using the Blast search tool for a sequence similarity search with the target sequence as query. A standard protein-protein Blast was performed on the 530 residues of target, resulting in a list of potential templates with available crystal structures. Most of the potential templates were GTs with a sequence identity of about 20%, but one partial structure of an UGT had very good homology and a high sequence identity. The partial structure (PDB id: 2O6L) consisted of the CT domain of the closely related enzyme human UDP-glucuronosyltransferase 2B7 (UGT2B7), and had a sequence identity of 82% with the query sequence. The CT domain included most of the residues that make up the binding site of the UDPGA cofactor. Consequently, this made the crystal structure useful as a template for a model of the CT domain of target, but also as a part of a multi template model where two templates are combined in the modelling process. The crystal structure of UGT2B7 had formed as a dimer, with chains designated A and B. Since chain B lacked some residues located close to the binding site, chain A was chosen for modelling.²⁰

Because of the good homology and sequence identity of UGT2B7 with target, combined with the low sequence identity in most of the other potential templates, a multiple template modelling procedure was the best option for an acceptable full length model of both domains. The rationale for building a model with both domains was based on studies indicating interactions between the co-factor UDPGA and residues in the NT-domain. In addition, a partial model CT domain of the enzyme was built based on UGT2B7 alone. This model could lack some residues of importance to UDPGA binding, but would have higher sequence identity.^{18,20}

To model both domains of the target protein there was a need for templates with acceptable sequence identity in the N-terminal, in addition to the partial structure of UGT2B7. A delta-Blast algorithm of the first 284 residues from the NT domain of target was performed, resulting in a long list of potential templates. GTs were marked for next iteration, and followed by a psi-Blast algorithm. This resulted in a new list of 48 potential templates. The templates obtained from the Blast search tool were shortlisted and investigated further based on: (1) conserved UDPGA binding site, (2) sequence

identity, (3) query cover, (4) resolution, and (5) expectation-value. All the chosen templates considered had most of the UDPGA binding site conserved, but ultimately this region which was the focus of this study would be modelled from the UGT2B7 template. The sequence identity for these templates were lot lower than of UGT2B7, but when combined together with the CT domain the sequence identity will be acceptable. The query cover of all templates were all over 80%, except from the partial structure with 31% query cover. The resolution of the chosen templates ranged from 1.7Å-2.59Å, which is considered reliable. The expectation value (e-value) is a parameter describing the number of different alignments expected to occur by chance in a database search, the lower the e-value, the better the alignment. Proteins with an e-value above 0.0001 was excluded.

Based on the criteria the following crystal structures were chosen as templates for homology modelling, PDB id: 3WAD, 4AMG, 4M83 and 2O6L, as shown in table 2.

Table 2 Templates chosen for homology modelling

PDB ID	Name	Sequence identity	Resolution	Deposition author
3WAD	Glycosyltransferase VinC	20%	2.00Å	Nango.E et al
4AMG	Glycosyltransferase SnogD	22%	2.59Å	Claesson.M et al ⁴⁰
4M83	Glycosyltransferase OleD	21%	1.70Å	Wang.F et al
2O6L	UDP-glucuronosyltransferase 2B7	82%	1.80Å	Miley.MJ et al ²⁰

3.2.2 Sequence Alignment

The templates selected for homology modelling were aligned with the sequence of UGT2B17 using the alignment tool in the ICM software. The sequence of 2O6L needed no adjustment because of the high sequence identity. The other templates had relatively low homology with target, and needed manual adjustment. By using a multiple sequence alignment of the templates combined with several other human UGTs, a basis for further adjustment was built, as shown in figure 13.

Some site-directed mutagenesis studies of human UGTs were available, giving insights to residues of importance. Many of these residues and the secondary structures associated with them of were conserved or semi-conserved, thereby helping in the alignment process. Residues H35 and D152 act as a catalytic dyad in the catalytic reaction initiating the glucuronidation mechanism of the enzyme. Residues R49 and H51 have a role in function and structural integrity required for optimal catalytic

activity, but are not directly involved in substrate binding. The residue F90 forms aromatic ring stacking interactions with phenolic substrates. The residue S121 is required for the ability to conjugate C19 steroids at the 3 α -OH position, thereby being involved in steroid specificity. The residues S309 and R339, in addition to many residues in the region 357-400 are involved in UDPGA binding, and forms the binding pocket.^{22,41-44}

Any gaps in the alignments were shifted to the loop regions where this was possible. The adjusted sequence alignments shown in figures 14, 15, 16 and 17 were used to build the models of UGT2B17.

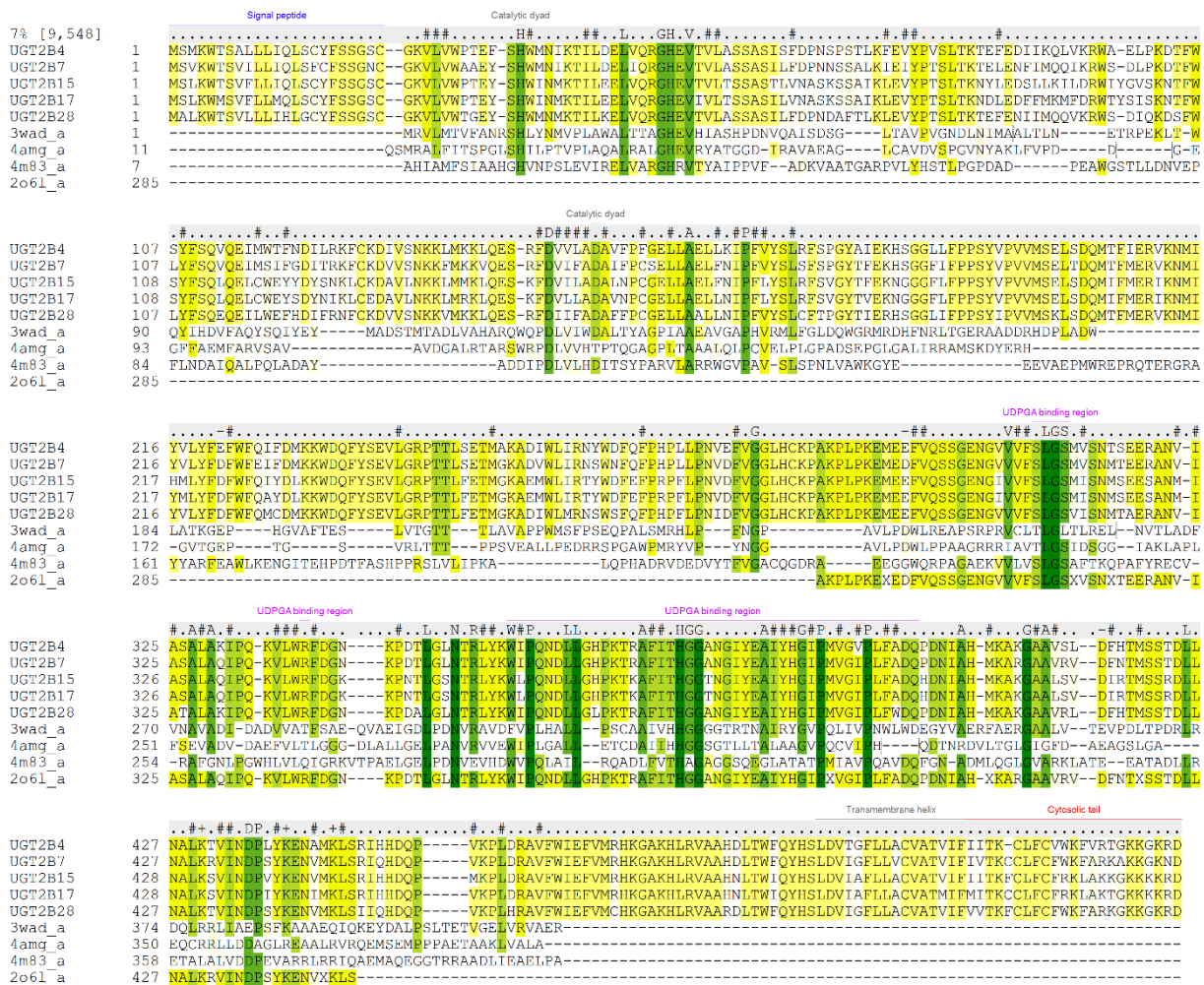


Figure 13 Multiple sequence alignment of several homologous UGTs and the chosen templates. Areas with dark green colour indicates fully conserved residues, yellow colour indicates semi conserved.

20% pP=3.9

```

.....#...###...SH##NM...##..L...GHEV###...
UGT2B17 1 MSLKWMSVFLMLQLSCYFSSGSCGKVLVWPTEYSHWIKMKTILEEVQRGHEVIVLTSSA
3wad_a 1 -----MRVLMTVFANRSHLYNMVPLAWAITTAGHEVHIASHPD

.##.....L.##P.....N.###.#.....#.K.T....#...-###.Y
UGT2B17 61 SILVNASKSSAIKLEVYETSLTKDLEDFMCMFDRWTYSISKNFWSYFSQLQELCWEY
3wad_a 39 NVQAISDSG---LTAVGVGNLNIAMALTLN-----ETRPEKLE---WQYIHDVFAQY

S.###...D.###L#...#...D####DA###G.##AE##.P##..L.....#
UGT2B17 121 SDYNIKLCEAVLNKKMRKLQESKFDVLLADAVNPCELLAEFLNIFLYSRFSVGYT
3wad_a 100 SQIYEYM-ASTMTADVAHARQWQPLVLIWDALTYAGPIAAAVGAHVRML-----FG

#-.G.....M.-#...#...ER#...#...###D.W#.....F...#
UGT2B17 181 VEKNGGFLFPSPSYVPVWSELSDQMIEMERIKNMIYMLYFENFQAYDLKKWDQFYSEV
3wad_a 154 LDQWGR-----VRDHFNRL-TGERAADRHDPLA---LATKGEPHGVAFTESL

#...TTL#...#...#...#...#...E.P.###.###F.G.....P.#.-###...
UGT2B17 241 LGRPTTLFETMGKAEMWLIRTYWDFEFERPFPLPNVDFVGLHCKPAKPLKEMEEFVQSS
3wad_a 200 VTGTTTAVAPPWM-----SFPSQFALSMRHLPEFC-----EAVLPDWLREA

.....V##.LG...#...#...A.##..AA.I...V##.F...#...#G...N.R##.#
UGT2B17 301 GENGIIVFSLGSMISNMSEESNMIASALAQIPQKLVWRDGGKKNPTL---SSTRLYKW
3wad_a 243 PSRPRVCLTLGLTLRELNVTLADFV-NAVADIDADVATSAEQVAEIDLPDNRVAVDF

#P...LL..P...A##.HGG.....AI.#G#P.#.#P.##...#A..#+GAAL..
UGT2B17 358 LQNDLGHFKTKAFITHGCTNGIYEAIYHGI PMVGIPLFADQHDNIE-HMKAKGAALSV
3wad_a 305 VPLHALL--SCANIVHGGGGTRTNAIRYGVVQLIVENWLWDEGYVVERFAERGAALVT

206L
-#...#...L...L+.#I.-P.#K...#...+.#...P.....V+##-R.....
UGT2B17 417 DIRTMSSRDLLNAKSVNDIYENIMKLSRIHHDQF-----KPLDRAVFWIEF
3wad_a 363 EVPDLTPDRLRDQRRRLIAEESFAAAEQIQKEYDALPSLTETVGELEVAER-----

.....
UGT2B17 468 VMRHKGAKHLRVAAHNLTWIIQYHSLDVIAFLLACVATMIFMITKCCFLCFRKLAKTGKKK
3wad_a -----

...
UGT2B17 528 KRD
3wad_a ---

```

Figure 14 Sequence alignment for UGT2B17 and 3WAD used for homology modelling. Areas with green colour indicates conserved residues. Red annotation marks region for multiple templates.

22% pP=3.3

```

.....S#...##...#SH##...##..L...GHEV.##T...
UGT2B17 1 MSLKWMSVFLMLQLSCYFSSGCGKVLVWPTEYSHWINMKTILEEIVQRGHEVIVLSSA
4amg_a 11 -----QSMRALFITSPGLSHILPTVPLAQAARALGHEVRYATGGD

..#V.....A#.#.#...Y#..#...D.E.FF#.MF.R#.....
UGT2B17 61 SILVNASKSSAIKLEV---SPTSLTKNDLEDFEMKMDRWTYSISKNTFWSYFSQLQELC
4amg_a 51 IRAVAEAGLCVVDVSPGVNAKLFVDPDGGGFFAEMFARVS-----

.....AV#...#R.#...+D#####.#...#G.L.A.#L.#P#...#.#
UGT2B17 118 WEYSYDNIKLCEDAVLNKKLMKQESKFDVLLADAVNPCCELELNIIFLYSLRFSV
4amg_a 104 -----AVAVDGALEARTASWRPDLVVHTPTQGAGPITAAALQLICV---ELPL

G#.....G.G#L#.....##-.-.+.....
UGT2B17 178 GYTVKNGGCFIFPPSYVPMSELSQMI FMERIKNMIYMLYFDFWFQAYDLKKWDQFY
4amg_a 148 GPADSEPLCALIRR-----AMSKDYERHG-----

.....V.G.PT...L..T#...#E##L#.....#W.#.#.....V.#.GG##.....
UGT2B17 238 SEVLCRPT---TIFEIMGKAMWIRT----YWDFFPRPFLPNDFVGLHCKPAKPL
4amg_a 173 --VTCEPTGSVRITTPPSVALLPEDRRSPGAWPMRY-----VPYNGAV-----

.....#.-##.....###.LGS#.S.....A.##.S.#A.#...###.#.G..#..LG
UGT2B17 290 PKEMEEFVQSSGENGIVVFSLGSIMINMSEESANMIASALRQIPQKVLWRFDEKKPNTLG
4amg_a 217 ---LPDWLPPAAGRRIAVTLGSIDSGGIAKLAPLF-SEVADVDAEFVLTGLGDLALLG

.....N.R##.W#P...LL.....A#I.HGG...##.A###G#P.#.IP#...Q..N..#
UGT2B17 350 ---SNTFLYKILQNDLGHPKTKFITHGGTNGIYEAIYHGIEMVGLFLFADQHDVIAH
4amg_a 275 ELPANVRVVENIELGALL--ETCDAILHCGSGTLLTALAAGVQCVELH---QDTNRDV

.....G#.#.#-#.#.....#+.##.D#.#+E.##+#...#.-.P..P#-.A#.#.#.
UGT2B17 407 MKAKGAALSVDIRTMSRDLLNALKSVINDPIYKENIMKLSRIHHDQVVKLDRAVFWIE
4amg_a 333 LTGLCIGFDAEAGSLGAEQ----CRLLDDAGLRFAALRVRQEMSEM-PAETAAKLVA

##.....
UGT2B17 467 FVMRHKGAKHLRVAAHNLTWIQYHSLDVIAFLLACVATMIFMITKCCLECFRKLAKTGKK
4amg_a 388 LA-----

....
UGT2B17 527 KKRD
4amg_a ----

```

Figure 15 Sequence alignment for UGT2B17 and 4AMG used for homology modelling. Areas with green colour indicates conserved residues. Red annotation marks region for multiple templates.

