# Ligand Recognition and Specificity of Metabolic Enzymes and Nuclear Receptors

**Inauguraldissertation**
zur
Erlangung der Würde eines Doktors der Philosophie
vorgelegt der
Philosophisch-Naturwissenschaflichen Fakultät
der Universität Basel

von

André Fischer

2023

Originaldokument gespeichert auf dem Dokumentenserver der Universität Basel

edoc.unibas.ch

Dieses Werk ist lizenziert unter einer Creative Commons Attribution-NonCommercial 4.0 International License.

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät
auf Antrag von

Prof. Dr. Daniel Ricklin
PD Dr. Martin Smieško
Prof. Dr. Amedeo Caflisch

Basel, den 19. Oktober 2021

Dekan
Prof. Dr. Marcel Mayor

# Acknowledgements

I am very grateful to my supervisor **PD Dr. Martin Smieško** who introduced me to the field of computational chemistry, offered unlimited consulting, and provided a pleasant working environment. His global expertise in computational medicinal chemistry fundamentally contributed to shape the researcher that I have become.

I would like to thank my secondary supervisor **Prof. Dr. Daniel Ricklin**, who was always there if I had a problem and enriched my multidisciplinarity by introducing his cutting-edge work around the pharmaceutical modulation of the complement system. The collaborative meetings and projects were a pleasure.

I am grateful to **Prof. Dr. Amedeo Caflisch** who kindly offered to be my external expert despite his busy schedule.

Further, I would like to express my gratitude to our group head **Prof. Dr. Markus A. Lill** who continuously offered scientific advice and introduced me to the power of artificial intelligence and deep learning.

I would like to thank **my colleagues** of the Computational Pharmacy group who were collaborative, always helpful with advice, and offered pleasant lunch conversations. Here I would like to mention **Manuel Sellner**, who constantly provided me with his advice.

Last, but not least, I would like to thank **my parents Suzanne and Franz Fischer** for their boundless support and allowing me to become who I am.

# TABLE OF CONTENTS

**Acknowledgements** . . . . . . . . . . . . . . . . . . . . . . . . . . . i

**1 Introduction 1**

1.1 Preamble: Ligand Recognition and Specificity . . . . . . . . . . . . . . . . 1

1.2 Pharmacological Background . . . . . . . . . . . . . . . . . . . . . . . . . 2

1.2.1 Drug Metabolism . . . . . . . . . . . . . . . . . . . . . . . . . 2

1.2.2 Cytochrome P450 Enzymes . . . . . . . . . . . . . . . . . . . . . . 2

1.2.3 Human Carboxylesterases . . . . . . . . . . . . . . . . . . . . . . . 4

1.2.4 Nuclear Receptors . . . . . . . . . . . . . . . . . . . . . . . . 5

1.3 Molecular Modeling . . . . . . . . . . . . . . . . . . . . . . . . . . . 7

1.3.1 The basis of molecular modeling in drug discovery . . . . . . . . 7

1.3.2 The most relevant interactions of drugs with proteins . . . . . . . 9

1.3.3 Molecular docking . . . . . . . . . . . . . . . . . . . . . . . . 13

1.3.4 MD simulations . . . . . . . . . . . . . . . . . . . . . . . . . . 15

1.3.5 The application of MD simulations . . . . . . . . . . . . . . . . . 18

1.3.6 Cheminformatics and machine learning . . . . . . . . . . . . . . 23

**2 Spontaneous Ligand Access Events to Membrane-Bound CYP2D6 Sampled at Atomic Resolution 42**

2.1 Supporting Information . . . . . . . . . . . . . . . . . . . . . . . . . . 62

**3 A Conserved Allosteric Site on Drug-Metabolizing CYPs: A Systematic Computational Assessment 88**

3.1 Supporting Information . . . . . . . . . . . . . . . . . . . . . . . . . . 111

**4 Computational Prediction of Ester Hydrolysis by Human Carboxylesterases 1 and 2 127**

4.1 Supporting Information . . . . . . . . . . . . . . . . . . . . . . . . . . 150

**5 Review: Ligand Pathways in Nuclear Receptors 164**

**6 Ligand Pathways in Estrogen-Related Receptors 189**

6.1 Supporting Information . . . . . . . . . . . . . . . . . . . . . . . . . . . 211

**7 Conformational Changes of Thyroid Receptors in Response to Antagonists 216**

7.1 Supporting Information . . . . . . . . . . . . . . . . . . . . . . . . . . . 240

**8 Allosteric Binding Sites On Nuclear Receptors: Focus On Drug Efficacy and Selectivity 247**

8.1 Supporting Information . . . . . . . . . . . . . . . . . . . . . . . . . . . 275

**9 Computational Assessment of Combination Therapy of AR Targeting Compounds 295**

9.1 Supporting Information . . . . . . . . . . . . . . . . . . . . . . . . . . . 318

**List of Publications** . . . . . . . . . . . . . . . . . . . . . . . . . . . 332

**Curriculum Vitae** . . . . . . . . . . . . . . . . . . . . . . . . . . . 334

## CHAPTER 1

# Introduction

## 1.1 Preamble: Ligand Recognition and Specificity

Molecular recognition is a key process in the formation of ligand-protein complexes concerning both selectivity and stability of binding. Hence, it is a fundamental principle for most biological processes in the human body. In the context of pharmaceutical chemistry, it covers ligand-protein interactions, effects of the surrounding solvent, allosteric regulation, and conformational adaptation - important effects for the design of drug molecules (Figure 1). The foundation for the principle of recognition was already proposed in the 19th century by Emil Fischer who formulated the lock-and-key principle, which remains significant until today although with slight adjustments [1, 2]. The formation of ligand-protein complexes is a dynamic event. While shallow binding sites often display comparatively simple molecular recognition, the access to buried binding pockets, as they for example occur in cytochrome P450 enzymes (CYPs) or nuclear receptors (NRs), is a complex event with high relevance for binding kinetics and specificity [3, 4]. The consideration of target specificity is a crucial aspect for the successful design of drugs with acceptable, or in the best case, no side effects at all. Correspondingly, binding to anti-targets is one of the most relevant reasons for economically devastating drug attrition in clinical development [5, 6].

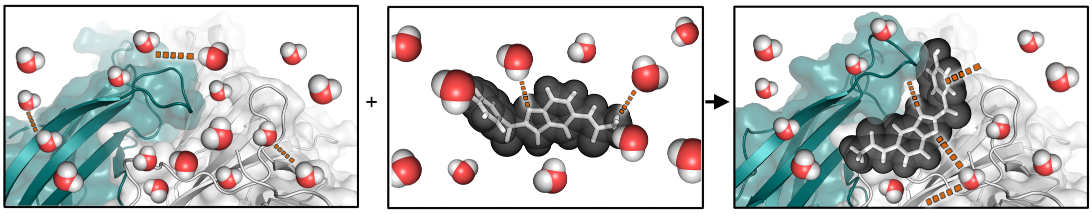


**Figure 1** Schematic depiction of several processes covered by molecular recognition including desolvation, water-mediated interactions, conformational adaptation of ligand and protein, and ligand-protein interactions.

Hence, the detailed understanding of molecular recognition is of pivotal importance. In this regard, computational methods provide a cost effective approach to study the underlying phenomena [7]. This thesis addresses various aspects of ligand recognition in the framework of drug-metabolizing enzymes and nuclear receptors including ligand-protein association, allosteric modulation and communication, specificity, as well as ligand-induced conformational adaptation.

## 1.2 Pharmacological Background

### 1.2.1 Drug Metabolism

A large share of clinical candidates fail due to a poor pharmacokinetic profile, which is closely related to the metabolic degradation of a compound, causing large economic damage and, potentially, leaves patients with reduced treatment options [8, 9]. Hence, there remains a challenge for drug discovery scientists to rationally design compounds with optimal properties in regards to their biotransformation. Generally, after a drug is ingested orally, it is subject to several physiological barriers before it can reach circulation and, ultimately, interact with its therapeutic target. Initially, it needs to undergo dissolution and permeate lipid bilayers in the gastrointestinal tract (GIT) without being too insoluble in water. After being absorbed from the GIT to the portal vein, oral therapeutics pass the liver, which is the primary organ for drug metabolism. Almost all drugs are subject to a process called first-pass metabolism, where the molecules pass the hepatocytes in the liver, which are rich in metabolic enzymes. Metabolic reactions, directed to facilitate the excretion of the compound from our body, can be divided into phase I and II depending on the involved enzymes and catalyzed reactions (Figure 2). Phase I reactions of small-molecules are mainly performed by cytochrome P450 enzymes (CYPs) in the liver along with smaller contributions by esterases and enzymes present in enterocytes. Phase II metabolism is targeted toward the conjugation of a molecule to a hydrophilic moiety such as glucuronic acid or glutathione to increase its water solubility and facilitate renal excretion. While the primary function of drug metabolism is to facilitate the excretion of potentially harmful compounds, some drugs are transformed into their active principle by metabolic enzymes. Furthermore, there are cases in which the metabolic transformation results in products with increased toxicity. Nevertheless, after passing the liver for the first time, the remaining portion of the original drug as well as the resulting metabolites are distributed through the circulation and can, ultimately, reach their designated target [10, 11, 12].