21% pP=2.3

```

.....#####.H#.....#.ELV.RGH.V.##....
UGT2B17 1 MSLKWSVFLMQLSYFSSGSCGKVLVWP-TEYSWINMKTILEELVQRGHEVIVLTSSA
4m83_a 1 -----AHIAMFSAIAAHGIVNPSLEVIRELVARGERTTYAIPP-

.###.....#+#.Y#..L...D#-.....#-..W...#.N...#.F...#Q.L#..
UGT2B17 61 SILVNASKSSAIKLEVPTSITKNLEDDFFMKMFDRITYSISKTFWSYF--SQLDELQWE
4m83_a 38 -VFADKVAATGARPVLYHSTIPGPDAD-----PEAVGSTLLDN--VEPELNDALQALPQL

#.Y.....-...D##L#D#...##..#LA...#P...#SL...#.#.#.
UGT2B17 120 YSDINIKLCEDAVLNKKLMRKLQESKFLVLAAVNPGCELLAELLNIFLYSLRFSVGYT
4m83_a 90 ADA-----ADDIPVLVHITSYPARVLAARRWGVF--AVSLSPNLVAW

..K.....#.P...#...E.....+...#Y##.F-#W#...#.#...#.#.#
UGT2B17 181 VEINGGGFLFPISYVPVMSLSDQMIFMERIKNMIIMLYDFDFQAYDLKKWDQFYSEVL
4m83_a 132 --GYEEVAEPMWREPRQTE-----RGRAYARAEALNKENGITEHPDTFASHP

.R...L.....I...###.R#...#.FVG...#.A...E...#...G.
UGT2B17 242 GPTTFETMGKAEMWLRITYWDFEFPFPLPNVDFVGLHCKPKPLPKMEEFVQSSGE
4m83_a 180 PSLV-----IPKALQPHADVDEVDVYTFVACQGR-----EGGWQRPAGA

..#V##SLGS##.....#...#A#...#P...#L#...G+K...P...LG...N...#W#
UGT2B17 303 NGIVFSLGSMISNMSEESANMIASALAQIEQKVLWRFDKK--INTLGS---NTRLYKWL
4m83_a 225 EKVLVSLGSFTK-QPAFYRECVRAFGNLGWHVVLQIGRVTAEELGELPDNVEVHDV

PQ...#L.....#F#TH.G..G..E...#PM#.#P...#DQ#.#N#...#G#A...#...
UGT2B17 359 PNDLGHPKTKAFITHGNTNGIYEAIYHGIEMVGIHLFADQHDIAHMKAKAALSVDIR
4m83_a 285 PQLAIE--RQADLEVTAGAGGSQGLATATPMIAVQAVDQFGADMLQGLVARKLATE

206L
.#...L...#...#DP...E...+L.RI.#-.#.....RA#.#IE#.#.....
UGT2B17 420 TMSRDNLNALKSVINDPIYKENIMKLSRIHHDQPVK-PLDRVFWLDFVM--RHKGAKHL
4m83_a 344 EATADLRETALALVDP---EVARRLRRIQAEMAQEGGTRRAADLLEALPA-----

.....
UGT2B17 478 RVAAHNLTWIQYHSLDVIAFLLACVATMIFMITKCLFCFRKLAKTGKKKKRD
4m83_a -----

```

Figure 16 Sequence alignment for UGT2B17 and 4M83 used for homology modelling. Areas with green colour indicates conserved residues. Red annotation marks region for multiple templates.

```

82% pP=35.9
UGT2B17 1 MSLKWMSVFLLMQLSCYFSSGSCGKVLVWPTEYSHWINMKTILEELVQRGHEVIVLTSSA
2o6l_a 285 -----

UGT2B17 61 SILVNASKSSAIKLEVYPTSLTKNDLEDFMCKMFDRTYSISKNTFWSYFSQLQELCWEY
2o6l_a 285 -----

UGT2B17 121 SDYNIKLCEDAVLNKKLMRKLQESKFDVLLADAVNPGCELLAE LLNIPFLYSLRFSVGYT
2o6l_a 285 -----

UGT2B17 181 VEKNGGGFLFPPSYVPVVMSELSDQMI FMERIKNMIYMLYFDWFQAYDLKKWDQFYSEV
2o6l_a 285 -----

UGT2B17 241 LGRPTTLFETMGKAEMWLIRTYWDFEFPRPFLPNVDFVGG LHC K P AKPLPKE.E-FVQSS
2o6l_a 285 ----- AKPLPKEMEEFVQSS
AKPLPKEXEDFVQSS

                UDPGA binding region                UDPGA binding region

UGT2B17 301 GENG#VVFSLGS.#SN..EE.AN#IASALAQIPQKVLWRFDG.KP.TLG.NTRLYKW#PQ
2o6l_a 300 GENGIVV FSLGSMISNMS EESANMIASALAQIPQKVLWRFDGK KENTLGSNTRLYKWL PQ
GENGVV FSLGSXVSNXTEERANVTASALAQIPQKVLWRFDGNKFDTLGLNTRLYKWI PQ

                UDPGA binding region

UGT2B17 361 NDLLGHPKT+AFITHGG.NGIYEAIYHGIP.VGIPLFADQ#DNIAH.KA+GAA#.VD#.T
2o6l_a 360 NDLLGHPKT KAFITHGGTNGIYEAIYHGIPVGIPLFADQHDNIAHMKAKGAALSVDIRT
NDLLGHPKTRAFITHGGANGIYEAIYHGIPXVGIPLFADQPDNIAFXKARGA AVRVD FNT

                .SS.DLLNALK.VINDP.YKEN#.KLS.....

UGT2B17 421 MSSR DLLNALKSVINDPIYKENIMKLSRIHHDQPVKPLDRAVFWIEFVMRHKGAKHLRVA
2o6l_a 420 X SST DLLNALRRVINDPSYKENVXKLS-----

UGT2B17 481 AHNLTWIQYHSLDVIAFLLACVATMIFMITKCLFCFRKLAKTGKKKKRD
2o6l_a -----

```

Figure 17 Sequence alignment for UGT2B17 and 2O6L used for homology modelling. Areas with green colour indicates conserved residues. Red annotation indicates the UDPGA binding region.

3.2.3 Model building

One partial model of the CT domain, and three full length models of both domains of UGT2B17 was built based on four different crystal structures. All structures belong to the GT family and were the most suitable candidates with regards to the criteria for template identification.

- Bacterial glycosyltransferase VinC in complex with magnesium ion
PDB id: 3WAD, resolution 2.0Å, chain A, length 398 residues
- Bacterial glycosyltransferase SnogD
PDB id: 4AMG, resolution 2.59Å, chain A, length 362 residues
- Bacterial glycosyltransferase OleD in complex with Erythromycin A and UDP
PDB id: 4M83, resolution 1.7Å, chain A, length 393 residues
- Human UDP glucuronosyltransferase 2B7
PDB id: 2O6L, resolution 1.8Å, chain A, length 162 residues

The partial model and three complete initial models were made using the Homology macro in ICM. Afterwards the Multi-Template Model Editor macro was used to improve the quality of the three GT based models by adding 2O6L as a second template for their CT domain. Since the target sequence contained a signal peptide, a transmembrane region and a cytosolic tail not present in the GT templates, the excessive carboxy and amino terminus tails generated by the modelling procedure were trimmed of the models to avoid them interacting with the secondary structures.

3.2.4 Model refinement

The Refine Model macro of the ICM software was used to refine the built models, a full refinement and optimization of backbone, sidechains and loops were carried out. This refinement macro included (1) Monte Carlo fast simulations for sampling of the conformational space of side chains, (2) repeated annealing of the backbone with tethers, and (3) a second run of Monte Carlo fast simulations on the side chains. Each repetition of Monte-Carlo fast samples the conformational space of the molecule with the ICM global optimization procedure, which consists of a random move followed by a local energy minimization, and then a complete energy calculation. Based on the energy and temperature, the repetition is either accepted or rejected.²⁸

3.2.5 Model validation

Since homology modelling has many aspects of uncertainty, the SAVES metaserver was used to analyse and validate the built models. Of the different programs available in the metaserver, ProCheck and WhatCheck was chosen for the validation. ProCheck investigates the stereo chemical quality of a protein structure by analysing the overall and residue-by-residue geometry, the result of analysis is

represented by a Ramachandran plot. WhatCheck did extensive checking of many stereo chemical parameters of the residues in the models.^{45,46}

In order to identify structural differences between the models and their templates, the root-mean-square-deviation (RMSD) was also calculated by ICM for the CT domain and for the binding pocket. RMSD is a measure of the degree of similarity of two protein 3D structures, and it calculates the average distance between equivalent backbone C_α atoms by superimposing the models on their templates.^{47,48}

3.3 Molecular Docking

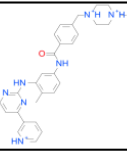
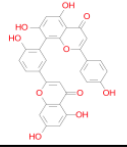

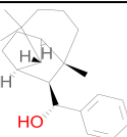
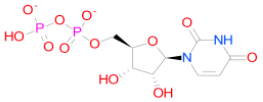
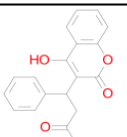
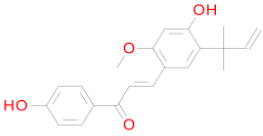
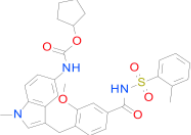
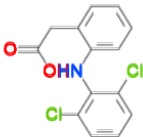

Molecular docking is a method used to predict protein ligand interactions within a targeted binding site, and score their potential complementarity. Exploring these interactions are important for our understanding of how the protein function, and for the development of new drug molecules. Studies on the co-factor binding site of UGT2B17 may provide insights on the formation of the ligand-protein complex, and the intermolecular forces deciding specificity and affinity of a ligand. The formation of a ligand-protein complex may lead to structural changes in both ligand and protein. Retrieving information about protein ligand interactions and can assist in designing new inhibitors with a good fit in the binding pocket.²⁴

3.3.1 Inhibitors and decoys

To validate the models ability to differentiate between inhibitors and decoys, a set of known inhibitors of UGT2B17 was needed. The activity databases PubChem and ChEMBL contained mostly bioassays with compounds binding to the catalytic site, instead of the UDPGA binding site which was the focus of this study. Consequently, inhibitors had to be obtained by examining studies where the UDPGA binding site in UGT2B17, or UGTs in general had been investigated. 17 inhibitors with varying ability to inhibit UGT2B17 were identified, of these were 10 selected based on known IC₅₀ or K_i values, as shown in table 3.

Ideally the docking would be performed with experimentally determined decoys for target, but none were available at time of this study. To acquire decoys, the known inhibitors were entered as templates into the Decoyfinder software. The software generated a set of 145 decoy substances with similar physiochemical properties as the inhibitors. The decoys were inserted into a chemical table with the inhibitors, giving a dataset of 155 substances ready for docking.⁴⁹⁻⁵⁷

Table 3 UGT2B17 inhibitors

ID	Structure	Chemical Composition	CHEMBL ID	PubMed ID	IC ₅₀ (μM)	K _i (μM)
1		C29H34N7O	CHEMBL941	26642944	0,8	0,4
2		C30H18O10	CHEMBL63354	29470958	2,1	2,1
3		C15H16O2	CHEMBL418971	23948605		19,9
4		C21H30O	CHEMBL376840	17474732		21,8
5		C9H12N2O12P2	CHEMBL130266	17998297		100,0
6		C19H16O4	CHEMBL1464	25393417		166,8
7		C21H22O4	CHEMBL139702	26875642		32,0
8		C31H33N3O6S	CHEMBL603	25834030		50,0
9		C14H11Cl2NO2	CHEMBL139	19643121		65,0
10		C13H17O2	CHEMBL521	19643121		1340,0

3.3.2 Ligand and model preparation

The ligands and the models had to be prepared by ICM before the docking procedure. Both the inhibitors and the decoys were converted from 2D structures to 3D conformations, and formal charges were assigned to physiological conditions (pH=7.0) The models had their hydrogens optimized, any missing side chains were hidden, and the residues Histidine, Proline, Asparagine, Glutamine and Cysteine were also optimized.

3.3.3 Identification of ligand binding pocket

Identifying the correct binding pocket to study is essential in the docking process. The UGT2B17 enzyme have an aglycone binding site, and a co-factor binding site, the latter being the focus of this study.

Comparison of the CT domain crystal structure of human UGT2B7 to other GT family enzymes revealed that UDPGA binds to the same site as the co-factor in these enzymes. Bacterial enzymes are part of the GT family and use UDP-glucose as co-factor substrate, while humans use UDP-glucuronic acid. One of the chosen templates, namely the structure of GT OleD (PDB id: 4M83) was crystalized in complex with UDP, this indicated the putative binding site.

Since UDP lacked the glucuronic acid moiety of UDPGA, the binding site determined from UDP by ICM would have been too short and missed several residues of importance. This was solved by an initial docking of UDPGA in the pocket of Model_2O6L indicated by the superimposed UDP, giving an excellent pose and a good score for the docked co-factor. This UDPGA pose was later superimposed on all models, and residues in the models in a 5Å vicinity to the superimposed ligand were selected, thereby defining the binding pocket to be used in the main docking procedure.^{17,18,20}

3.3.4 Docking of inhibitors and decoys

Docking of known inhibitors of UGT2B17 and decoys into the putative binding pocket of the models was carried out to investigate the accuracy of the models. A semi-flexible docking approach was used in this study. This keeps the ligands fully flexible, and the homology models are represented as rigid structures. Protein structure backbone and sidechains of enzymes are considered flexible in nature, with an approach using a rigid binding pocket in the docking, this flexibility is not taken into account.

The binding pocket used in the docking is visualized as an energy grid, with pre-calculated energy maps representing ligand binding interactions such as van der Waals, hydrogen-bonding, electrostatics, and hydrophobic interactions. The box defining the energy grid maps was set to include the entire binding pocket, and to exclude neighbouring cavities which could disturb the docking. The

ligand binding probe in the binding pocket was kept at default, as predicted by ICM using the Monte Carlo global optimization procedure.

The chemical table of inhibitors and decoys was docked into the binding pocket using the docking macro of ICM. Three parallel dockings runs were done on all four models. Once the docking was finished, a collection of the most energetically favourable poses of the ligands were collected and could be displayed interactively inside the binding pocket.

3.3.5 Evaluation of docking

The docking was evaluated using ROC curves, giving insights to the overall predictability of the built models. The scores obtained by the docking process were analysed using the inbuilt ROC-curve command in ICM. The positives (inhibitors) docked were labelled as 1, while negatives (decoys) were labelled as 0. The results were displayed as ROC curves, the AUC was calculated and interpreted. The model with the best AUC values was most capable of discriminating between inhibitors and decoys, and was chosen for further work with VLS.

3.4 Virtual Ligand Screening

The most accurate homology model based on the ROC curve evaluation was selected for virtual ligand screening. The inhibitor with the best activity according to the experimental data combined with the best score from docking (compound 2), was used to screen the database eMolecules for potential hit compounds. A structure similarity search was performed with the cut off set to 50% similarity with compound 2. The screening resulted in a chemical table of 47000 structures. The table was shortened to 36000 after removing compounds with unwanted chemical properties, such as pan-assay interference compounds (PAINS), and compounds predicted to be toxic by the ICM software. PAINS are reactive chemical compounds that often give false readouts in screenings, since they non-specifically react with numerous targets.⁵⁸ The chemical table of 36000 structures was docked in the Model_4AMG with a score cutoff at -25, producing a hitlist of the compounds with better score than the cutoff.

In an effort to obtain isoform specific hit compounds with high affinity to UGT2B17, and not all the other UGT isoforms, compounds making hydrogen bonds with the sidechain of T378 were marked in the hitlist. Residue T378 was specific to the androgen metabolising isoforms UGT2B15 and UGT217, as shown in figure 18. Those compounds not making interactions with T378 were not excluded from the hitlist, but kept as a negative control, as the main focus was to identify high affinity ligands binding to the protein.

Finally, the compounds in the hitlist were clustered by physiochemical properties to get a diverse set for *in vitro* testing, by selecting the best scoring compounds with drug-like properties from each cluster.

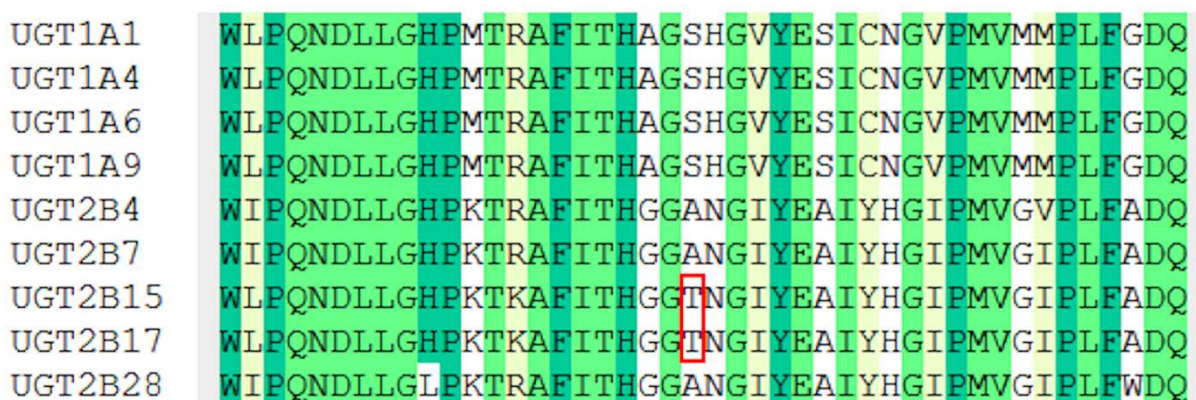


Figure 18 Multiple sequence alignment of binding pocket residues 357-400 for UGTs. Red box shows isoform specific residue T378.

4. Results and discussion

4.1 Homology modelling

The 3D crystal structure of a protein is a valuable tool for investigating the binding properties between a ligand and the binding site. When this study was done, the crystal structure of UGT2B17 was not yet experimentally determined and consequently homology models were constructed as working tools to gain further insights of the interactions between a ligand and the binding site.

4.1.1 Sequence alignment and model building

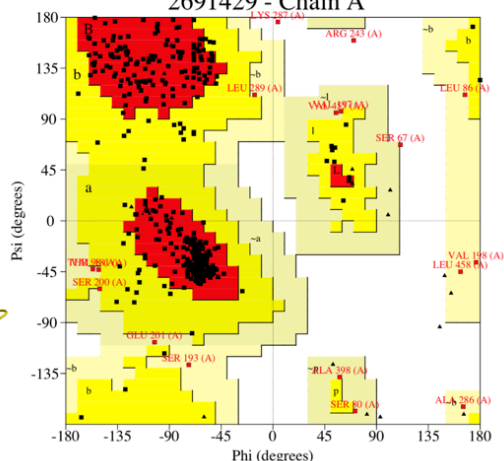
Homology modelling was used to construct and refine four 3D models of the target enzyme UGT2B17 based on known crystal structures of homologous proteins. Three full length models containing both NT and CT domains, and one partial model of the CT domain were built using the homology modelling macro inbuilt in ICM. The enzymes used as templates were all part of the GT enzyme superfamily, namely UGT2B7 (PDB id: 2O6L), GT VinC (PDB id: 3WAD), GT SnogD (PDB id: 4AMG), and GT OleD (PDB id: 4M83). Their resolution was 2.0Å, 2.59Å, 1.7Å, and 1.8Å, as shown in table 2.

A multiple sequence alignment with the templates combined with several other human UGTs was done to have a basis for further adjustment. The three bacterial GT templates were about 130 residues shorter than target, and had low sequence identity with target, leading to a difficult alignment process. The membrane interacting region in the NT domain added extra complexity, since the GT templates lacked this region. This resulted in several gaps in the alignments, these were shifted to the loop regions if possible. The main focus was to get the core region of the NT domain aligned correctly. A few site-directed mutagenesis studies were available, these aided in the alignment process by highlighting regions of importance. The sequence identities for the adjusted alignments were 82% for UGT2B7, 20% for GT VinC, 22% for GT SnogD, and 21% for GT OleD, as shown in table 2. The sequence identity between target and template strongly correlates with model accuracy, and three of the alignments had a low sequence identity. At the time of this study only one known UGT crystal structure of was available, namely the partial structure of UGT2B7. Because of the low sequence identity between GT templates and target, a multi template model procedure was done. After an initial construction of models based solely on their GT template, the partial structure of UGT2B7 was added as a second template for the CT domain, improving the quality of the models in this region. To be able to utilize homology models for VLS, a sequence identity above 60% would be preferred. By combining two templates, the overall sequence identity was raised to approximately 55% for the three full length models.

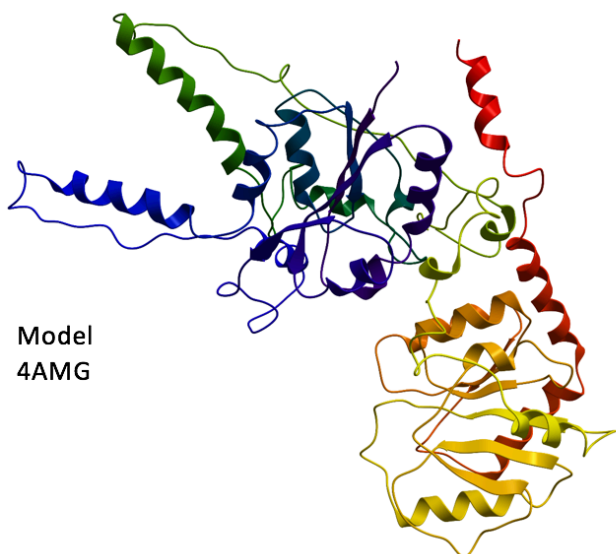
Model
3WAD



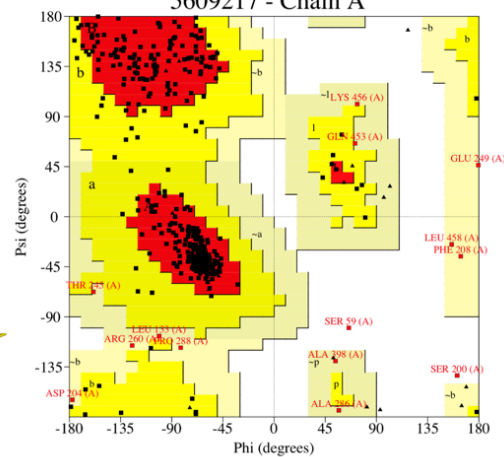
Ramachandran Plot
2691429 - Chain A



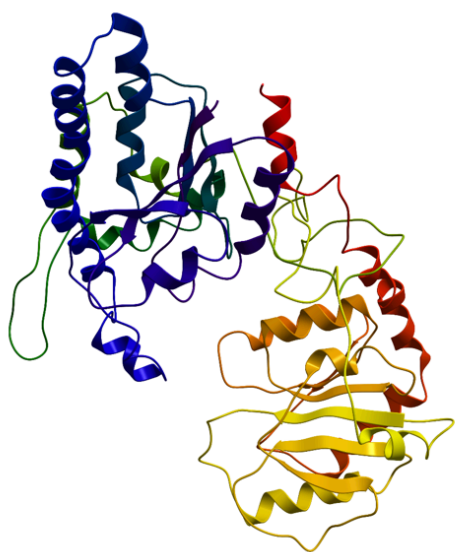
Model
4AMG



Ramachandran Plot
5609217 - Chain A



Model
4M83



Ramachandran Plot
8757137 - Chain A

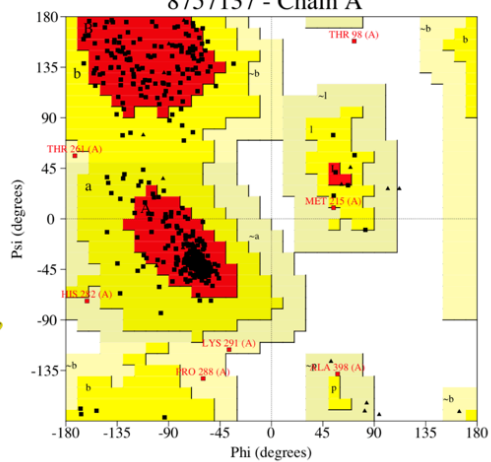


Figure 19 Homology models and their corresponding Ramachandran Plot. Models are visualized as ribbons, with the protein chain colour scheme of a rainbow, from blue at the amino-terminus to red at the carboxy-terminus. Ramachandran Plot was generated by ProCheck, showing residues in the most favoured regions (red), additionally allowed regions (yellow), generously allowed regions (beige) and disallowed regions (white).

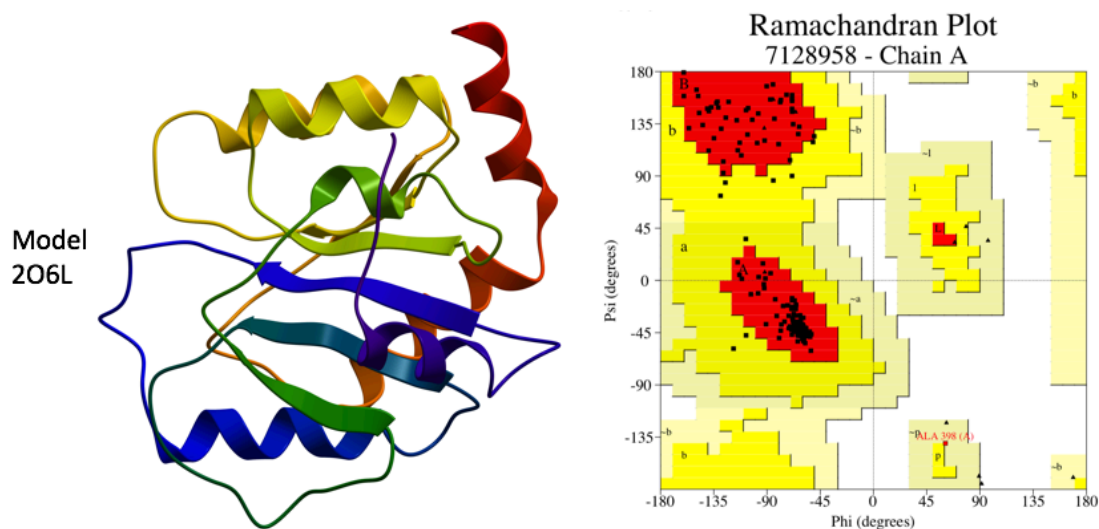


Figure 20 Homology model 2O6L and its corresponding Ramachandran Plot. Model is visualized as ribbon, with the protein chain colour scheme of a rainbow, from blue at the amino-terminus to red at the carboxy-terminus. Ramachandran Plot was generated by ProCheck, showing residues in the most favoured regions (red), additionally allowed regions (yellow), generously allowed regions (beige) and disallowed regions (white).

The final models were called Model_3WAD, Model_4AMG and Model_4M83, and Model_2O6L, as shown in figure 19 and 20. The partial model contained the co-factor binding site, and all full length models contained the typical structural characteristics of GTs and UGTs, a highly diverse NT domain with the aglycone binding site, a conserved CT domain with the co-factor binding site, and a catalytic cleft between them.

Ideally the homology models should have been based on templates with the appropriate conformational state. The purpose of this study was to aid in the development of inhibitors of UGT2B17, and thus the templates should preferably all have been in complex with inhibitors for high quality models. Only one template in complex with a ligand, combined with acceptable sequence identity in the NT domain, was available when identifying potential templates.

Homology models are dependent of the sequence identity between target and template, and on the quality of the template crystal structures. A modelled structure should be considered as a guide to be used as a working tool, and never as reliable as an experimentally derived structure.

4.1.2 Model validation

The models and their corresponding Ramachandran plots generated by ProCheck are displayed in figure 19 and 20, their ProCheck statistics are shown in table 4. The Ramachandran plot visualizes the stereo chemical quality of the models, the overall and the residue-by-residue geometry. A good quality model is expected to have over 90% of the residues within the most favoured regions. According to the Ramachandran plots, Model_2O6L was within the threshold of a good quality model with a percentage of 94.2% The other models had 83.1% 86.0% and 87.5% respectively, which is just below the limit for good quality models. Of the residues making up the presumed co-factor binding pocket, none were in the disallowed regions, and only A398 was in the generously allowed region.

The models were also evaluated using the WhatCheck tool, which did extensive checking of many stereochemical parameters of the residues in the models. All models passed the overall summary rapport, confirming that the models were of satisfactory quality. In conclusion, Model_2O6L was of good quality, the others were of acceptable quality.^{45,46}

Table 4 Ramachandran plot statistics generated by ProCheck

Model	Most favoured regions	Additionally allowed regions	Generously allowed regions	Disallowed regions
3WAD	83.1%	12.6%	3.6%	0.8%
4AMG	86.0%	10.8%	2.8%	0.5%
4M83	87.5%	11.0%	1.3%	0.3%
2O6L	94.2%	5.0%	0.7%	0.0%

To investigate if the CT domain of models resemble their template, the models were superimposed on their UGT template, as shown to the right in figures 21-24. In addition, the RMSD of backbone C α was calculated for the CT domain and for the binding pocket, as shown in table 5. RMSD describes the degree of similarity between superimposed structures, and RMSD values are presented in Ångstrom (Å). Low RMSD values below 2Å means the two structures are similar, while a value of 0Å implicate that two structures are identical in conformation.

RMSD values for the CT domain and the residues forming the binding pocket are shown in table 5. The models were decently conserved on the CT domain template, with RMSD values ranging from 0.157Å to 1.759. The binding pocket was better conserved, with RMSD values ranging from 0.149Å to 0.840Å, this indicated that the models were correct in this region.

Table 5 Calculated RMSD for the models

Model	RMSD for the CT domain	RMSD for the binding pocket
3WAD	1.264Å	0.445Å
4AMG	1.759Å	0.840Å
4M83	1.184Å	0.685Å
2O6L	0.157Å	0.149Å

Some degree of uncertainty in the models will always be present, since the templates have structural similarities with target, but are not identical. The uncertainty in these models were most profound in the NT domains, due to gaps in the alignments and low sequence identity. This resulted in substantial structural differences, with some long secondary structures pointing out of the NT domain of the models. Model_4AMG had two long helices pointing outwards, while the other two models had a few extra loop regions. The models superimposed on their corresponding GT templates are shown to the left in figures 21-23. Since the co-factor binding site built up by the CT domain and some residues in the core of the NT domain was the main interest in this study, the uncertainty in the peripheral secondary structures of the NT domain were of less importance.