### 1.2.2 Cytochrome P450 Enzymes

As mentioned above, CYPs are responsible for the majority of phase I drug metabolism. The most relevant enzymes include CYP1A1, CYP2A6, CYP2B6, CYP2C8, CYP2C9,

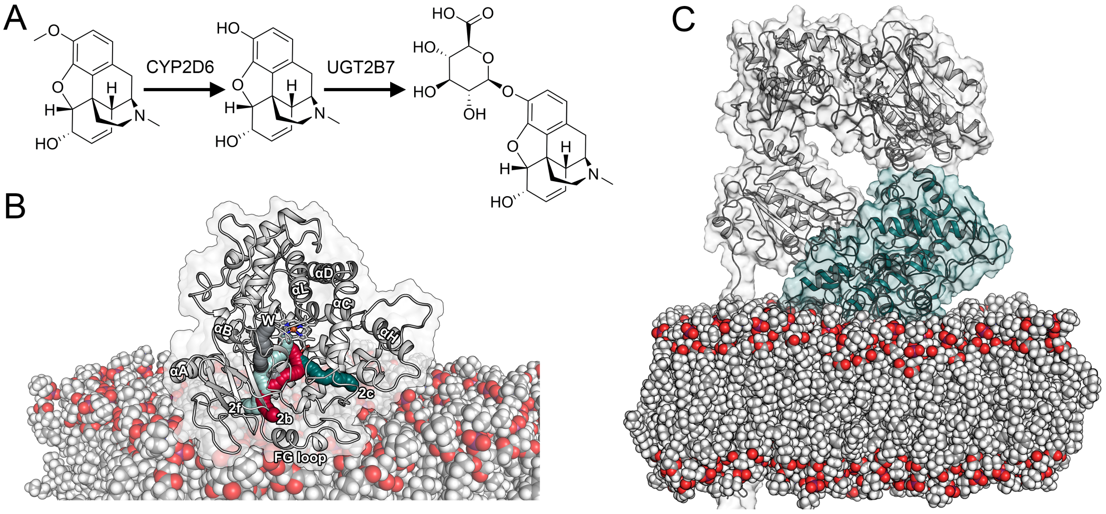


**Figure 2** Drug metabolism. (A) Illustration of the interplay between phase I and II metabolism by the example of codeine. The first reaction catalyzed by Cytochrome P450 2D6 (CYP2D6) is an O-demethylation of the aromatic methoxy group. In a next step, a glucuronic acid moietiy is conjugated to the free hydroxy group by UDP-glucuronosyltransferase 2B7 (UGT2B7), increasing the water solubility of the compound to allow renal elimination. (B) Depiction of CYP2D6 embedded in a membrane with several secondary structure elements and ligand tunnels indicated. (C) Complex of CYP2D6 in complex with its redox partner Cytochrome P450 reductase embedded in a 1-palmitoyl-2- oleoylphosphatidylcholine membrane.

CYP2C19, CYP2D6, CYP2E1, and CYP3A4, covering approximately 90% of drug metabolism [12]. Reactions catalyzed by CYPs include oxidation, hydroxylation, N-desalkylation, O-desalkylation, and desamidation [10]. Thereby, the reactions follow a conserved mechanism in which the substrate is oxidized by the introduction of an oxygen atom that is activated by the heme prosthetic moiety in the active site. The electrons for the reaction are transferred from their redox partner Cytochrome P450 reductase (Figure 2). Each enzyme exhibits a distinct substrate specificity with, for example, CYP2D6 preferably accepting lipophilic bases with a hydrophobic ring and a nitrogen atom that can be protonated under physiological conditions [10, 11, 12, 13, 14, 15]. The specificity of CYPs is highly relevant as drug-drug interactions, caused by enzyme inhibition or two molecules being metabolized by the same CYP, can alter the metabolic clearance and result in severe adverse effects. In analogy, interindividual differences in metabolic performance due to genetic polymorphism of CYP1A2, CYP2C9, CYP2C19, and CYP2D6 can significantly alter plasma levels of drugs. For example, there are allelic variants of CYP2D6 that can lead to a more than 5-fold increase in the metabolic transformation of a drug compared to the wild-type enzyme [10, 11, 16]. While their intricate substrate specificity can be be partially deduced from structural differences in the active sites of CYPs, other structural features such as differences along the selected route of ligands to access the binding site have been evidenced to influence it [17, 18, 19, 20]. The active site of CYPs is buried within the core of the enzyme and is connected by tunnels to the surrounding environment. Single amino acid mutations along these tunnels have been shown to influence the binding kinetics of different ligands and, hence, contribute to specificity of CYPs. Until today, experimental techniques have failed to provide atomistic detail into complex ligand access and egress phenomena. Computational methods, on the other hand, can provide detailed insight into associated conformational changes and dynamic events as elaborated in the following sections and chapters [18, 20, 21, 22, 23, 24].

In contrast to prokaryotic CYPs, their mammalian counterparts are anchored to the membrane of the endoplasmic reticulum. In addition to a helical transmembrane anchor, their globular domain is partially immersed in the membrane lipids with an insertion depth depending on the respective isoform. Due to the lipophilicity of a large share of CYP substrates it was postulated that hydrophobic ligands primarily enter the active site through tunnels protruding into the membrane while hydrophilic products prefer solvent-exposed tunnels [18, 20, 21, 22, 23, 24]. Ultimately, the complex machinery of these flexible enzymes including their genetic polymorphism, ligand tunnels, conformational changes, and differences in their active sites have consequences on drug design as well as pharmacotherapy in general. Its complete understanding as well as cost-effective ways to make predictions for rational design are a highly relevant scientific topic. As it is detailed in this thesis, computational methods can fill several knowledge gaps and contribute to safe and efficacious therapeutics.

### 1.2.3 Human Carboxylesterases

Besides CYPs, hydrolytic enzymes including human carboxylesterases (hCE), among other minor types of esterases such as acetylcholinesterase and butyrylcholinesterase, are of major relevance for phase-I drug metabolism of drugs containing amides, esters, and related functional groups [25, 26]. The esterification of compounds featuring a free carboxylic acid or alcohol function is a frequently exploited technique to overcome limitations such as poor bioavailability due to limited passive transport. The resulting molecule is referred to as a prodrug and is, in the optimal case, released in circulation by esterases in the blood plasma. The primary enzymes hydrolyzing ester-containing drugs in humans are hCE-1 and hCE-2 with well-established substrates such as angiotensin-converting enzyme inhibitors, $\beta$-blockers, and cholesterol-lowering drugs [26, 27, 28]. Furthermore, the same enzymes are also involved in the elimination of compounds and responsible for endogenous processes such as lipid homeostasis [26, 29]. Interestingly, the expression pattern of hCE enzymes is highly different, with hCE-1 primarily expressed systemically and hCE-2 mostly limited to the intestine. Hence, to achieve controlled release of the active principle, the selectivity of a prodrug for hCE enzymes is of pivotal importance, as its premature hydrolysis in the GIT would render the esterification approach obsolete [26, 30]. According to the literature, the substrate specificity of hCE-1 and hCE-2 largely depends on the size of the acyl and alcohol moieties of potential ligands (Figure 3A). While hCE-1 seems to prefer compounds with a small alcohol moiety such as methylesters or ethylesters, hCE-2 primarily cleaves compounds with a small acyl moiety, likely due to differences in their active sites and resulting steric limitations [26, 27, 30, 31, 32]. The main structural difference between hCE-1 and hCE-2 is a missing loop close to the active site in hCE-2 (Figure 3B). In addition to their role as drug-metabolizing enzymes and their endogenous functions, hCEs have been implicated as drug target for hypertriglyceridemia and diabetes [27, 33, 34].

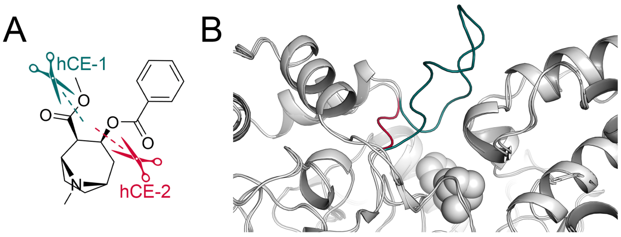


**Figure 3** Comparison of hCE-1 and hCE-2. (A) The rationale for the substrate specificity of hCE-1 and hCE-2 based on the size of the resulting hydrolysis products is shown with the structure of cocaine as an example. (B) The main structural differences between hCE-1 and hCE-2 by comparing a crystal structure of hCE-1 to a modeled structure of hCE-2.

The experimental determination of the hCE selectivity requires the use of recombinant proteins and comes with high costs. Computational methods, on the other hand, can serve as predictive tools to estimate which enzyme is responsible for the hydrolysis of a functional group [31, 35, 36]. Moreover, the prediction of hCE metabolism is important to avoid potential drug-drug interactions, similar to the situation for CYPs [26, 35]. In total, approximately 10% of marketed drugs follow a prodrug principle of which around half of them are activated by hydrolysis, rendering the estimation of the substrate specificity of hCEs an important task in the field of predictive metabolism. In Chapter 4 of this thesis, a predictive tool relying on machine learning algorithms is introduced to address this challenge.

### 1.2.4 Nuclear Receptors

Nuclear receptors (NRs) are a large family of ligand-inducible transcription factors, meaning that the interaction with a small-molecule leads to the direct regulation of gene transcription. They are involved in highly important physiological processes including cell proliferation, development, immunity, metabolism, and reproduction. Moreover, they have been implicated in several diseases such as hormone-dependent cancers, diabetes, and obesity [4, 37, 38, 39]. Thus, a large share of NRs have been implicated as drugs targets, such as the androgen receptor (AR) in prostate cancer or the estrogen receptors (ERs) in breast cancer, leading to several NR-targeting therapeutics available on the market today [40, 41, 42, 43]. However, as cancer cells have a comparatively unstable genome, they can acquire mutations leading to drug resistance, such as distinct changes in binding site of the AR [44, 45, 46]. Therefore, there remains an unmet need for novel therapeutics that circumvent the resistance mechanisms developed by cancers.

NRs share a common structural architecture consisting of three domains: the N-terminal domain, which is highly variable in different receptors, the DNA-binding domain mediating interactions with the DNA, and the ligand-binding domain (LBD) responsible for their regulation with small-molecules. While all domains have been considered as targets by drug discovery scientists [47, 48], the LBD is the primary target for pharmacological intervention. The LBD exhibits a common fold among hormonal NRs (Figure 4A). Similar to CYPs, the orthosteric binding pocket for small-molecules in NRs is buried within the core of the protein. However, in NRs they are referred to as pathways instead of tunnels [49, 50]. Using computational methods, experimental or clinical findings regarding the binding kinetics could be revealed in atomistic detail, underlining their suitability as predictive tools in this context [49]. Furthermore, the transport of ligands can be influenced by auxiliary proteins delivering lipophilic compounds to the soluble receptors, such as CRABP2 for the retinoic acid receptor [51].