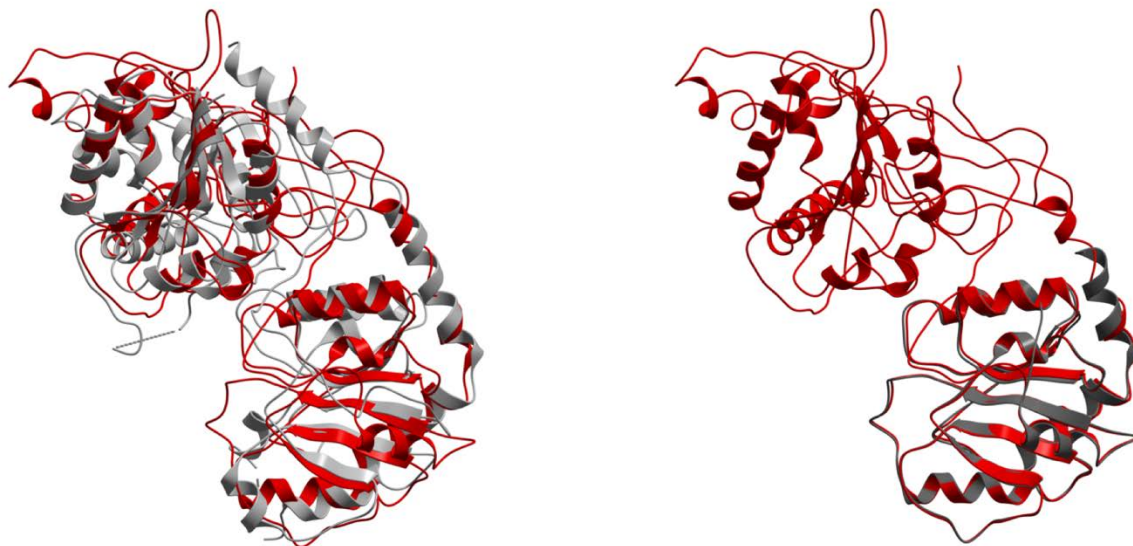


Figure 21 Model_3WAD superimposed on templates. Model shown in red, templates in grey. Left figure shows 3WAD template, and right figure shows 2O6L template

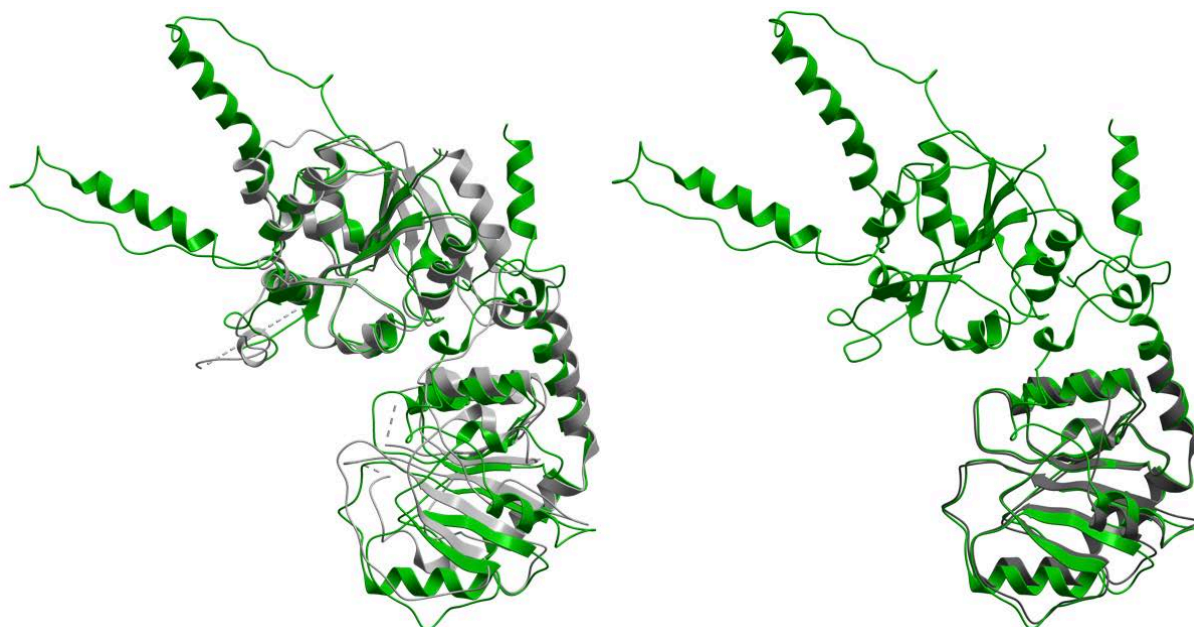


Figure 22 Model_4AMG superimposed on templates. Model shown in green, templates in grey. Left figure shows 4AMG template, and right figure shows 2O6L template.

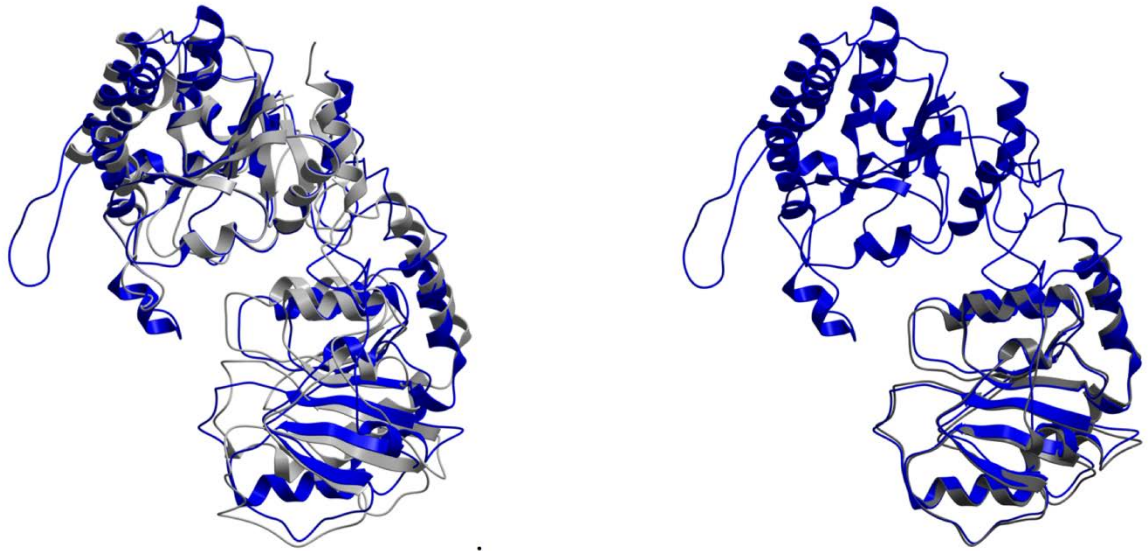


Figure 23 Model_4M83 superimposed on templates. Model shown in blue, templates in grey. Left figure shows 4M83 template, and right figure shows 2O6L template.

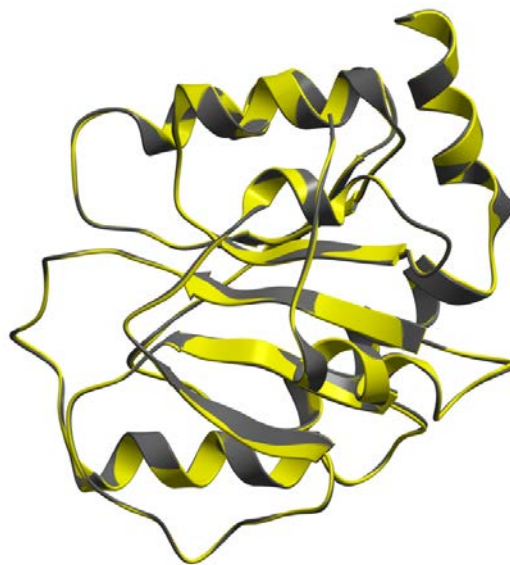


Figure 24 Model_2O6L superimposed on template. Model shown in yellow, template in grey.

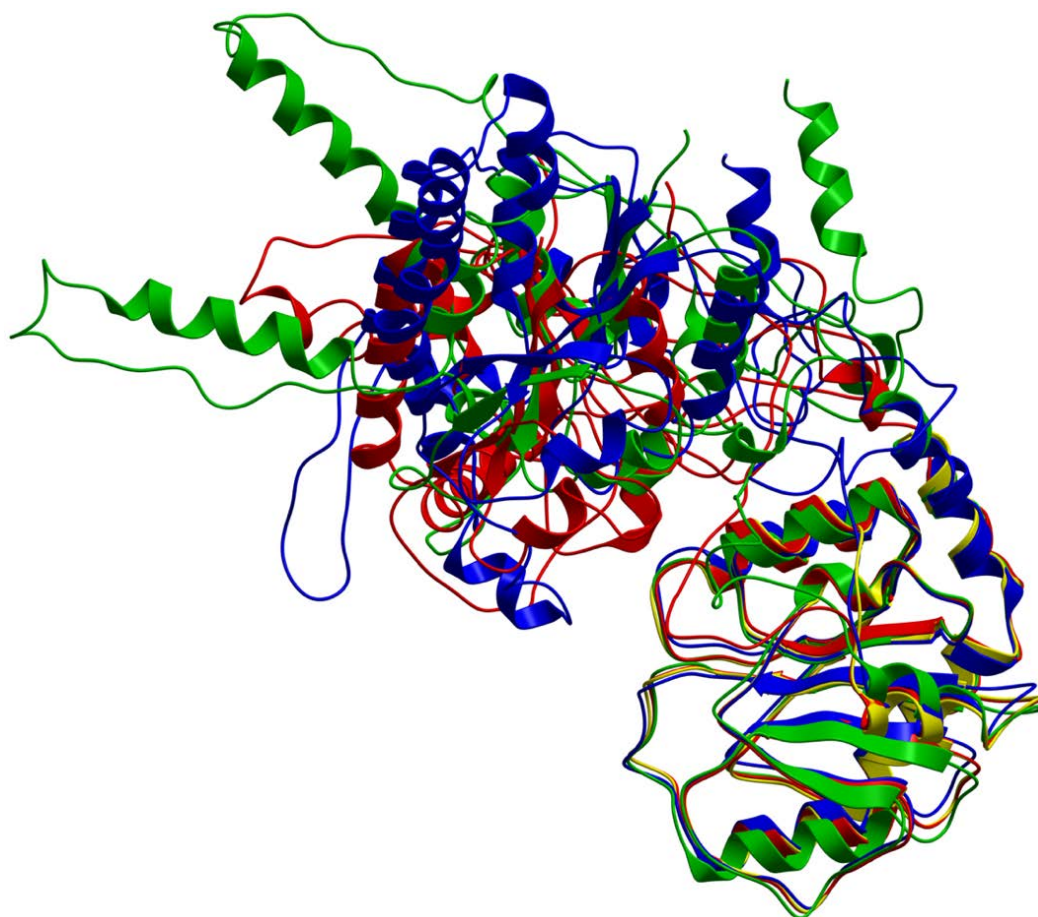


Figure 25 The four homology models superimposed. Model_3WAD shown as red, Model_4AMG shown as green, Model_4M83 shown as blue, and Model_2O6L shown as yellow.

When the four models were superimposed, as shown in figure 25, the structural differences and similarities of the models became evident. The NT domain of the templates were highly variable, resulting in very different models. The CT domains were highly conserved in all the templates, adding a second template for this region made the structural similarity even better.

Some of the residues from the site-directed mutagenesis studies were conserved in the models, as seen in the alignments. The two residues involved in the catalytic reaction of the enzyme (H35 and D152) were positioned close to the UDPGA binding site in the core of all the models, as needed to initiate the catalysis, as shown in figure 26. The residue F90 responsible for ring stacking interactions with the aglycone was also in close proximity in Model_4AMG and Model_4M83. These conserved residues at key positions increased the possibility that the built models were similar to the target in the core region of the enzymes, despite the relatively low sequence identity in the NT domain.

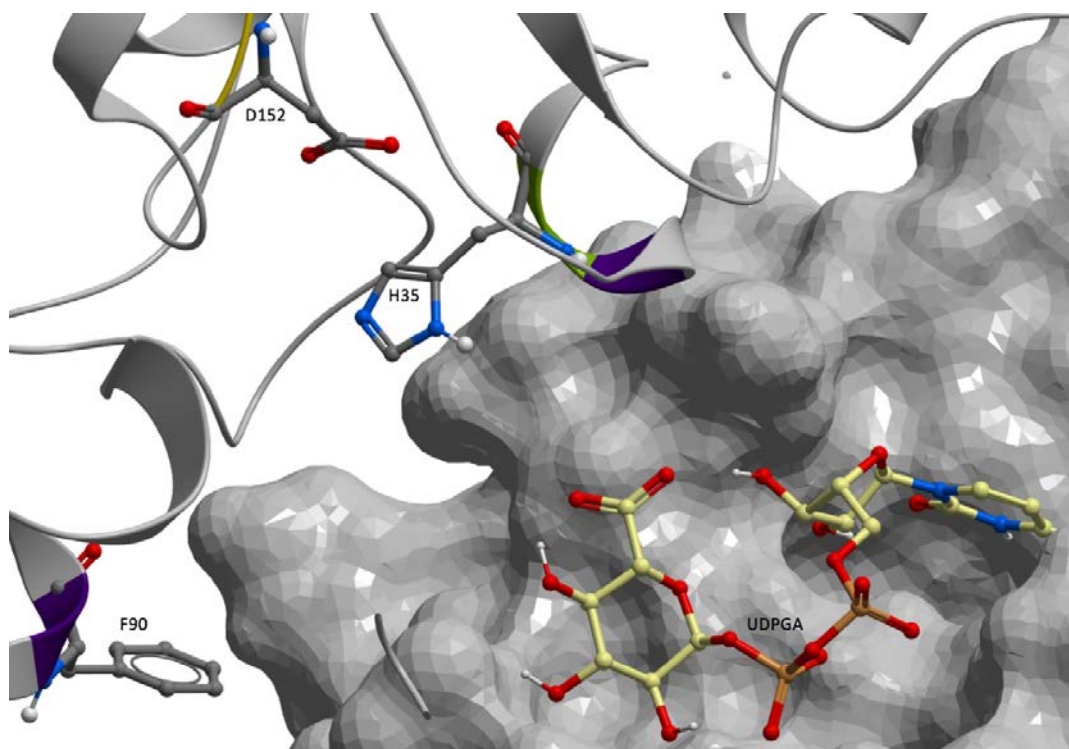


Figure 26 Residues H35 and D152 forming the catalytic dyad, and residue F90 crucial for interactions with the aglycone, all in close proximity to UDPGA situated in the binding pocket of Model_4AMG. NT domain shown as ribbon, CT domain shown as mesh.

4.2 Molecular docking

4.2.1 Identification of ligand binding pocket

Enzymes that are part of the GT family share a common co-factor binding site located in the CT domain of the protein. Several experimentally determined crystal structures have confirmed this, including the templates used in this study. Superimposing the structure of GT OleD in complex with UDP on the models indicated the putative binding site. An initial docking run of UDPGA in the binding pocket of Model_2O6L indicated by the superimposed UDP was performed. This gave the co-factor UDPGA an excellent pose and fit in the binding pocket, with a good docking score of -49, as shown in figure 27.

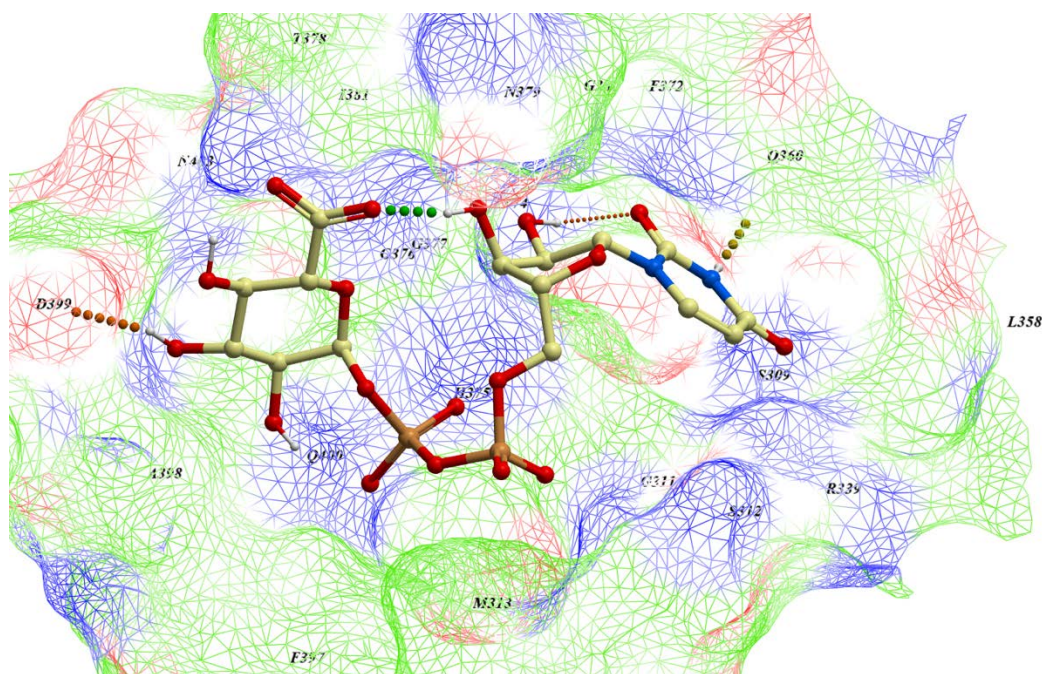


Figure 27 UDPGA docked into binding pocket of Model_2O6L. The surface of the pocket is coloured by the binding properties. Aromatic lipophilic shown as white. Aliphatic lipophilic shown as green, HBA potential shown as red, HBD potential shown as blue.

This UDPGA pose was superimposed on all models, and residues in a 5Å vicinity to the superimposed UDPGA ligand were selected, thereby defining the binding pocket to be used in the docking process. Model_3WAD, Model_4AMG, Model_4M83 and Model_2O6L had 37, 32, 33 and 27 residues defining the binding pocket respectively. Model_3WAD had the linker region between the two domains in close vicinity to the co-factor binding pocket, adding extra residues and narrowing the pocket. Table 6 lists all residues forming the binding pocket of the models.

Table 6 Residues forming the co-factor binding site in the models. Residues most likely to have contact with ligands are shown as bold.

Model	Residues forming the co-factor binding site
3WAD	E32, Y33, S34 , H35, I37, N38, V278, G279, G280, L281, S309 , G311, S312, M313, R339 , K356, W357 , L358, P359, Q360 , N361, L364, F372, T374 , H375 , G376, G377, T378, N379 , G380 , I381, E383 , F397, A398, D399 , Q400 , N403
4AMG	E32, Y33, S34 , V278, K284, S309 , G311, S312, M313, R339 , K356, W357 , L358, P359, Q360 , N361, L364, F372, T374 , H375 , G376, G377, T378, N379 , G380 , I381, E383 , F397, A398, D399 , Q400 , N403
4M83	D88, F90, M93, H282, K284, S309 , G311, S312, M313, R339 , K356, W357 , L358, P359, Q360 , N361, L364, T374 , H375 , G376, G377, T378, N379 , G380 , I381, Y382, E383 , F397, A398, D399 , Q400 , N403
2O6L	S309 , G311, S312, M313, R339 , K356, W357 , L358, P359, Q360 , N361, L364, F372, T374 , H375 , G376, G377, T378, N379 , G380 , I381, E383 , F397, A398, D399 , Q400 , N403

Comparing several crystal structures of GTs in complex with ligand (including GT OleD) with crystal structures without a ligand, have shown a conformational change of W357, moving the residue closer to the ligand. The conformational change, presumably initiated by co-factor binding, could make ring stacking interactions possible between the aromatic ring of W357 and the uracil of UDPGA. This conformational change makes the residue important for ligand binding, despite its initial peripheral placement in the binding pocket.^{20,59}

4.2.2 Docking of inhibitors and decoys

A chemical table of 10 inhibitors and 145 decoys were docked in semi-flexible mode into the putative binding site of UGT2B17, to evaluate the ability of the homology models to differentiate between them. The experimentally determined binding affinities of the inhibitors are shown in table 3. The inhibitors were a diverse set of compounds, with different degree of inhibition of the target enzyme. The binding poses of the docked inhibitors were investigated, and the score values analysed. Residues of importance for binding of UDPGA are shown as bold in table 6. Most of the inhibitors were docked into the electronegative centre of the pocket, as shown in figure 28. When the most accurate model (Model_4AMG) was docked, the score values for the inhibitors ranged from +0.9 to -23.5.

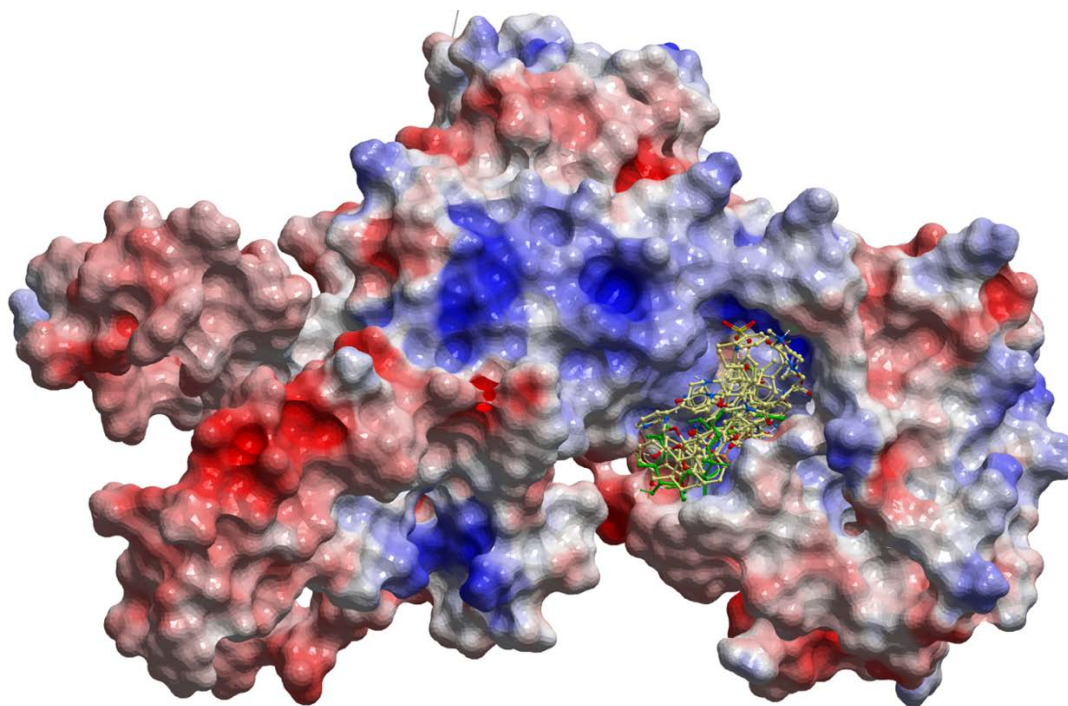


Figure 28 Model_4AMG shown as electrostatic potential. Areas coloured blue represent positive areas, red represents negative areas, and white represents neutral areas. The ten docked inhibitors shown in binding pocket. UDPGA added as reference, shown in green.

Compound 1 was according to the experimental data the best inhibitor, but it failed to bind properly, with the best score value of -0.3 from three parallel dockings. When investigating the binding poses of this structure, it was evident that it had penetrated the binding pocket wall and partially docked into neighbouring cavities, in all parallels. This indicated that the defined binding pocket was too narrow for the high MW molecule. Compound 8 was another high MW structure with a poor docking score of +0.9. A possible explanation could be that the preferred binding poses for these structures included both the co-factor binding pocket and the neighbouring aglycone binding pocket, since a marginal expansion of the energy grid box used in docking failed to improve the results. Apart from these, all other inhibitors had good negative score values. Compound 6 had a score value of -23.5 from the docking, which was the best score of the ten inhibitors. This compound did not have the highest binding affinity, being a non-competitive inhibitor with a K_i of $166.7\mu\text{M}$ according to the experimental data, excluding it from further investigation. Compound 5 had the second best score value of -21.2, but this structure included uridine, making it undesirable as a screening probe since similar structures could adversely affect the critical biological processes utilizing nucleotide sugars. Compound 2 was the second best inhibitor according to the experimental data. The docking gave a score value of -14.8, which was the third best score. The compound had a binding pose in the centre of the pocket, involving interactions with many of the residues confirmed important by site-directed mutagenesis studies. This compound was used to investigate the binding pocket of the most accurate model, and for further studies with VLS.

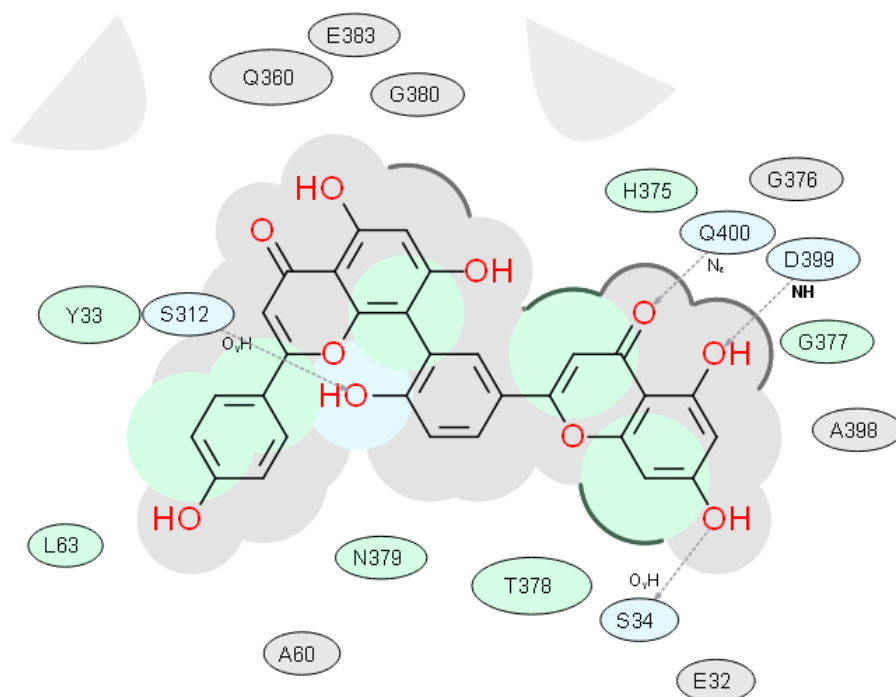


Figure 29 Ligand 2D interaction diagram for compound 2. Green shading represents hydrophobic region. Blue shading represents hydrogen bond acceptor. Grey dashed arrows indicates hydrogen bonds. Size of residue ellipse represents the strength of contact. Distance between ligand and residue label represents proximity. Broken thick line around ligand shape indicates accessible surface. Grey parabolas represents accessible surface for large areas.

Compound 2 was investigated with regard to ligand binding interactions with the binding pocket, as seen in figure 29 and 30. The structure could make hydrophobic interactions with Y33, L63, H375, G377, T378, and N379. One of the aromatic rings was in a good position to make ring stacking interactions with Y33. Hydrogen bonds were observed to S34, S312, D399 and Q400, the distance being 2.9Å, 2.4Å, 3.2Å and 2.5Å respectively. There was also a possibility that hydrogen bonds could also be formed with residues Q360, T378 and N379.

The crystal structure of UGT2B7 shows the presence of water molecules in the binding pocket, suggesting that some of these could be involved in hydrogen bond interactions between ligands and the protein. The presence of water in the pocket was not accounted for in the docking process, but when superimposing the water molecules of the UGT2B7 template on Model_4AMG, water mediated interactions were possible with R339 and T374.

The ligand binding interactions described are not as static as the figures show. In reality, both the enzyme and the ligand have natural structural flexibility and motion, making it easier to interact with residues in the binding pocket. The presumed conformational change upon ligand binding may also affect the binding pocket interactions. These figures are snapshots of how the ligand-protein complex could appear in reality. The lack of this flexibility in the docking is one of the main drawbacks of molecular docking studies.

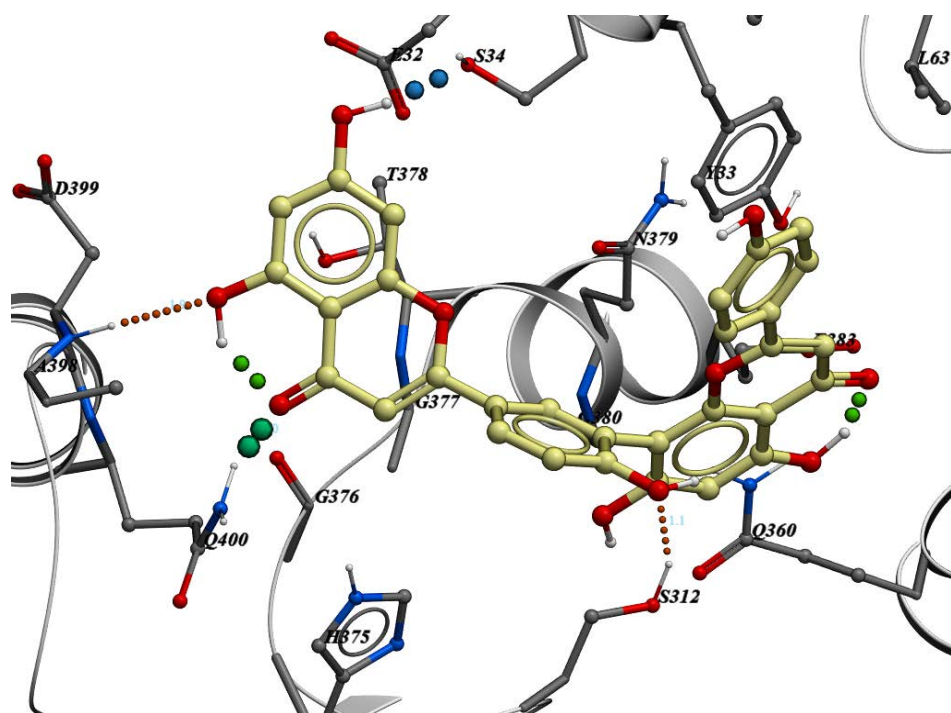


Figure 30 Presumed binding interactions between compound 2 and the binding pocket.

Site directed mutagenesis studies were available for other UGTs. These confirmed that residues S34, S309, R339, W357, Q360, E373, T374, H375, N379, G380, E383, D399 and Q400, positioned in the binding pocket, were involved in co-factor binding. Several of these residues were involved in binding interactions with compound 2, indicating that the models were correct. This study has shown that residues Y33, L63, S312, G377, T378 may also be involved in ligand binding interactions. These residues along with the other NT domain residues listed in table 6 could be interesting to study further in future site-directed mutagenesis studies. If experimental studies through a crystal structure determination, or site-directed mutagenesis studies could confirm these residues as important, it would be possible to conclude that the models were partially correct, and that the proposed residues were involved in ligand binding.^{20,22}

4.2.3 Evaluation of docking

ROC curves were made to evaluate the ability of the models to differentiate between binders (inhibitors) and non-binders (decoys). The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the model. A diagonal curve represents a model which is not able to discriminate between true positives and false positives. The calculated AUC for the ROC curves is a measure of the accuracy of the models, shown in percentage. Three parallel docking runs were performed for the results to be statistically viable. The ROC curves of all docking runs are shown in figures 31-34, the calculated AUC is shown in table 7.