In addition to their orthosteric binding site located in the core of the LBD, superficial allosteric sites denoted as activation function-2 (AF-2) and binding function-3 (BF-3) (Figure 4B) have been implicated as drug targets, especially for the treatment of castrate-resistant prostate cancer. At this stage of the disease, structural changes such as amino acid mutations in the binding pocket or alternative splicing can render classical antiandrogens obsolete [52, 53, 54]. Another incentive to target alternative binding sites in NRs is the lack of selectivity of antagonists for hormonal receptors due to the common steroid scaffold of their natural ligands. The AF-2 site corresponds to a protein-protein interaction surface responsible for the binding of coactivator proteins necessary for downstream signaling (Figure 4C). On the other hand, the BF-3 site has been implicated in the allosteric regulation of the protein-protein interactions at the AF-2 site as well as in the interaction with auxiliary proteins such as chaperones. Interestingly, most efforts toward allosteric NR antagonists were initially based on computational design. Compounds binding to the allosteric sites of AR, ER$\alpha$, thyroid receptors, as well as the glucocorticoid receptor have been reported. However, this class of compounds has not yet reached clinical application to this date [52, 55, 56, 57, 58, 59, 60].

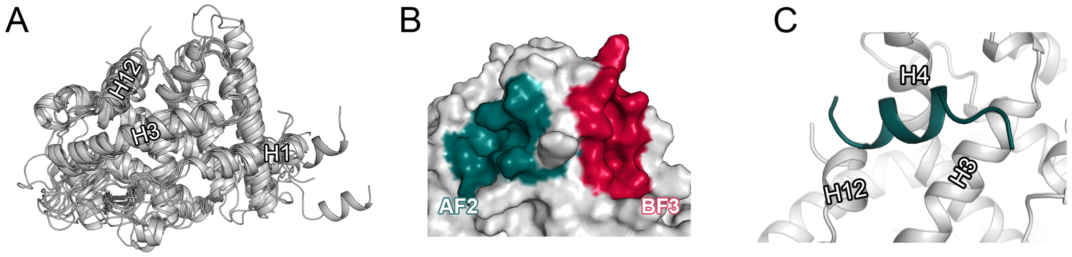


**Figure 4** Structural aspects of NRs. (A) Structural superposition of the androgen receptor, estrogen receptors $\alpha$ and $\beta$, the glucocorticoid receptor, mineralocorticoid receptor, progesterone receptor, as well as the thyroid receptors $\alpha$ and $\beta$ with three helices specifically highlighted. (B) The AF-2 and BF-3 allosteric sites depicted in two different colors with surface representation. (C) Fragment of a coactivator protein bound to the AF-2 site of the androgen receptor.

## 1.3 Molecular Modeling

### 1.3.1 The basis of molecular modeling in drug discovery

Computational chemistry, molecular modeling, computer-aided drug design, and cheminformatics are terms used for computational techniques applied in academia as well as the chemical or pharmaceutical industry to conduct research and provide novel insights. The boundary between the individual fields is sometimes ambiguous. While computational chemistry generally focuses on molecular mechanics and quantum-chemical calculations, cheminformatics primarily deals with the generation and interpretation of chemical data using machine learning and data science techniques, including the correlation of structure and activity [61, 62]. Molecular modeling encompasses theoretical or computational methods that provide a simplified or idealized description of the behavior of molecules [63]. Lastly, the term computer-aided drug design (CADD) covers all computational approaches applied to drug discovery and development [64]. Methods in CADD can be generally classified as ligand-based drug design (LBDD) and structure-based drug design (SBDD). Again, many techniques in LBDD are related to cheminformatics such as similarity searching and quantitative-structure activity relationships (QSAR) [61, 65]. While structural information from laboratory experiments can be highly useful in LBDD, especially in regards to crystal structures of small organic compounds, SBDD is strongly dependent on the availability of experimentally determined structures of the respective biological target. Among the most prominent approaches leveraging structural information of macromolecules are molecular docking applied to structure-based virtual screening and molecular dynamics simulations, modeling processes related to molecular recognition [66, 67].

As already mentioned, SBDD methodology depends on experimental structure determination by structural biologists. The most prominent method to obtain structures of proteins is X-ray crystallography contributing to over 150000 entries of the Protein Data Bank (PDB) [68, 69], which is the primary database for macromolecular structures. In principle, the method involves the crystallization of purified protein molecules, resulting in crystals which diffract X-ray beams producing a unique diffraction pattern that can be detected. Using a mathematical method termed Fourier transformation the obtained data can be transformed into an electron density map, which is ultimately fitted to the atomic structure of the protein using computational algorithms [70, 71] (Figure 5). In structures derived from X-ray crystallography, hydrogen atoms are generally not visible as they only have one electron. During the pre-processing steps for the application of SBDD methodology, hydrogen atoms are usually generated by computer algorithms [72].

The second largest group of deposited structures in the PDB comes from nuclear magnetic resonance (NMR) spectroscopy of liquid samples [68, 69, 70, 73]. In this technique, the sample is exposed to a magnetic field and the atomic nuclei are excited with radio waves. In response, the local magnetic field around the nuclei changes depending on their electronic environment, allowing to obtain structural information. Inter-atomic distances of up to 6 Å are derived from Nuclear Overhauser Effect experiments, relying on the transverse magnetization between nuclei. Longer distances, on the other hand, can be determined with paramagnetic relaxation. By the use of J-couplings, one can obtain information about dihedral angles. For protein structure determination, the resulting empirical parameters are computationally interpreted as restraints to represent the structure of a protein [73, 74].

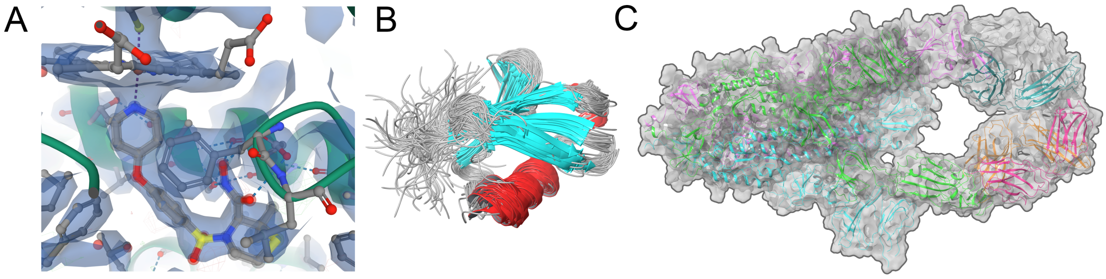


**Figure 5** Visualization of protein structures determined by the three most common structure elucidation methods. (A) Electron density map of the active site of CYP2D6 bound to the inhibitor prinomastat (PDB ID: 3TDA). (B) Structural ensemble explaining the recognition dynamics of ubiquitin captured by protein NMR spectroscopy (PDB ID: 2K39). (C) Cryo-EM structure of the spike glycoprotein of the severe acute respiratory syndrome coronavirus-2 (PDB ID: 7CZQ) with a total of 1283 amino acids.

The third largest group of over 8000 structures are derived from cryo-electron microscopy (cryo-EM), representing the most recent technique applied to protein structure determination [68, 70]. In cryo-EM, electrons focused by a magnetic field are projected at the sample, which interact with the Coulomb potential of each atom, resulting in a diffraction pattern that can be detected and interpreted. In addition, a magnified image of the molecules can be directly registered at the back-focal plane of the transmission electron microscope. Importantly, cryo-EM can be used to analyze non-crystalline samples. This circumvents one of the bottlenecks of X-ray crystallography, as it can be challenging to obtain well-ordered crystals for certain proteins and macromolecular assemblies [70, 71].

In conclusion, the availability of protein structures derived from X-ray crystallography, NMR spectroscopy, or cryo-EM form the inherent foundation of SBDD. However, the number of available structures is comparatively low regarding the approximately 106 million of non-redundant of protein sequences annotated in the National Center for Biotechnology Information database. In the case of a lacking 3D structure of a particular target, computational modeling can fill the gap [75]. In a technique referred to as homology modeling, a template structure of a homologous protein is used to derive a model for the target protein. The technique relies on the fact that the protein sequence determines its three-dimensional structure and that their structure is conserved despite differences in the sequence [75, 76]. In particular, a sequence similarity of at least 25% is recommended between target and template. Essentially, the first step is to align the sequences of target and template, followed by the building of a rough model of the protein backbone. In the next steps, the structure is iteratively refined by loop modeling, side-chain addition, and general optimization [75, 76]. While classical homology modeling remained the gold standard in protein structure prediction for many years, advancements in artificial intelligence methods fueled by improvements in computational architecture have gained a lot of attention. In general, they allow to obtain contact maps between protein residues and, thus, to predict the spatial proximity of individual amino acids. Only recently, an algorithm termed AlphaFold was introduced that, despite the lack of a homologous structure, allowed to predict models with atomic accuracy [76, 77]. The combination of template-based homology modeling together with contact maps derived from deep learning will likely lead to further improvements in protein structure prediction and, therefore, will allow to address novel therapeutic targets using computational techniques [76].

### 1.3.2 The most relevant interactions of drugs with proteins

In the course of this thesis, various types of ligand-protein interactions will be mentioned, which will be briefly addressed in this paragraph. Ligand-protein interactions constitute the basis of molecular recognition of drug compounds and are a fundamental component to be addressed in SBDD. They are of major importance for the estimation of the binding strength of a compound, which is one of the central tasks in computer-assisted drug design. According to the Gibbs equation, the binding free energy of a ligand depends on enthalpic and entropic terms. The thermodynamic profile of different ligands may show high variation, can be driven by either a gain in enthalpy or entropy, and strongly depends on the thermodynamic cycle of a ligand binding event [78, 79, 80, 81].

The most abundant intermolecular interaction is the van der Waals force based on the electrostatic attraction (or repulsion) between permanent and induced dipoles (Figure 6A). In particular, they are considered as a combination of London dispersion forces, Debye forces, and Keesom forces. London dispersion forces arise from non-polar atoms due to temporary fluctuations of their electron cloud and the resulting altered charge distribution. This temporary dipole induces a redistribution of the charge distribution of neighboring molecules, leading to electrostatic interactions between them. Keensom forces result from the electrostatic interaction of permanent dipoles. Lastly, Debye forces are based on a temporary dipole induced by a permanent one. Even though van der Waals forces are comparatively weak, they are additive and, thus, contribute significantly to molecular recognition [82].

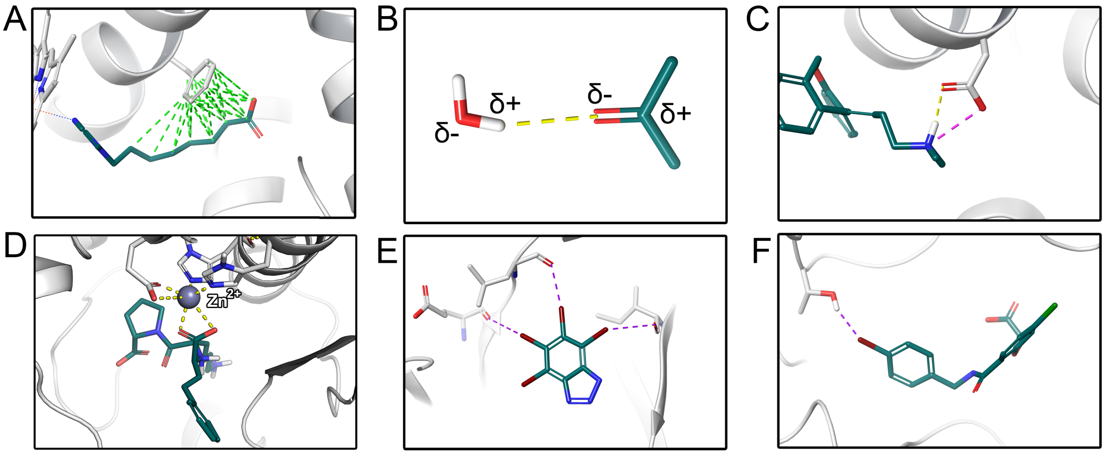


**Figure 6** Molecular interactions. (A) Van der Waals interactions between $\omega$-imidazolyl octanoic acid bound to the active site of CYP2E1 (PDB ID: 3KOH). (B) Schematic depiction of a hydrogen bond between a water molecule and a carbonyl group with partial charges indicated. (C) Charge-assisted hydrogen bond of doxepin in complex with the human histamine H1 receptor (PDB ID: 3RZE). (D) Ligandmetal interaction of lisinopril with angiotensin-converting enzyme (PDB ID: 1O86). (E) 4,5,6,7-Tetrabromobenzotriazole acting as halogen bond donor in complex with cyclclin A2 (PDB ID: 1P5E). (F) 2-(4-Bromobenzyl)carbamoyl-5-chlorophenoxy acetic acid acting as halogen bond acceptor in complex with aldole reductase (PDB ID: 4LAU).

Hydrogen bonds are attractive interactions between a hydrogen atom in a polarized bond and a neighboring electronegative atom (Figure 6B). They arise from electrostatic forces between the so-called donor (containing the hydrogen atom) and the acceptor resulting from a permanent multipole and, in contrast to van der Waals interactions, present strict geometrical preferences depending on the involved atoms. The angle between donor and acceptor in hydrogen bonds is generally greater than 150° along the direction of the free electron pair of the acceptor atom. When distances of hydrogen bonds consisting of nitrogen and oxygen atoms were measured in crystal structures, the distribution of the median lengths were 2.9 and 2.8 Å for the donors N-H and O-H, respectively. Due to their geometry and the resulting directionality, they are important for the specificity of ligand-protein association [78, 83]. In general, electrostatic interactions follow Coulomb's law and, therefore, decrease inversely proportional to the distance between the charges [63]. As the force is stronger when the point charges are large, hydrogen bonds are stronger if the donor and acceptor is charged (Figure 6C). In this situation, the interaction is referred to as charge-assisted hydrogen bond or salt bridge [63, 80, 84]. Furthermore, electronegative atoms or charged groups can interact with metal ions that are bound to the protein (Figure 6D), such as angiotensin-converting enzyme inhibitors which are clinically used to treat high blood pressure [80, 85].

In addition to conventional hydrogen bonds of moieties containing oxygen and nitrogen atoms, the halogen atoms bromine, chlorine, and iodine have unique electronic properties that allow for weak electrostatic interactions termed halogen bonds. Their electrostatic potential is anisotropic with a positive region at the tip of the halogen, referred to as $\sigma$-hole, and a negative region. Thus, whereas the $\sigma$-hole can act as a hydrogen bond donor along the axis of the covalent bond of the halogen, the negative region can act as an acceptor perpendicular to the covalent bond (Figures 6E and F). Halogen bonds are weaker and longer than typical hydrogen bonds, while their strength increases with the size of the halogen atom. On the other hand, fluorine atoms are less polarizable and more strongly electronegative resulting in the absence of a pronounced $\sigma$-hole [18, 78, 86].

It has been proposed, that the single best parameter for a correlation with binding affinity is the amount of hydrophobic ligand surface that is buried upon its association with a protein. If a non-polar molecule is surrounded by water molecules, the entropy decreases due to the ordering of the solvent around it. This can be avoided if multiple non-polar molecules aggregate resulting in an entropically favorable contribution, which was defined as the hydrophobic effect [78]. In addition to the entropic contribution, enthalpically favorable interactions involving $\pi$-electrons, hence termed $\pi$-interactions, of aromatic rings can occur due to their specific shape and electronic properties. In particular, the electron-rich $\pi$-system above and below of a benzene ring bears a negative partial charge, while the connected hydrogen atoms are electropositive [87, 88]. This leads to the possibility for aryl-aryl interactions as well as additional phenomena involving aromatic rings and other polarized moieties (Figures 7A-D). Between two aryl rings, three different configurations are known: (i.) a T-shaped interaction between a hydrogen atom and the negative charge of the $\pi$-system of orthogonal rings, (ii.) a parallel-displaced orientation following the same principle as a T-shaped one, and (iii.) a sandwich or face-to-face configuration with two parallel rings, relying on van der Waals forces between the atoms due to poor electrostatics resulting from the repulsion of the $\pi$- electron clouds [88, 89, 90]. Interactions with non-aryl components include cation-$\pi$ interactions, where a positively charged atom such as a basic nitrogen or an ion undergoes an electrostatic interaction with the partial negative charge of the $\pi$-system resulting in a strong attraction. Furthermore, alkaline metals can interact with $\pi$-systems in a similar fashion [87, 88].

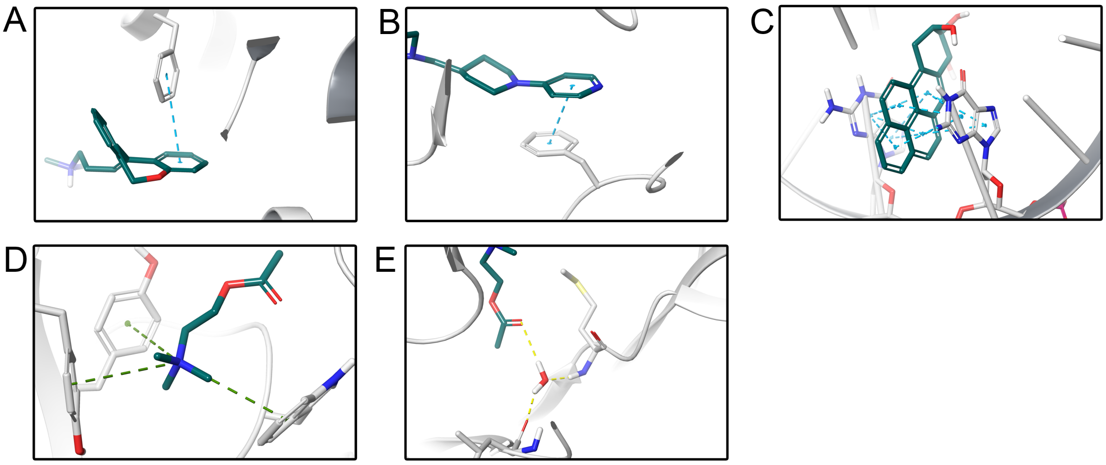


**Figure 7** Molecular interactions. (A) T-shaped $\pi$-interaction of the human histamine H1 receptor in complex with doxepin (PDB ID: 3RZE). (B) Parallel-displaced $\pi$-interaction between 4-[(6-Chloro-2-naphthalenyl)sulfonyl]-1-[[1-(4-pyridinyl)-4-piperidinyl]methyl]-2-piperazinecarboxylic acid ethyl ester and the human coagulation factor Xa (PDB ID: 1IQJ). (C) 1,2,3-Trihydroxy-1,2,3,4-tetrahydrobenzo[a]pyrene interacting with a DNA fragment through $\pi$-stacking interactions. (D) Cation-$\pi$ interaction between acetylcholine and the acetylcholine binding protein. (E) Water-mediated molecular interaction between acetylcholine and the acetylcholine binding protein (PDB ID: 3WIP)

Besides direct contacts, the interaction of a ligand and a protein may be mediated by water molecules (Figure 7E), especially in solvent-exposed binding sites [78, 91, 92]. Furthermore, the binding process of a ligand involves the desolvation of both the ligand and its binding site. Whether the release of a water molecule from a binding site is favorable regarding the ligand binding free energy inherently depends on enthalpic and entropic contributions. While it is considered that the release of water molecules from the binding site is entropically favorable, some of them might strongly interact with the binding pocket. For example, the removal of a water molecule from a charged functional group such as a carboxylate is accompanied by a high desolvation penalty, which consequently has to be compensated by strong ligand-protein interactions. Thus, detailed understanding of solvation thermodynamics is needed for the accurate prediction of the ligand binding free energy [92, 93, 94]. In Chapter 8 of this thesis, the characterization of hydration sites was considered to examine the efficacy of allosteric NR modulators.