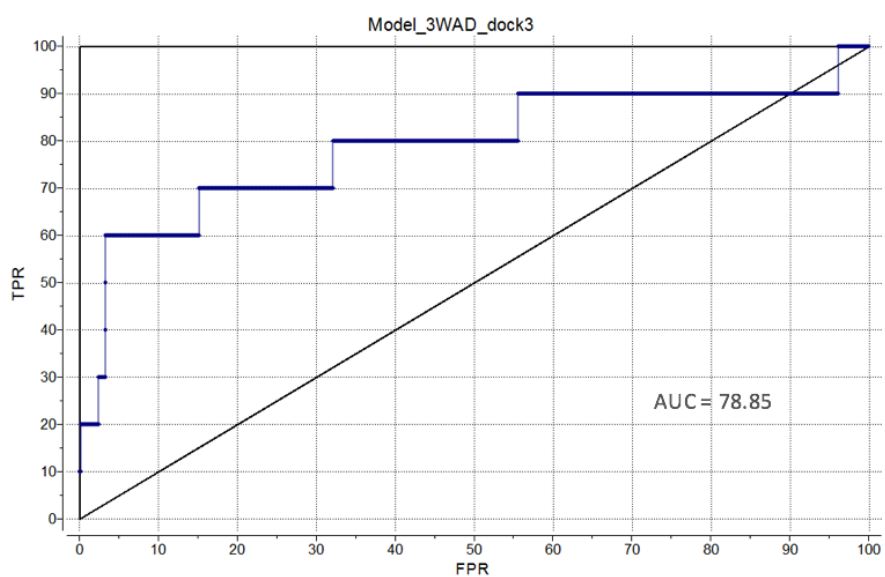
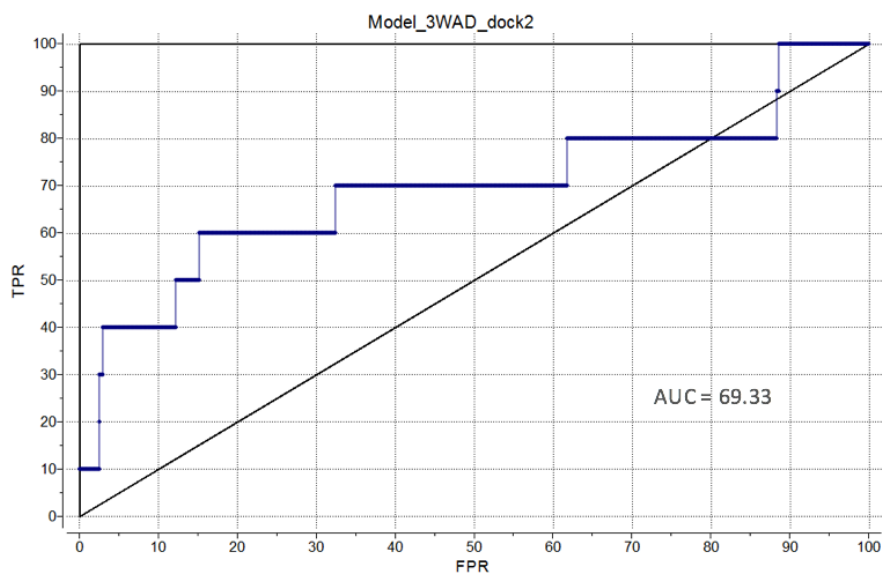
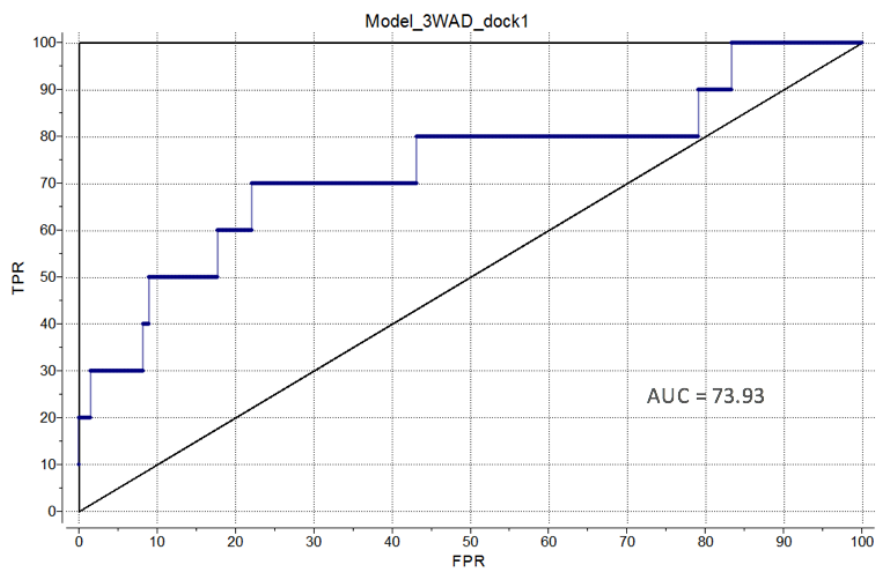


Figure 31 ROC curves for 3 parallel dockings on Model_3WAD. True positive rate on y-axis, false positive rate on x-axis.

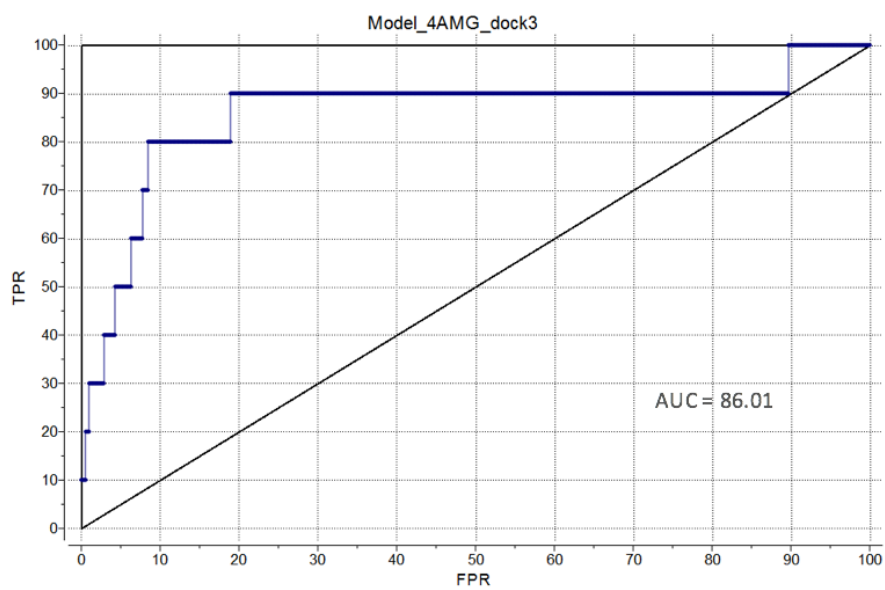
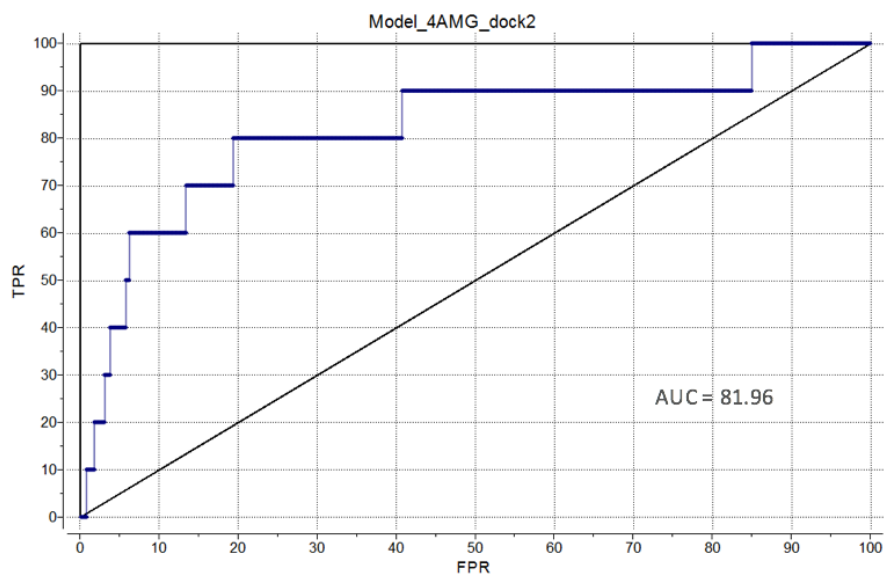
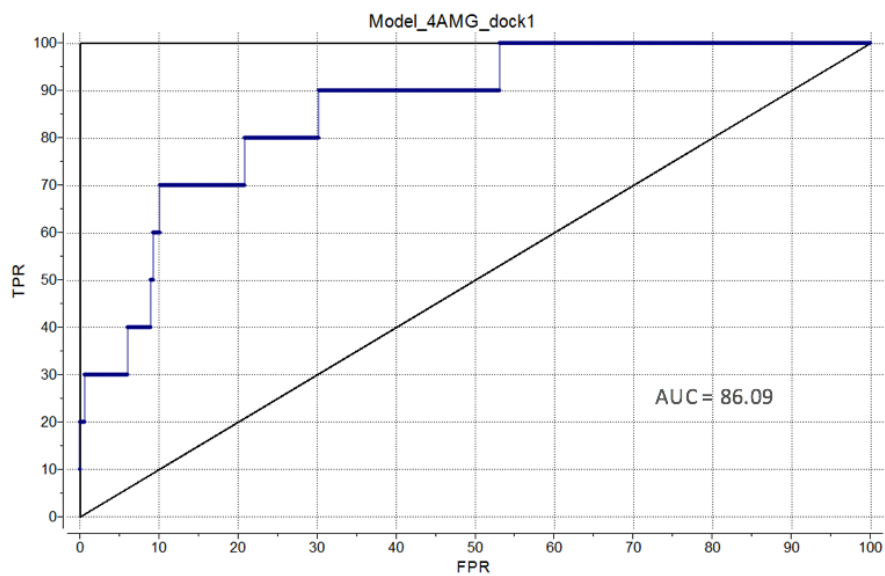


Figure 32 ROC curves for 3 parallel dockings on Model_4AMG. True positive rate on y-axis, false positive rate on x-axis.

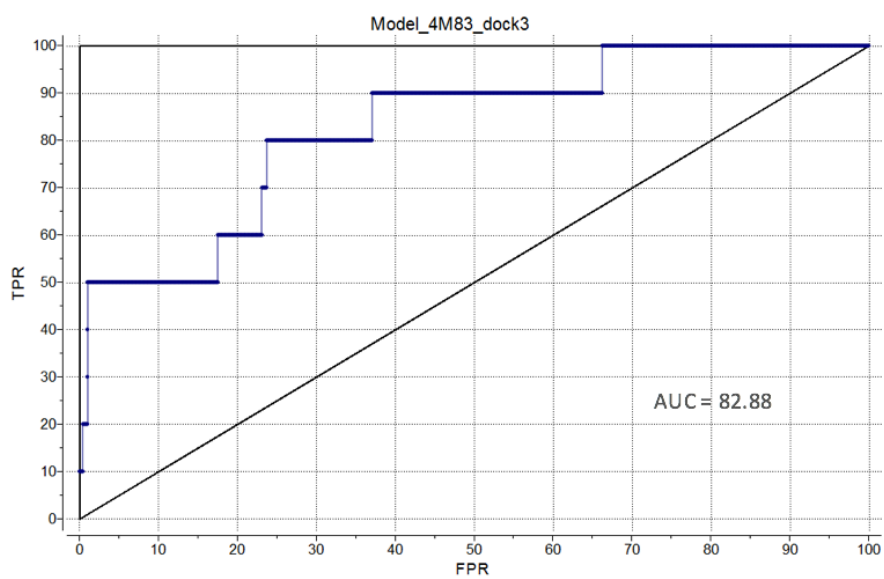
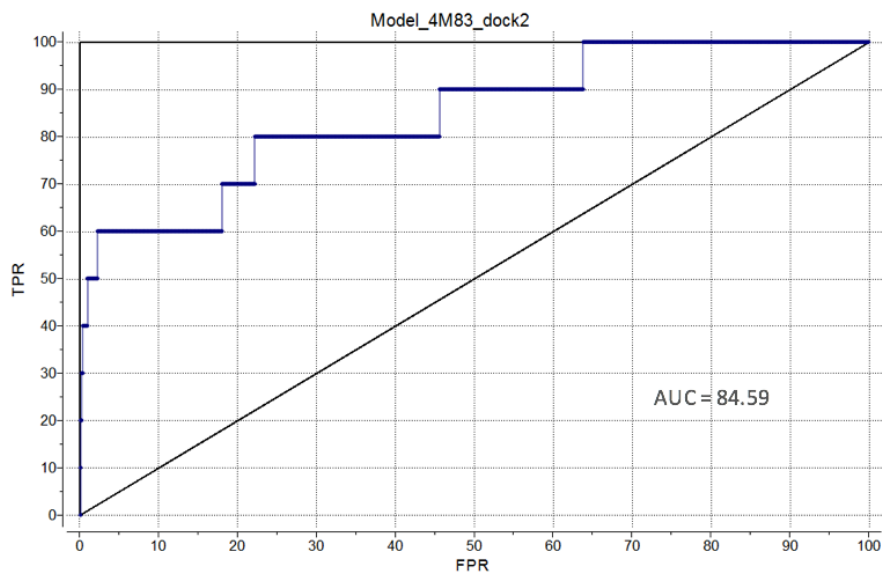
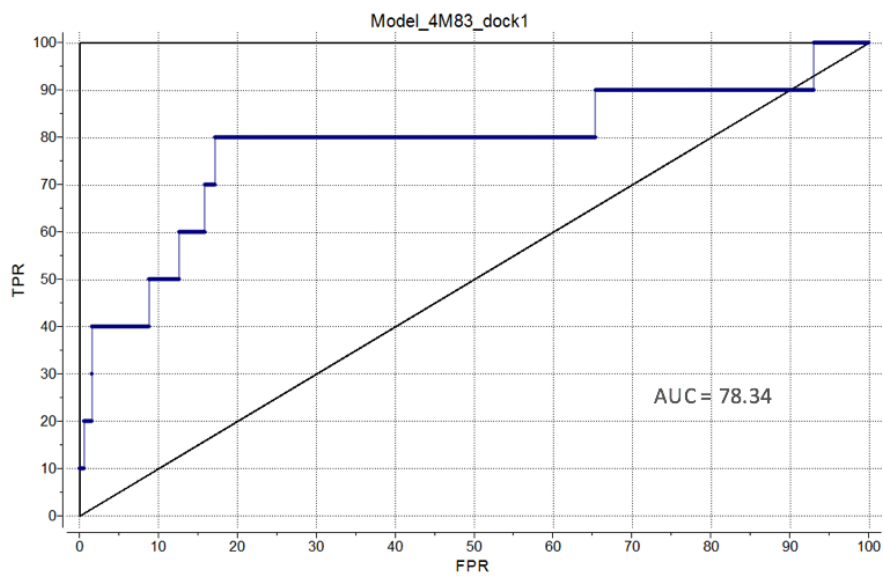


Figure 33 ROC curves for 3 parallel dockings on Model_4M83. True positive rate on y-axis, false positive rate on x-axis.