### 1.3.3 Molecular docking

The technique referred to as molecular docking, which is one of the most frequently exploited approaches in CADD, relies on the above-mentioned key-and-lock principle describing molecular complementarity. Essentially, it predicts the interaction of a ligand with a target structure. In the classical case, the ligand represents a small-molecule, in analogy to the majority of drugs on the market. However, algorithms to predict the interaction of peptide, protein, or nucleic acid ligands have been developed as well [1, 2, 81, 95, 96, 97]. In the above-described overview of CADD methodology, it can be assigned to the SBDD techniques and, therefore, strongly relies on high-quality structural information of the target macromolecule. It is one of the most popular methods applied in drug discovery applicable to the screening of virtual libraries and the development of obtained hits into drug candidates, leading to several success stories in the past. [80, 81, 96, 98].

Generally, docking can be divided into two stages: a sampling stage probing the orientation of the ligand, also referred to as pose, within a predefined binding pocket followed by a scoring stage estimating its binding free energy. As modern algorithms for small-molecule docking treat the ligand as flexible entity, conformational sampling of the ligand is combined with its translational and rotational degrees of freedom. This results in a large number of possible solutions [81]. While it is often time-intensive, the sampling stage can be optimized, for example, by matching pharmacophores between the binding pocket and the ligand, limiting the number of orientations to the ones with a potential for favorable ligand-protein interactions [97]. Scoring of the individual orientations generated in the sampling stage is used to select the putative correct pose using a so-called scoring function. These mathematical representations can be roughly divided into empirical, force-field based, and knowledge-based scoring functions. Empirical scoring functions consist of the sum of various energy terms, such as the ones accounting for intermolecular interactions introduced in the previous section, weighted by coefficients. These coefficients are optimized to correlate with binding affinity data that is used during training. Force-field scoring functions follow the concept of a molecular mechanics force-field approximating the potential energy of a system. A force-field combines bonded and non-bonded terms, although scoring functions primarily address non-bonded terms describing intermolecular interactions. As force-fields are a central component of molecular dynamics (MD) simulations, they will be discussed in more detail in the following section. Lastly, knowledge-based scoring functions leverage databases by associating an energy component with frequently observed ligand-protein contacts to ultimately sum up the energy components for a given pose [81]. The validation of docking protocols prior to their application for SBDD is highly recommended, especially as the performance of an algorithm can depend on the target under investigation [80, 99, 100, 101]. The high number of available crystal structures for certain targets allows for an assessment of the pose prediction accuracy of docking protocols by comparing the predicted poses to the ones that were experimentally determined (Figure 8). Most frequently, the root-mean square deviation (RMSD) between the heavy atoms of the two molecules is used as a performance metric, although this approach comes with certain limitations. For example, the RMSD strongly depends on the number of atoms in a particular ligand, and consequently, is often higher for large ligands [100, 101, 102, 103]. As ligand-protein interactions are the central component of ligand recognition, it was suggested that a comparison of the intermolecular interactions in the experimental and docked pose is a more useful metric to be assessed [100]. In addition to the pose prediction accuracy, the ranking performance can be explored by the correlation of docking scores to binding affinities of known ligands [104]. Another common way to validate scoring is to compile a library of presumably inactive decoy compounds, optimally with physicochemical properties matching a set of known actives. From the resulting docking scores of the actives and decoys, a Receiver Operating Characteristic (ROC) curve can be computed by plotting the true positive rate of detecting an active against the false positive rate. The area under the curve (AUC) in such an ROC plot often serves as a quantitative measure describing the performance of a docking application [105, 106].

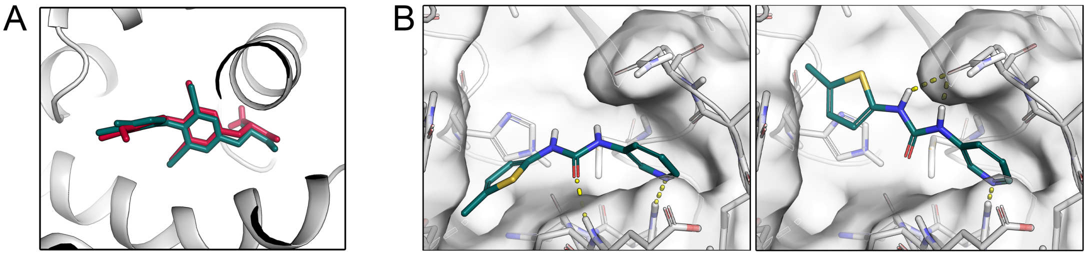


**Figure 8** Docking poses compared to experimentally determined binding modes in crystal structures. (A) Redocking of triiodothyronine in the orthosteric binding site of thyroid receptor $\alpha$. (B) 1-(5-methylthiophen-2-yl)-3-pyridin-3-ylurea bound to the main protease of SARS-CoV-2 (PDB entry 5RH0). The native pose is shown left, while a slightly different pose from docking is displayed at the right. The figure was adapted from a binding pose prediction challenge in a recent review [80].

While early expectations for the impact of docking were not completely fulfilled, current algorithms achieve a pose prediction accuracy of over 80% for a near-native pose among the top-5 [107]. However, the accurate scoring and ranking of the obtained orientations still remains a challenge due to the simplistic mathematical representation of the complex binding thermodynamics behind ligand-protein association. Due to these inaccuracies, it is important to keep in mind that even though the use of computers advanced the rational design and discovery of drug candidates, human intervention is often required to post-process results from docking by visualizing the generated complexes. While the analogy to poses of similar ligands in crystal structures, shape-complementarity, interactions with specific residues, and general electrostatics can be evaluated during this visual inspection, automatized techniques including interaction fingerprints, scaffold docking, or the automatic comparison of congeneric series can support the process. Furthermore, docking poses can be subjected to post-scoring methods based on molecular mechanics/generalized Born surface area (MM/GBSA), alternative scoring functions, alchemical binding free energy calculations, or machine learning algorithms [80].

Since the postulation of the key-and-lock principle, additional insight into molecular recognition could be obtained regarding the flexibility of binding events. Both the ligand and the protein can mutually adapt upon binding as it was shown in over 90% of structures in the PDB, for which apo and holo structures were available [80, 96, 108]. Thus, the key-and-lock principle was revised to the hand-and-glove principle by Koshland [109]. In order to address these phenomena, several flexible docking algorithms have been introduced, allowing either side chains to adapt, subject the complex to a molecular mechanics minimization, dock the ligand to an ensemble of structures, or post-process the pose with MD simulations [96, 97, 110, 111]. Many of these algorithms require intensive computation, limiting their applicability to the screening of smaller libraries as they are typical during lead optimization procedures [80].

### 1.3.4 MD simulations

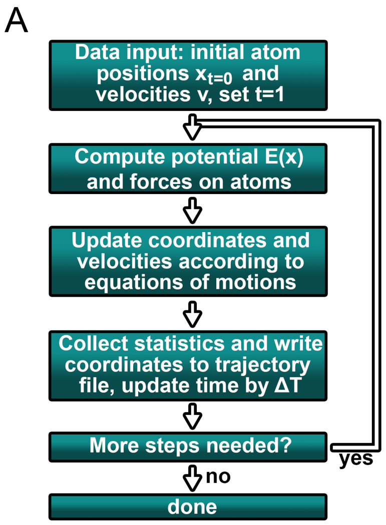
*"The living cell has as many macromolecules as the United States have citizens. And that is a very good comparison, because these molecules in the cell form a society. They assemble and they work together. And that is what life sciences and medicine tries to understand. The problem is, to look at all this detail, to see what all these molecules, citizens, do in the cell, there are many microscopes that have been tried - for example the famous light microscope, but it can not resolve the citizens. And it is a computer that is actually today finally permitting us to see the citizens at work. The citizens that, in some cases, just build pipes and, in some other cases, amazing machines that read the genetic information and turn it into new proteins of the cell or to harvest the sun*

*light to solve the energy problem of nature. So, this computational microscope is not made of glass and metal, but it is made of software."*

This is an inspiring quote of a talk by Klaus Schulten, one of the pioneers in MD simulations who sadly deceased in 2016, given at the University of Illinois in 2010. The computational microscope he was referring to in his talk is based on MD simulations of biological macromolecules [112].

As described in the previous section on molecular docking, molecules are flexible entities capable of changing their conformation rapidly. Although three-dimensional structures are able to support our understanding of these dynamic motions, obtaining them is costly and sometimes not possible. Thus, researchers have sought for alternative techniques to predict protein motions [113]. This led to the discovery of MD simulations, during which the input atomic coordinates of a molecular system are transformed to represent a point later in time (Figure 9A) based on classical mechanics described by equations I and II [113, 114]. Equation I corresponds to Newton's second law [63], while equation II descends from the arithmetic mean of the common acceleration in classical mechanics and is used to transform the atomic coordinates [63, 114, 115, 116]. The potential that enacts on each atom is computed by a molecular mechanics force-field considering bonded and non-bonded interactions (equations III to VII in Figure 9B). Among the bonded interactions, bond stretching ($E_{bond}$) and angle bending ($E_{angle}$) have the same functional form as they are described by harmonic potentials. Dihedral angles ($E_{torsion}$) are described by a series of cosine functions as shown in equation V. The non-bonded interactions include a term for electrostatic interactions described by the Coulomb law and a Lennard-Jones term quantifying van der Waals interactions [63, 114, 116]. Parameters of the force-field such as bond lengths are generated to fit experimental or quantum-mechanical data in a process denoted as force-field parameterization [113]. In general, an MD algorithm starts with the input of the atomic coordinates of the system, which are attributed a randomized initial velocity sampled from a Boltzmann distribution. In order to evaluate the statistical uncertainty of a simulation, it is common practice to conduct multiple runs with different seeds used to compute the random initial velocities [116, 117]. In a next step, the potential to translate the atoms is obtained by force field calculations, followed by the translation of the atoms and updating of the time by the selected time step as shown in Figure 9. In a certain time interval during the simulation, atomic coordinates are deposited resulting in a trajectory of the system that can be analyzed and visualized [116].

When attempting to simulate atomic motions, one has the consider the extremely short timescales of atomic events. The vibrations of bonds, for example, take place in the femtosecond (fs) timescale, meaning that the time steps that have to be applied to accurately simulate proteins need to align with these fast motions. While solvent-exposed side chains can rotate in the picosecond to nanosecond timescale, global conformational changes or protein folding may only take place within seconds [115, 116, 118]. Generally, the interval for the integration of the differential term in equation I (Figure 9B) is selected to be 1-2 fs based on a compromise between accuracy and efficiency. By using SHAKE algorithms, distance constraints are applied to the fastest degrees of freedom in a bio-molecular system such as vibrations of bonds to hydrogen atoms, ensuring their accurate representation [114, 119]. Importantly, the above-mentioned time step is referred to as the inner timestep, which is applied for the computation of the fastest components of a particular system. A second outer time step is used for slower components, resulting in a speed-up as the computationally demanding calculations of long-range forces have to be performed less frequently [120]. Additional algorithms such as particle-mesh Ewald apply a specific cutoff for electrostatic interactions after which the contributions are evaluated in Fourier space, further reducing the computation time [121].


A
Data input: initial atom positions $x_{t=0}$ and velocities v, set t=1
Compute potential E(x) and forces on atoms
Update coordinates and velocities according to equations of motions
Collect statistics and write coordinates to trajectory file, update time by ΔT
More steps needed?
yes
no
done

B

$$F(t) = ma(t) = -\frac{\partial E(x(t))}{\partial x(t)} \tag{I}$$

$$x(t+\delta t) = x(t) + v(t)\delta t + \frac{1}{2}\,a(t)\delta t^2 \tag{II}$$

$$E = E_{\text{bond}} + E_{\text{angle}} + E_{\text{torsion}} + E_{\text{non-bonded}} \tag{III}$$

$$E_{\text{bond}} = \sum_{bonds} K_r (r - r_0)^2 \tag{IV}$$

$$E_{\text{angle}} = \sum_{angles} K_\theta (\theta - \theta_0)^2 \tag{V}$$

$$E_{\text{torsion}} = \frac{V_1}{2}[1 + \cos(\Phi + \Phi_1)] + \frac{V_2}{2}[1 - \cos(2\Phi + \Phi_2)] + \frac{V_3}{2}[1 + \cos(3\Phi + \Phi_3)] \tag{VI}$$

$$E_{\text{non-bonded}} = \sum_i \sum_j \left[\frac{q_i q_j e^2}{r_{ij}} + 4\varepsilon\left(\frac{\sigma_{ij}{}^{12}}{r_{ij}{}^{12}} - \frac{\sigma_{ij}{}^{6}}{r_{ij}{}^{6}}\right)\right] f_{ij} \tag{VII}$$

**Figure 9** Force-field terms and MD algorithm. (A) A simplified scheme of an MD algorithm according to Lindahl [116]. (B) The fundamental equations behind an MD algorithm are presented. While equation I corresponds to Newton's second law, equation II illustrates how atomic coordinates are transformed according to classical mechanics. The force-field terms (equations III to VIII) are shown at the example of the OPLS force-field [122].

To accurately reproduce physical conditions of a microscopic system in which neither matter or energy are exchanged to its surroundings and reproduce its macroscopic behaviour, so-called thermodynamic ensembles were introduced [63, 114]. The isothermal-isobaric ensemble (NPT), keeps the number of particles (N), the pressure (P), and the temperature (T) constant facilitated by thermostat and barostat algorithms.

Thermostats ensure that the macroscopic temperature of a system remains constant by adequate adaptions to the equations of motion, while barostats appositely scale the system volume. The NPT ensemble reflects laboratory conditions most closely. Two other ensembles are the canonical ensemble (NVT) with constant volume and temperature as well as the microcanonical ensemble (NVE) with constant energy and volume. [63, 114, 123]. In bio-molecular simulations, the molecules under investigation are immersed in a box of solvent molecules. As ligand-protein association most frequently takes place in an aqueous environment, water is commonly used as solvent. To allow the explicit treatment of water, several water models have been developed to reproduce its behavior. Among the most common rigid water models are the transferable intermolecular potential 3P (TIP3P), simple point charge (SPC), and simple point charge extended (SPC/E) models. These models comprise of Lennard-Jones and Coulomb terms and possess three interaction points with point charges [114, 124]. In addition to water molecules, other biologically relevant constituents such as membranes can be added to the system to accurately reproduce physiological conditions. To attain bulk properties and a constant number of particles and allow for the application of electrostatic cutoff algorithms, periodic boundary conditions are employed, for which a central unit cell is replicated in all three dimensions, forming an infinite lattice of copies [114, 125, 126].

Following the description of the time steps, a high number of iterations have to be conducted to reach a timescale of bio-molecular relevance. If an interval of 2 fs is selected, a microsecond simulation requires half a billion steps. In addition, systems of ligand-protein complexes, as they are often studied during structure-based drug design can consist of more than a million atoms including the solvent. Thus, early applications of MD simulations were limited to small systems due to the demanding computational task at hand, despite the above-mentioned techniques to improve efficiency [127, 128]. However, due to recent developments in computer hardware, especially graphics processing units (GPUs), the timescale accessible to MD simulations has massively increased. Nowadays, it is possible to routinely conduct simulations with a duration of multiple microseconds [114, 129]. Furthermore, dedicated supercomputers, such as the one developed by D.E. Shaw Research, allow to capture events taking place in the second timescale with a peak performance of over 200 microseconds per day [130].

### 1.3.5 The application of MD simulations

As described in the previous section, MD simulations allow to incorporate flexibility into a molecular system by transforming its coordinates according to classical mechanics supported by a force-field evaluating the bonded and non-bonded interactions. Thus, as the treatment of flexibility is often limited in docking calculations, MD simulations can overcome this limitation, allowing to study the time-evolved stability of a docking pose and potential conformational rearrangements (Figure 1A and 1C)). This is one of their main applications in the field of SBDD [114].

A primary objective of nearly all drug discovery projects is the design of strongly binding (high affinity) ligands while maintaining favorable properties regarding toxicology and pharmacokinetics. Hence, the estimation of the binding affinity of a particular compound for a pharmaceutically relevant protein is one of the major tasks in CADD, as previously discussed. Docking scores obtained from scoring functions often present a low correlation with experimentally determined binding affinities due to the simplified description of the underlying thermodynamics [78, 80, 93, 114, 131, 132, 133]. As the accurate estimation of ligand-protein binding free energy requires the consideration of all components of molecular recognition, including description of conformational adaptation and flexibility as well as effects of the surrounding solvent, MD simulations are of particular interest for this task [78, 131]. In analogy to other techniques, the accuracy of methods to estimate binding free energies correlates with their computational expense. These methods can be generally divided into endpoint methods and pathway methods. Endpoint methods sample the ligand and protein in both the unbound and bound state, computing the energy difference between these states. These methods include linear interaction energy, molecular mechanics Poisson–Boltzmann surface area (MM/PBSA) molecular mechanics generalized born surface area (MM/GBSA) calculations [131, 134, 135]. Even though endpoint methods often show better performance than a scoring function [80, 136, 137], their missing consideration of conformational entropy and the simplification of water thermodynamics by implicit treatment can still affect their accuracy [134]. In Chapter 6 of this thesis, MM/GBSA calculations were employed to post-process MD simulations of a combination of ligands bound to the AF-2 and BF-3 allosteric sites of NRs with orthosteric antagonists. The methodology was used to estimate if a combination of these compounds results in a gain in binding free energy for allosteric inhibitors [55].

Pathway methods include steered MD (introduced below), umbrella sampling, thermodynamic integration, and free energy perturbations (FEP), of which the latter is considered the current gold standard [80, 135, 138]. In principle, FEP rely on specialized MD simulations during which the relative binding free energy between a set of similar or congeneric ligands is estimated. According to the thermodynamic cycle, the relative binding free energy between two ligands can be either obtained by comparing their free energies derived from (un-)binding simulations or by transforming them into each other in the bound and unbound state. While it is computationally highly intensive to simulate the former (as detailed below), the latter method is used in FEP. This means, that different pairs of ligands are compared by alchemically transforming them into each other, both bound to the protein and in solvation with explicit water molecules. The alchemical transformation, also called perturbation, is performed in so-called $\lambda$-windows in which one state is gradually transformed into another. Since the perturbation from one ligand to another needs to be relatively small to retain accuracy, the methodology is frequently applied in lead optimization, as congeneric series are common at this stage. To increase the accuracy and give an estimate for the statistical confidence of a result, a ligand is compared with multiple other ligands resulting in a closed cycle. Today, the FEP+ technology developed by Schrödinger can achieve an accuracy of roughly 1 kcal/mol due to recent advancements of force-fields, enhanced sampling techniques, computational capabilities, and a clever simulation setup. It is likely that further developments in hardware as well as the accuracy and efficiency of FEP calculations will increase the integration of such SBDD techniques in drug discovery workflows [134, 135, 138].