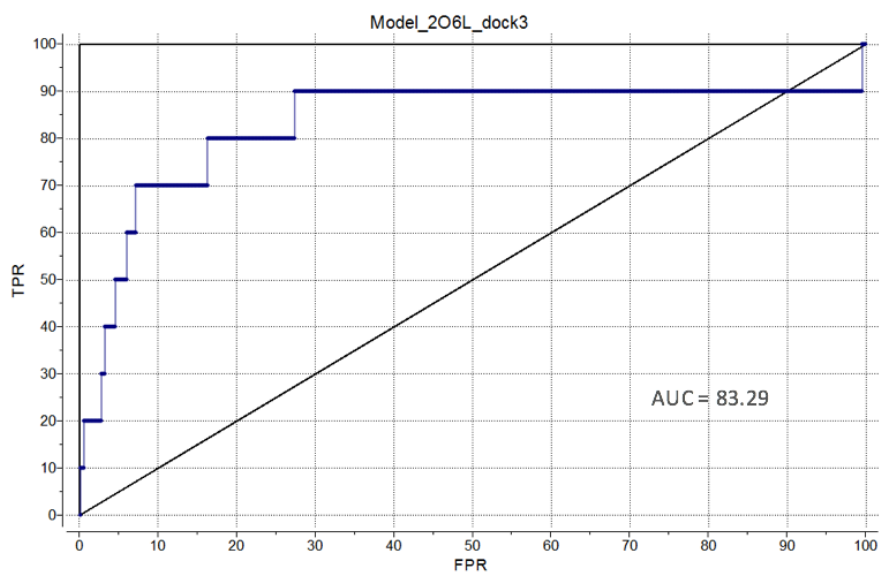
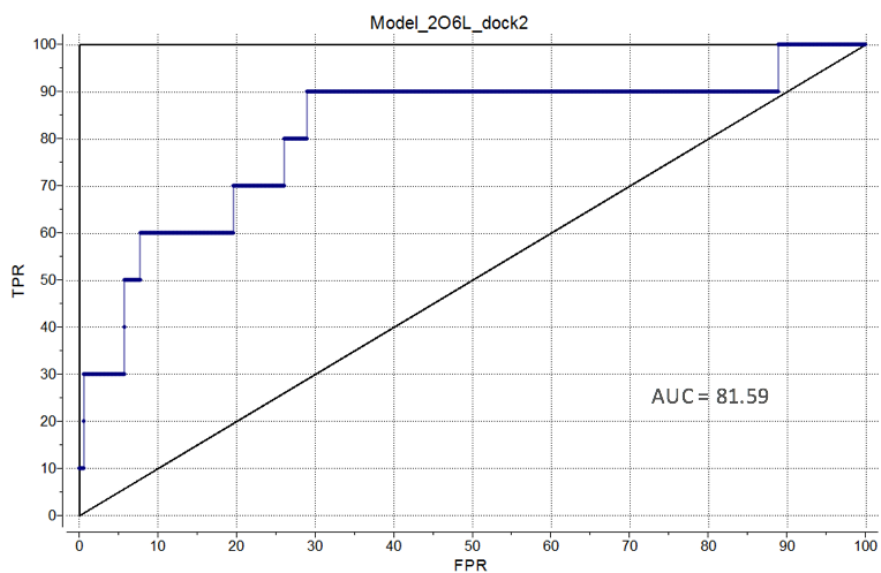
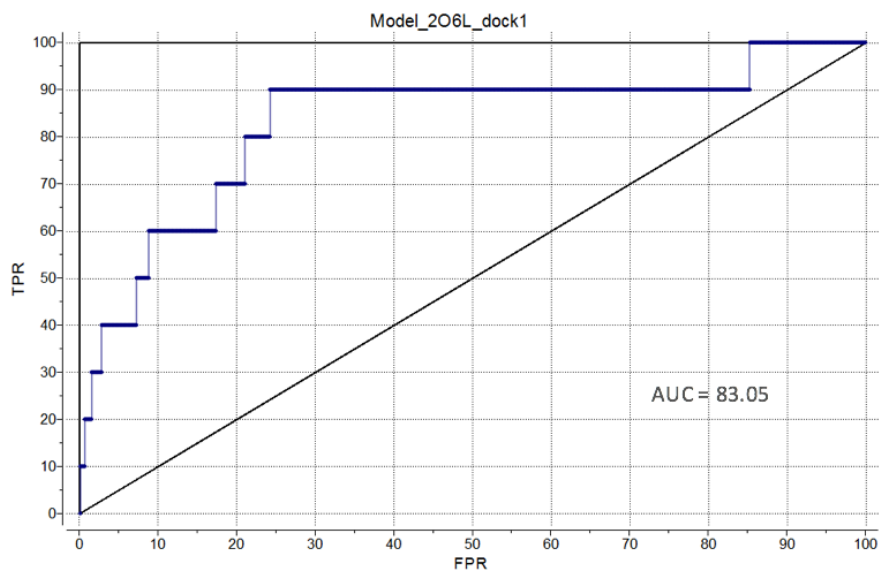


Figure 34 ROC curves for 3 parallel dockings on Model_2O6L. True positive rate on y-axis, false positive rate on x-axis.

Table 7 Calculated AUC for all docking runs

Model	Docking parallel	AUC	mean AUC
3WAD	1	73.61	73.93%
	2	69.33	
	3	78.85	
4AMG	1	86.09	84.68%
	2	81.96	
	3	86.01	
4M83	1	78.34	81.93%
	2	84.59	
	3	82.88	
2O6L	1	83.05	82.61%
	2	81.59	
	3	83.29	

Model_3WAD had ROC curves closer to the diagonal than the other, a mean AUC value of 73.93, indicating its fairly good at differentiating true binders from decoys. Model_4M83 had a mean AUC value of 81.93. Model_2O6L had a mean AUC value of 82.61. These models were more accurate than Model_3WAD, but had ROC curves crossing the diagonal line. Model_4AMG had ROC curves furthest away from the diagonal line curve and never crossing it. This model had the highest calculated AUC of 86.09, and mean AUC value was 84.68, which was the highest mean value of the models, indicating an accurate model. In conclusion, Model_4AMG was the most accurate in distinguishing between true binders and non-binders, and was studied further with VLS.

Model_4AMG was according to the ROC curves more accurate than Model_2O6L, despite the lower sequence identity and RMSD of binding pocket. This indicates that the NT domain residues of Model_4AMG could be important for ligand binding.

A bad value from the calculated AUC can indicate that a model is unable to distinguish between decoys and inhibitors. However, this does not necessarily imply that the model is inaccurate. In general, the ROC curve depends heavily on the choice of decoys. Decoys for the ROC curves were generated by the Decoyfinder software, since none were available from experimental data. These decoys are compounds with similar physiochemical properties and MW, presumed to be inactive. Without experimentally determined decoys, some of the theoretical decoys could actually be true binders, and may have generated false negatives, affecting the TPR in the curves.

The defined binding pocket used in the docking was quite open, increasing the possibility for decoys to bind, and consequently it could be difficult to distinguish between inhibitors and decoys. This could possibly lead to false positives affecting the FPR.

The use of ROC curves as a statistical method to evaluate a model is not optimal with few known inhibitors. Only ten inhibitors with known affinity were available at the time of this study. Ideally the enzyme should have at least fifty known binders to make a detailed graphical plot for evaluation.

4.3 Virtual ligand screening

Compound 2 was according to the experimental data the second best inhibitor. The docking confirmed its binding mode in the pocket, with a score value of 14,8. Consequently, this structure was chosen as drug candidate to screen a chemical database for potential hit compounds. The screening resulted in a chemical table of 47000 structures with a 50% similarity with compound 2. The chemical table was shortened to 36000 after removing compounds with unwanted chemical properties. These structures were docked in the Model_4AMG, producing a preliminary hitlist of the 250 compounds with score values better than -25.

The structures in the hitlist were visually inspected, confirming that their binding poses were in the centre of the putative binding pocket of Model_4AMG where the known inhibitors did bind. The residues from the site-directed mutagenesis studies were involved in ligand binding interactions with the hitlist structures. In addition, the residues proposed in this study were also involved. Many structures made hydrogen bonds with S312 and T378, while ring stacking interactions with Y33 were also observed for several structures. These residues may prove to be important for ligand binding to UGT2B17.

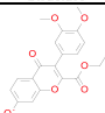
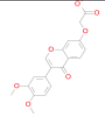
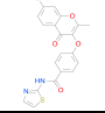
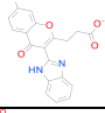
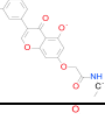
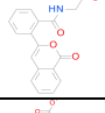
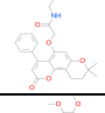
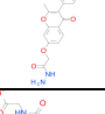
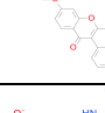
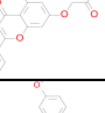
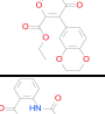
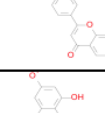
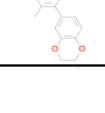
The hitlist was clustered based on physicochemical properties to get a diverse set of potential lead compounds for *in vitro* testing. The 25 compounds with the best docking scores from each cluster, combined with drug-like properties, were selected for experimental studies, as shown in table 8.

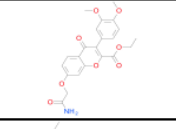
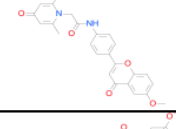
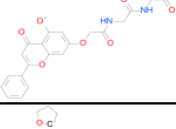
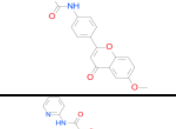
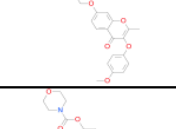
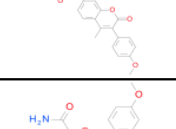
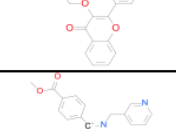
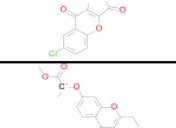
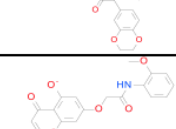
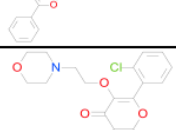
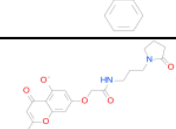

Good docking score for hitlist structures does not imply high affinity, only experimental testing can confirm if a structure binds to the pocket. As long as a structure is found in the hitlist, the relative ranking becomes less important. Focusing solely on structures with good score values may lead to discrimination of structures posed in a conformation with lower score. The VLS method is a theoretical approach to be used as a guide for discovering rational compounds for *in vitro* testing.

The drug-like properties of the structures were predicted by ICM, and shown as a druglikeness score. Druglikeness above zero indicates that the structure may have drug-like properties, but does not predict if a structure is pharmacologically active.

In the hitlist, the structures forming hydrogen bonds with T378 were marked as potential isoform specific inhibitors for UGT2B15 and UGT2B17. The polar sidechain T378 is specific to the two androgen metabolising enzymes within the UGT2 family, and should be the focus of lead optimization studies if experimental testing confirms activity.

Table 8 Hitlist with proposed UGT2B17 inhibitors selected from VLS

Structure	Molecular formula	Score	MW	LogP	PSA	HBA	HBD	RB	Druglikeness	T378	eMolecules ID
	C20H18O7	-35,3	369,1	4,2	73,9	10	0	6	0,72	✓	748274
	C19H16O7	-35,1	355,1	3,1	72,0	10	0	6	0,26	✓	4810962
	C21H16N2O5S	-35,0	408,1	3,8	68,8	8	1	6	0,74		17106338
	C19H14N2O5	-34,9	349,1	2,9	85,8	9	2	4	0,32		1303729
	C22H19NO9	-34,0	438,1	2,3	112,8	14	1	7	0,65		298143056
	C18H13NO5	-33,7	322,1	1,8	75,4	9	1	5	0,29		42976602
	C25H25NO7	-33,2	450,2	3,6	86,7	11	1	8	0,46		4843851
	C20H20N2O6	-32,8	384,1	1,9	91,8	9	3	7	0,71		4595739
	C20H16ClNO6	-32,4	400,1	3,1	80,7	10	1	7	1,16	✓	17653094
	C19H15NO7	-31,7	367,1	2,2	98,7	12	1	7	0,70		17673402
	C20H16O7	-31,6	367,1	4,0	74,5	10	0	4	1,00	✓	1107581
	C24H17NO5	-31,3	398,1	5,0	72,6	9	1	5	0,59		182622610
	C18H14O6	-30,9	325,1	3,2	70,3	8	1	1	0,82	✓	4788008

Structure	Molecular formula	Score	MW	LogP	PSA	HBA	HBD	RB	Druglikeness	T378	eMolecules ID
	C22H21NO8	-30,9	427,1	3,2	97,2	11	2	9	0,83	✓	2211387
	C25H22N2O5	-30,3	430,2	3,6	67,0	8	1	6	0,83		182651162
	C21H18N2O8	-29,8	424,1	1,3	123,2	14	2	10	0,70	✓	17672928
	C21H19NO5	-28,5	364,1	4,2	60,8	7	1	5	0,87		182386540
	C24H20N2O6	-27,5	432,1	3,5	75,2	9	1	8	0,72		13584317
	C22H21NO6	-26,3	395,1	3,1	58,9	8	0	5	0,63		5528175
	C18H15NO5	-25,6	325,1	2,6	69,0	7	2	5	1,47		1530112
	C25H17ClN2O5	-25,3	459,1	4,8	67,4	9	0	5	1,08		4662678
	C23H22O7	-25,2	409,1	4,1	64,3	9	0	6	1,05	✓	642667
	C24H19NO6	-25,1	416,1	4,5	75,6	9	1	7	0,71		13583015
	C21H20ClN2O4	-25,1	385,1	4,0	39,6	6	0	5	1,51	✓	13140587
	C24H24N2O6	-25,1	435,2	3,0	87,4	10	1	9	0,90		4847976

4.4 Future aspects

In this study, homology models were constructed as working tools to aid in the design of experimental studies related to UGT2B17. Drug discovery and development is a lengthy process, as seen in figure 35, and the work done here were the beginning stages of the process. Through VLS, 25 compounds have been identified as potential lead candidates for UGT2B17. These compounds should be further tested *in vitro* to determine their binding affinity to UGT2B17. In case the compounds bind to the enzyme as proposed, the next step would be optimization of target interactions and pharmacokinetic properties, and subsequent *in vivo* testing for affinity. Any compounds able to pass through all these stages could be considered as a potential drug candidates ready for clinical trials.

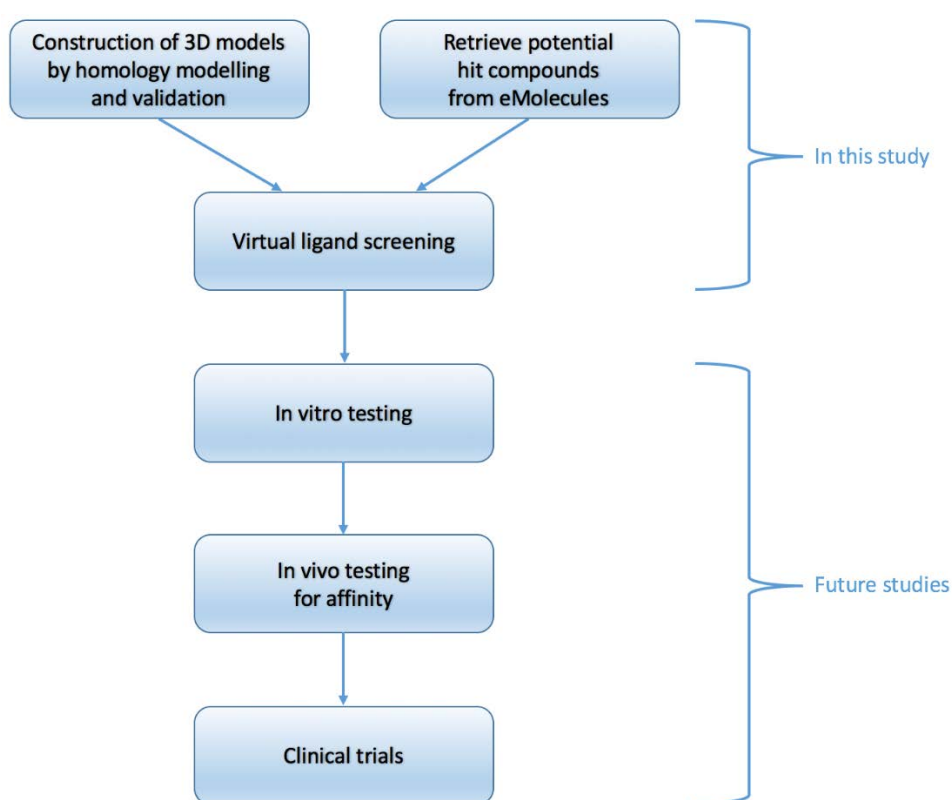


Figure 35 Steps in drug development

5. Conclusion

In the absence of a crystal structure for UGT2B17, four homology models based on different templates were constructed to improve the understanding of the ligand binding properties of the enzyme.

Modelling a membrane bound protein without a full length human template was a challenge, but structural analysis showed that the models were in agreement with experimental data, with residues vital for ligand binding present in all models. The models were of acceptable quality for docking studies, with Model_4AMG the most accurate in discriminating between inhibitors and decoys. This model was used as a working tool to gain insights in the interactions between ligand and binding pocket, allowing us to propose residues that may be involved in ligand binding interactions. These residues could be good candidates for future site-directed mutagenesis studies. Virtual ligand screening generated a hitlist of 25 compounds for future *in vitro* testing to determine their binding affinity for UGT2B17.

6. References

1. Nussey S, Whitehead S. *Principles of Endocrinology*. BIOS Scientific Publishers; 2001.
2. Sand O, Haug E, Toverud KC, Sjaastad Ø V. *Menneskets Fysiologi*. 2nd ed. Oslo: Gyldendal akademisk; 2014.
3. Melmed S, Michael Conn P. *Endocrinology : Basic and Clinical Principles*. 2nd ed. Totowa, N.J.: Humana Press; 2005.
4. Svartberg J. Skal eldre menn behandles med testosteron? *Tidsskr Nor Legeforen*. 2005;7(125):879-882.
5. Svartberg J, Midtby M, Bønaa KH, Sundsfjord J, Joakimsen RM, Jorde R. The associations of age, lifestyle factors and chronic disease with testosterone in men: the Tromsø Study. *Eur J Endocrinol*. 2003;149(2):145-152.
6. Dankers ACA, Roelofs MJE, Piersma AH, et al. Endocrine Disruptors Differentially Target ATP-Binding Cassette Transporters in the Blood-Testis Barrier and Affect Leydig Cell Testosterone Secretion In Vitro. *Toxicol Sci*. 2013;136(2):382-391. doi:10.1093/toxsci/kft198.
7. Joensen UN, Veyrand B, Antignac J-P, et al. PFOS (perfluorooctanesulfonate) in serum is negatively associated with testosterone levels, but not with semen quality, in healthy men. *Hum Reprod*. 2013;28(3):599-608. doi:10.1093/humrep/des425.
8. Rang HP, Dale MM, Flower RJ (Rod J., Henderson G (Graeme). *Rang and Dale's Pharmacology*. 8th ed. Edinburgh: Elsevier Churchill Livingstone; 2016.
9. Patrick GL. *An Introduction to Medicinal Chemistry*. 6th ed. Oxford: Oxford University Press; 2017.
10. Meech R, Mackenzie PI. Structure and function of uridine diphosphate glucuronosyltransferases. *Clin Exp Pharmacol Physiol*. 1997;24(12):907-915. doi:10.1111/j.1440-1681.1997.tb02718.x.
11. Luukkanen L, Taskinen J, Kurkela M, Kostianen R, Hirvonen J, Finel M. Kinetic characterization of the 1A subfamily of recombinant human UDP-glucuronosyltransferases. *Drug Metab Dispos*. 2005;33(7):1017-1026. doi:10.1124/dmd.105.004093.
12. Kuuranne T, Saugy M, Baume N. Confounding factors and genetic polymorphism in the evaluation of individual steroid profiling. *Br J Sports Med*. 2014;48(10):848-855. doi:10.1136/bjsports-2014-093510.

13. Bernard JP, Opdal MS, Khiabani H. Generelle farmakodynamiske prinsipper. *Tidsskr den Nor Laegeforening*. 2006;126(16):2107-2109. doi:10.04.2015.
14. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev*. 2001;46(1-3):3-26.
15. Veber DF, Johnson SR, Cheng H-Y, Smith BR, Ward KW, Kopple KD. Molecular properties that influence the oral bioavailability of drug candidates. *J Med Chem*. 2002;45(12):2615-2623.
16. Brändén C-I, Tooze J. *Introduction to Protein Structure*. 2nd ed. New York: Garland Pub; 1999.
17. Breton C, Šnajdrová L, Jeanneau C, Koča J, Imberty A. Structures and mechanisms of glycosyltransferases. *Glycobiology*. 2006;16(2):29R-37R. doi:10.1093/glycob/cwj016.
18. Radomska-Pandya A, Czernik P, Little JM, Battaglia E, Mackenzie PI. Structural and functional studies of UDP-glucuronosyltransferases. *Drug Metab Rev*. 1999;31(4):817-899. doi:10.1081/DMR-100101944.
19. Gauthier-Landry L, Bélanger A, Barbier O. Multiple roles for udp-glucuronosyltransferase (UGT)2B15 and UGT2B17 enzymes in androgen metabolism and prostate cancer evolution. *J Steroid Biochem Mol Biol*. 2015;145:187-192. doi:10.1016/j.jsbmb.2014.05.009.
20. Miley MJ, Zielinska AK, Keenan JE, Bratton SM, Radomska-Pandya A, Redinbo MR. Crystal structure of the cofactor-binding domain of the human phase II drug-metabolism enzyme UDP-glucuronosyltransferase 2B7. *J Mol Biol*. 2007;369(2):498-511. doi:10.1016/j.jmb.2007.03.066.
21. Mackenzie PI, Owens IS, Burchell B, et al. The UDP glycosyltransferase gene superfamily: recommended nomenclature update based on evolutionary divergence. *Pharmacogenetics*. 1997;7(4):255-269.
22. Radomska-Pandya A, Bratton SM, Redinbo MR, Miley MJ. The crystal structure of human UDP-glucuronosyltransferase 2B7 C-terminal end is the first mammalian UGT target to be revealed: the significance for human UGTs from both the 1A and 2B families. *Drug Metab Rev*. 2010;42(1):133-144. doi:10.3109/03602530903209049.
23. Sliwoski G, Kothiwale S, Meiler J, Lowe EW, Jr. Computational methods in drug discovery. *Pharmacol Rev*. 2014;66(1):334-395. doi:10.1124/pr.112.007336.
24. Orry AJW, Abagyan R. *Homology Modeling : Methods and Protocols*. Humana Press; 2012.
25. Leach AR. *Molecular Modelling : Principles and Applications*. 2nd ed. Harlow: Prentice Hall;

- 2001.
26. Buenavista MT, Roche DB, McGuffin LJ. Improvement of 3D protein models using multiple templates guided by single-template model quality assessment. *Bioinformatics*. 2012;28(14):1851-1857. doi:10.1093/bioinformatics/bts292.
 27. Anfinsen CB. Principles that govern the folding of protein chains. *Science*. 1973;181(4096):223-230.
 28. Abagyan R, Totrov M. Biased Probability Monte Carlo Conformational Searches and Electrostatic Calculations for Peptides and Proteins. *J Mol Biol*. 1994;235(3):983-1002. doi:10.1006/jmbi.1994.1052.
 29. Kitchen DB, Decornez H, Furr JR, Bajorath J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat Rev Drug Discov*. 2004;3(11):935-949. doi:10.1038/nrd1549.
 30. Hajian-Tilaki K. Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Casp J Intern Med*. 2013;4(2):627-635.
 31. Klebe G. Virtual ligand screening: strategies, perspectives and limitations. *Drug Discov Today*. 2006;11(13-14):580-594. doi:10.1016/j.drudis.2006.05.012.
 32. Abagyan R, Totrov M, Kuznetsov D. ICM—A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation. *J Comput Chem*. 1994;15(5):488-506. doi:10.1002/jcc.540150503.
 33. Neves MAC, Totrov M, Abagyan R. Docking and scoring with ICM: the benchmarking results and strategies for improvement. *J Comput Aided Mol Des*. 2012;26(6):675-686. doi:10.1007/s10822-012-9547-0.
 34. Berman HM, Westbrook J, Feng Z, et al. The protein data bank. *Nucleic Acids Res*. 2000;28(1):235-242. doi:10.1093/nar/28.1.235.
 35. Pundir S, Martin MJ, O'Donovan C. UniProt Protein Knowledgebase. In: Humana Press, New York, NY; 2017:41-55. doi:10.1007/978-1-4939-6783-4_2.
 36. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215(3):403-410. doi:10.1016/S0022-2836(05)80360-2.
 37. Kim S, Thiessen PA, Bolton EE, et al. PubChem Substance and Compound databases. *Nucleic Acids Res*. 2016;44(D1):D1202-13. doi:10.1093/nar/gkv951.
 38. Gaulton A, Bellis LJ, Bento AP, et al. ChEMBL: a large-scale bioactivity database for drug

- discovery. *Nucleic Acids Res.* 2012;40(D1):D1100-D1107. doi:10.1093/nar/gkr777.
39. Cereto-Massagué A, Guasch L, Valls C, Mulero M, Pujadas G, Garcia-Vallvé S. DecoyFinder: an easy-to-use python GUI application for building target-specific decoy sets. *Bioinformatics.* 2012;28(12):1661-1662. doi:10.1093/bioinformatics/bts249.
 40. Claesson M, Siitonen V, Dobritzsch D, Metsä-Ketelä M, Schneider G. Crystal structure of the glycosyltransferase SnogD from the biosynthetic pathway of nogalamycin in *Streptomyces nogalater*. *FEBS J.* 2012;279(17):3251-3263. doi:10.1111/j.1742-4658.2012.08711.x.
 41. Patana A, Kurkela M, Finel M, Goldman A. Mutation analysis in UGT1A9 suggests a relationship between substrate and catalytic residues in UDP-glucuronosyltransferases. *Protein Eng Des Sel.* 2008;21(9):537-543. doi:10.1093/protein/gzn030.
 42. Xiong Y, Bernardi D, Bratton S, et al. Phenylalanine 90 and 93 Are Localized within the Phenol Binding Site of Human UDP-Glucuronosyltransferase 1A10 as Determined by Photoaffinity Labeling, Mass Spectrometry, and Site-Directed Mutagenesis [†]. *Biochemistry.* 2006;45(7):2322-2332. doi:10.1021/bi0519001.
 43. Senay C, Ouzzine M, Battaglia E, et al. Arginine 52 and histidine 54 located in a conserved amino-terminal hydrophobic region (LX2-R52-G-H54-X3-V-L) are important amino acids for the functional and structural integrity of the human liver UDP-glucuronosyltransferase UGT1*6. *Mol Pharmacol.* 1997;51(3):406-413.
 44. Dubois SG, Beaulieu M, Lévesque E, Hum DW, Bélanger A. Alteration of human UDP-glucuronosyltransferase UGT2B17 regio-specificity by a single amino acid substitution. *J Mol Biol.* 1999;289(1):29-39. doi:10.1006/jmbi.1999.2735.
 45. Laskowski RA, MacArthur MW, Moss DS, Thornton JM, IUCr. PROCHECK: a program to check the stereochemical quality of protein structures. *J Appl Crystallogr.* 1993;26(2):283-291. doi:10.1107/S0021889892009944.
 46. Hooft RWW, Vriend G, Sander C, Abola EE. Errors in protein structures. *Nature.* 1996;381(6580):272-272. doi:10.1038/381272a0.
 47. Maiorov VN, Crippen GM. Significance of Root-Mean-Square Deviation in Comparing Three-dimensional Structures of Globular Proteins. *J Mol Biol.* 1994;235(2):625-634. doi:10.1006/jmbi.1994.1017.
 48. Kufareva I, Abagyan R. Methods of protein structure comparison. *Methods Mol Biol.* 2012;857:231-257. doi:10.1007/978-1-61779-588-6_10.
 49. Zhang N, Liu Y, Jeong H. Drug-Drug Interaction Potentials of Tyrosine Kinase Inhibitors via

- Inhibition of UDP-Glucuronosyltransferases. *Sci Rep*. 2015;5:17778. doi:10.1038/srep17778.
50. Lv X, Zhang J-B, Wang X-X, et al. Amentoflavone is a potent broad-spectrum inhibitor of human UDP-glucuronosyltransferases. *Chem Biol Interact*. 2018;284:48-55. doi:10.1016/j.cbi.2018.02.009.
 51. Jiang H-M, Fang Z-Z, Cao Y-F, et al. New insights for the risk of bisphenol A: Inhibition of UDP-glucuronosyltransferases (UGTs). *Chemosphere*. 2013;93(6):1189-1193. doi:10.1016/j.chemosphere.2013.06.070.
 52. Bichlmaier I, Kurkela M, Joshi T, et al. Isoform-Selective Inhibition of the Human UDP-glucuronosyltransferase 2B7 by Isolongifolol Derivatives. *J Med Chem*. 2007;50(11):2655-2664. doi:10.1021/jm061204e.
 53. Fujiwara R, Nakajima M, Yamanaka H, Katoh M, Yokoi T. Product Inhibition of UDP-Glucuronosyltransferase (UGT) Enzymes by UDP Obfuscates the Inhibitory Effects of UGT Substrates. *Drug Metab Dispos*. 2008;36(2):361-367. doi:10.1124/dmd.107.018705.
 54. Sun H, Zhang T, Wu Z, Wu B. Warfarin is an Effective Modifier of Multiple UDP-Glucuronosyltransferase Enzymes: Evaluation of its Potential to Alter the Pharmacokinetics of Zidovudine. *J Pharm Sci*. 2015;104(1):244-256. doi:10.1002/jps.24250.
 55. Xin H, Qi X-Y, Wu J-J, et al. Assessment of the inhibition potential of Licochalcone A against human UDP-glucuronosyltransferases. *Food Chem Toxicol*. 2016;90:112-122. doi:10.1016/j.fct.2016.02.007.
 56. Oda S, Fujiwara R, Kutsuno Y, et al. Targeted Screen for Human UDP-Glucuronosyltransferases Inhibitors and the Evaluation of Potential Drug-Drug Interactions with Zafirlukast. *Drug Metab Dispos*. 2015;43(6):812-818. doi:10.1124/dmd.114.062141.
 57. Sten T, Finel M, Ask B, Rane A, Ekström L. Non-steroidal anti-inflammatory drugs interact with testosterone glucuronidation. *Steroids*. 2009;74(12):971-977. doi:10.1016/J.STEROIDS.2009.07.004.
 58. Baell J, Walters MA. Chemistry: Chemical con artists foil drug discovery. *Nature*. 2014;513(7519):481-483. doi:10.1038/513481a.
 59. Qasba PK, Ramakrishnan B, Boeggeman E. Substrate-induced conformational changes in glycosyltransferases. *Trends Biochem Sci*. 2005;30(1):53-62. doi:10.1016/J.TIBS.2004.11.005.