Interestingly, the development of GPU-accelerated simulations allowed to completely reproduce binding events of ligands from the solvated state of both ligand and protein to the formation of a ligand-protein complex [18, 139]. As such binding events usually take place in the microsecond timescale, they are considered rare molecular events and their investigation requires high computational efforts. To overcome this limitation, biased sampling techniques such as steered MD (SMD), random-accelerated MD (RAMD), accelerated MD, and metadynamics simulations have been developed. In these techniques, the potential derived from the force-field is adapted to encourage sampling of nearby regions on the potential energy surface. SMD simulations, in particular, apply an external force to the ligand in the form of a directional vector attached by a harmonic spring to steer the ligand in a predefined direction [49, 140, 141]. While RAMD simulations follow a similar approach, the direction of the force is not predefined, but randomly adapted if the ligand does not cover a certain distance along the trajectory [49, 142]. In accelerated MD simulations, a boost potential is added on top of the force-field, which was classically applied to the torsional term while more recent methods also modify the total potential. In this way, the energy surface is sampled more extensively by encouraging the system to escape local minima [143, 144]. Lastly, metadynamics simulations are based on the definition of collective variables (CVs) describing the reaction coordinate of a particular process within a molecular system. In this approach, the normal evolution of a system is biased by a history-dependent potential defined by a Gaussian centered on the trajectory followed by the CVs. The free energy surface of a particular process can be iteratively reconstructed based on the sum of the Gaussian potentials deposited during the process. Due to the possibility to sample rare molecular events in combination with the estimate for the free energy associated to a particular process, metadynamics simulations are widely applied to study biophysical phenomena [131, 145, 146]. In the course of this thesis, metadynamics simulations were applied to study the dissociation of ligands from a particular binding site (Chapters 3 and 6). In this context, the CV describing the reactions coordinate of the system was selected as the distance between the ligand and atoms in the binding site based on geometrical centroids. Furthermore, the progress of GPU-accelerated MD algorithms allowed to employ conventional unbiased MD simulations to study the complete association process of CYP2D6 substrates. As amino acid mutations resulting from genetic polymorphism can influence the metabolic efficiency of CYP2D6, their influence on this association process could be highlighted. The use of unbiased simulations can serve as a blueprint to evaluate outcomes from biased sampling techniques [18].

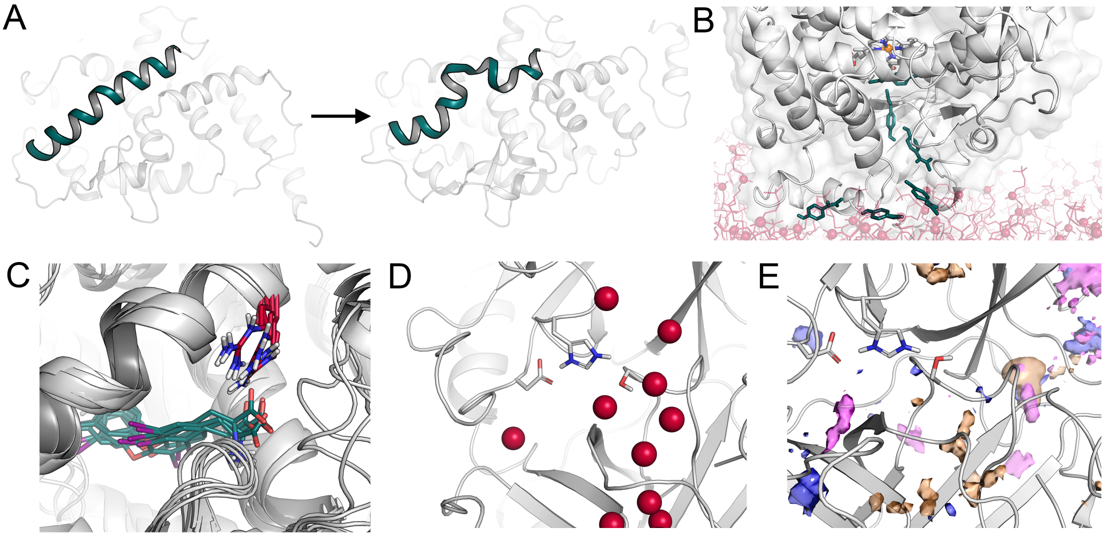


**Figure 10** Applications of MD simulations. (A) Distortion of helix 3 in thyroid receptor $\alpha$ induced by the antagonist dronedarone revealed by MD simulations [147]. (B) The recognition process of acetaminophen to CYP2D6 in unbiased MD simulations. (C) Structural ensemble of triiodothyronine bound to thyroid receptor $\alpha$ obtained from an MD simulations. In addition, a protein residue involved in a salt bridge is depicted. (D) Hydration site detected in the active site of thrombin [148]. (E) Density map obtained from a cosolvent MD simulation of prostasin [148].

Another application of MD simulations is the detection and profiling of hydration sites of a protein, often focused on its binding site. As described in previous sections, hydration phenomena are of keen relevance for ligand recognition and, therefore, the examination of the thermodynamic properties of water molecules solvating binding sites can support the design of potent and selective compounds [78, 92, 114]. Although experimental structure elucidation can be useful to locate hydration sites [149], the approach suffers from two limitations: several structures with low resolution are necessary and water molecules are often not visible in structures derived from X-ray crystallography. Since MD simulations include explicit water, methods were developed monitoring the location of water molecules by processing MD trajectories [92, 133, 150, 151]. A common technique is to subject the atomic coordinates of water molecules in an MD trajectory to a clustering algorithm such as 3D-DBSCAN resulting in a map of hydration hotspots with a certain occupancy [55, 133, 151]. In a next step the enthalpy and entropy are estimated to compute the binding free energy of a particular hydration site. The enthalpic contribution can be derived from the difference of electrostatic and van der Waals interactions in the hydration site compared to bulk solvent. The entropic contribution of removing the water molecule from the binding site can be computed by considering its degrees of freedom from a probability density function in the hydration site and the bulk solvent [133, 152] The incorporation of hydration data into scoring functions can result in a higher percentage of successfully ranked ligands. Recently, methods emerged to obtain the same data from a single static structure by training deep neural networks with hydration data from MD simulations [91].

Similar to the characterization of hydration sites, cosolvent MD simulations can be used to monitor the association of small organic probe molecules with the target of interest. The organic probes are selected to cover fragments of potential small-molecular ligands for the protein. Common probes include acetonitrile, isopropanol, and pyridine due to their low potential for aggregation. In contrast to adding multiple complete copies of a ligand, fragments can exchange more rapidly with binding sites resulting in less computational effort for binding site detection. The methodology originates from a technique in structure elucidation termed multiple-solvent crystal structure method where organic solvents are added to the crystallization solution to detect binding sites [153, 154]. In addition to the detection of binding sites of the protein, the obtained density maps from cosolvent simulations can be leveraged to design novel ligands for the protein by considering the densities as pharmacophoric features. Furthermore, these features can be incorporated into pharmacophore-based screening methodology to identify hits from chemical libraries [154, 155]. Another application of cosolvent simulations is the detection of cryptic binding pockets of targets with a low druggability which often involve the competition of small-molecules with protein-protein interaction sites. Cryptic binding pockets are typically occluded and only become apparent in the context of an interaction partner, which makes the method suitable for their identification and characterization [156, 157, 158]. Even though alternative methods were developed independent of MD simulations, the intrinsic treatment of flexibility and explicit solvation allows for conformational adaptations and reduces the dependence on high quality input structures [55, 154, 159]. In the course of this thesis the combination of hydration site data and density maps from cosolvent simulations was used to compare the AF-2 and BF-3 allosteric sites of several hormone-binding NRs. In this context, an algorithm for the hydration site detection in crystal structures of NRs was developed by the use of the 3D-DBSCAN clustering method [55] and combined with an MD-based technique [92]. The obtained data was used to systematically evaluate the druggability of the sites as well as the efficacy and selectivity of known binders.

### 1.3.6 Cheminformatics and machine learning

The area of cheminformatics emerged due to the large amount of chemical data and the need for representation of chemical information for computer processing. First introduced by Frank Brown in 1998 [61, 160], it was defined as follows:

*"The use of information technology and management has become a critical part of the drug discovery process. Chemoinformatics is the mixing of those information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and organization."*

Thus, cheminformatics deals with the representation of compounds in a format suitable for computational processing, the computation of chemical properties, and the association of properties with experimental outcomes, among other applications. One of the best-known cheminformatics methods is the development of QSAR models discerning mathematical relationships between molecular properties and activity to ultimately extrapolate them to predict novel compounds. In a similar fashion, quantitative structure-property relationships (QPSR) models can be derived. In this regard, the term machine learning should be contextualized, as the training and application of QSAR and QSPR models fall within this definition [61, 161]. Machine learning is considered a subset of artificial intelligence, which can be defined as human intelligence exhibited by machines. In principle, machine learning algorithms learn from experience and improve their performance during the learning process. Such algorithms can be applied to a variety of problems and are not limited to drug discovery. For example, a common dataset used for teaching purposes is the so-called Iris dataset with features of flowers such as their petal length and width, which can be used as metrics for classifying different species of flowers [162].

In general, machine learning can be roughly divided into supervised and unsupervised learning approaches. While the former methods rely on predefined labels that are associated with the training data, such as a class in classification tasks or a numerical value in regression, the latter describes methods that autonomously learn patterns directly from unlabeled data including clustering and dimensionality reduction. Among the most common methods in the field of supervised learning are random forest, support vector machine (SVM), k-nearest neighbors (k-NN), linear discriminant analysis (LDA), logistic regression, and artificial neural networks. If an artificial neural network has more than one hidden layer, it is also called a deep neural network and can be assigned to the field of deep learning. Random forest is a so-called ensemble method as it includes multiple decision trees and employs majority voting, meaning that the majority of trees decides the result in a classification task and their average is used for regression [161, 163, 164]. The randomness of the random forest method comes from bootstrapping where only a random subset of the dataset is used for each new tree as well as from the fact that only a random subset of features are considered at the nodes of a tree [163, 164]. SVM algorithms map features such as molecular descriptors in a multi-dimensional feature space and attempt to determine a hyperplane separating the inputs based on their respective labels. When the k-NN method is applied to classification tasks, it relies on the principle that a compound that is part of a particular class can be defined on its neighboring (most similar) compounds. The method considers weighted similarities between an object and its nearest neighbors, while the ”k” in k-NN comes from an integer number, that defines the number of neighbors that is considered [161, 164]. LDA determines a linear combination of features which separates two or more classes by establishing hyperplanes within the dataset maximizing the separation of categories [165, 166]. Regression methods are closely related to LDA and can be divided into linear and logistic regression. While linear regression is used for continuous data, logistic regression is applied to categorical data. Both methods establish a linear relationship for a given set of training points that can be directly used to predict outcomes in new data [161].

In the context of machine learning, principles for best practice were established due a number of studies in which models have been over-interpreted without appropriate validation. Among these principles are the curation of the dataset, the use of internal cross-validation as well as external validation sets, label randomization to exclude chance predictions, and the definition of an applicability domain. Furthermore, the combination of multiple machine learning techniques (combi-QSAR), allowing to detect a consensus among them, is recommended. Lastly, the descriptors that are used to characterize the compounds and serve as an input for the model should be non-redundant and physicochemically relevant [167, 168, 169].

Importantly, the field of cheminformatics is not limited to QSAR/QSPR predictions. For example, in order to compute molecular descriptors, compounds need to be represented in a machine-readable format [61]. A compound can be characterized either by its name, a two-dimensional drawing, or by its three-dimensional atomic coordinates (Figure 11). One of the most simple and widely used representations in cheminformatics is a linear notation termed Simplified Molecular Input Line Entry Specification (SMILES), which is human readable and represents a unique chemical structure including stereochemistry as well as charges [61, 170]. The most commonly used way to represent a chemical structure among different disciplines is its structure diagram (2D structure) which, from a mathematical perspective, can be interpreted as a molecular graph with nodes (atoms) and edges (bonds) [61, 171]. From such a graph, adjacency matrices can be computed and used as inputs to train machine learning models [161, 172]. Furthermore, many chemical descriptors used in QSAR/QSPR are derived from topological molecular graphs and, hence, are termed topological descriptors [61]. Another representation of a compound is the molecular fingerprint, encoding physicochemical or structural properties in a vector. A widely used class of fingerprints are Extended Connectivity Fingerprints (ECFP) [172, 173]. Many methods in CADD such as molecular docking or MD simulations require a three-dimensional structure, which can be directly obtained from a molecular graph. However, the conformational flexibility of a molecule needs to be sampled in order to reach an energy minimum, typically facilitated by a molecular mechanics force-field [61, 174].

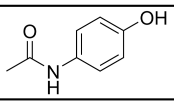
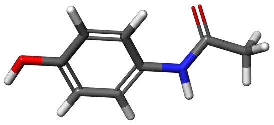
| **Type** | **Example** |
|---|---|
| Common name | Acetaminophen |
| Synonyms and trivial names | paracetamol, 4-Acetamidophenol, tylenol |
| IUPAC Name | N-(4-hydroxyphenyl)acetamide |
| Empirical formula | C8H9NO2 |
| 2D Structure |  |
| SMILES string | CC(=O)NC1=CC=C(C=C1)O |
| Adjacency matrix | $\left(\begin{array}{cccccccccccc} & v_1 & v_2 & v_3 & v_4 & v_5 & v_6 & v_7 & v_8 & v_9 & v_{10} & v_{11} \\ \hline v_1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ v_2 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ v_3 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ v_4 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ v_5 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ v_6 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ v_7 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ v_8 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ v_9 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ v_{10} & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ v_{11} & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{array}\right)$ |
| Fingerprint (EC-FP3) | 00004002 00000008 20000000 00000600 00000000 00000000<br>02000000 00000000 00000000 00000840 00280004 40008000<br>00008022 00100000 00000040 00000008 00040000 00000000<br>08000000 00020000 40100000 08000600 20002000 00000000<br>00000000 00000000 00040000 00000080 00008000 00060001<br>00000000 00000200 |
| 3D Structrure |  |

**Figure 11** Chemical representations. Different chemical representations are shown from names, the empirical formula, SMILES string, adjacency matrix (adapted after Rajesh *et al.* [175]), fingerprint (EC-FP3 computed in Open Babel [176], and the 3D structure by the example of acetaminophen.

Molecular representations such as SMILES strings and fingerprints are commonly used in another discipline within cheminformatics that covers search methods. It is potentially the earliest one covering methods for storing and searching of chemical structures. Typically, these structures are stored in chemical databases such as PubChem, ChEMBL, or the CAS Registry [61, 62, 171]. Finding and extracting data from such databases is also referred to as data mining [177]. Two widely used techniques in this context are (sub-)structure and similarity searches [61]. Search methods for whole structures rely on molecular identifiers such as graphs or linear notations, while substructure search is often done based on graph-derived pattern languages like like SMILES Arbitrary Target Specification (SMARTS) [61, 171]. The definition of a certain substructure is often insufficient to detect a family of structurally related compounds, especially if a number of common structural features are desired [61]. However, under the premise that similar structures exhibit similar characteristics, this is one of the fundamental applications of cheminformatics in drug discovery and, thus, algorithms for chemical similarity analysis were developed [61, 161, 178]. Different techniques can be applied for similarity searching, such as considering the maximum common substructure, molecular descriptors, or fingerprints [61]. The latter approach is frequently used in virtual screening where compounds are prioritized based on similarity coefficients such as the Tanimoto coefficient computing the fraction of bits shared between two feature vectors [161, 178].

With the rapid expansion of big data, cheminformatics techniques will continue to play an essential role in academic and industrial research [161]. In the context of this thesis, cheminformatics techniques were primarily applied in a project aimed at the development of a robust model to predict the metabolic fate of ester compounds. The obtained machine learning model, relying on a multi-scale modeling approach to compute various features of a query compound, can distinguish substrates of the enzymes hCE-1 from hCE-2 substrates with high accuracy as detailed in Chapter 4.

## References

[1] H.-J Böhm and G Schneider. *Protein-Ligand Interactions: From Molecular Recognition to Drug Design*. 1 2005. ISBN 9783527305216.

[2] Hans-Joachim Böhm and Gerhard Klebe. What Can We Learn from Molecular Recognition in Protein–Ligand Complexes for the Design of New Drugs? *Angewandte Chemie International Edition in English*, 35(22):2588–2614, 12 1996.

[3] Artur Gora, Jan Brezovsky, and Jiri Damborsky. Gates of enzymes. *Chemical Reviews*, 113(8):5871–5923, 2013.

[4] André Fischer, Martin Smiesko, and Martin Smieško. Ligand Pathways in Nuclear Receptors. *Journal of Chemical Information and Modeling*, 59(7):3100–3109, 2019.

[5] Joanne Bowes, Andrew J. Brown, Jacques Hamon, Wolfgang Jarolimek, Arun Sridhar, Gareth Waldron, and Steven Whitebread. Reducing safety-related drug attrition: The use of in vitro pharmacological profiling. *Nature Reviews Drug Discovery*, 11(12):909–922, 2012.

[6] Michael J. Waring, John Arrowsmith, Andrew R. Leach, Paul D. Leeson, Sam Mandrell, Robert M. Owen, Garry Pairaudeau, William D. Pennie, Stephen D. Pickett, Jibo Wang, Owen Wallace, and Alex Weir. An analysis of the attrition of drug candidates from four major pharmaceutical companies. *Nature Reviews Drug Discovery*, 14(7):475–486, 2015.

[7] Stephani Joy Y. Macalino, Vijayakumar Gosu, Sunhye Hong, and Sun Choi. Role of computer-aided drug design in modern drug discovery. *Archives of Pharmacal Research*, 38(9):1686–1701, 2015.

[8] Han van de Waterbeemd, Eric Gifford, and Ann Arbor. ADMET in silico modelling: towards prediction paradise? *Nat Rev Drug Discov*, 2(3):192–204, 3 2003.

[9] Ismail Kola and John Landis. Can the pharmaceutical industry reduce attrition rates? *Nature Reviews Drug Discovery*, 3(8):711–715, 2004.

[10] Omar Abdulhameed Almazroo, Mohammad Kowser Miah, and Raman Venkataramanan. Drug Metabolism in the Liver. *Clinics in Liver Disease*, 21(1):1–20, 2017.

[11] Takashi B T International Review of Cytology Iyanagi. Molecular Mechanism of Phase I and Phase II Drug-Metabolizing Enzymes: Implications for Detoxification. volume 260, pages 35–112. Academic Press, 2007. ISBN 0074-7696.

[12] Ulrich M. Zanger and Matthias Schwab. Cytochrome P450 enzymes in drug metabolism: Regulation of gene expression, enzyme activities, and impact of genetic variation. *Pharmacology and Therapeutics*, 138(1):103–141, 2013.

[13] Shu-Feng Zhou, Jun-Ping Liu, and Xin-Sheng Lai. Substrate specificity, inhibitors and regulation of human cytochrome P450 2D6 and implications in drug development. *Current medicinal chemistry*, 16(21):2661–2805, 2009.

[14] D Werck-Reichhart and R Feyereisen. Cytochromes P450: a success story. *Genome biology*, 1(6):REVIEWS3003, 2000.

[15] Manuel Sellner, André Fischer, Charleen G. Don, and Martin Smieško. Conformational landscape of cytochrome P450 reductase interactions. *International Journal of Molecular Sciences*, 22(3):1–14, 2021.

[16] T. S. Tracy, A. S. Chaudhry, B. Prasad, K. E. Thummel, E. G. Schuetz, X.-b. Zhong, Y.-C. Tien, H. Jeong, X. Pan, L. M. Shireman, J. Tay-Sontheimer, and Y. S. Lin. Interindividual Variability in Cytochrome P450-Mediated Drug Metabolism. *Drug Metabolism and Disposition*, 44(3):343–351, 2016.

[17] Maximilian J L J Fürst, Filippo Fiorentini, and Marco W Fraaije. Beyond active site residues: overall structural dynamics control catalysis in flavin-containing and heme-containing monooxygenases. *Current Opinion in Structural Biology*, 59:29–37, 2019.

[18] André Fischer and Martin Smieško. Spontaneous Ligand Access Events to Membrane-Bound Cytochrome P450 2D6 Sampled at Atomic Resolution. *Scientific Reports*, 9(1): 16411, 2019.

[19] Philippe Urban, Thomas Lautier, Denis Pompon, and Gilles Truan. Ligand Access Channels in Cytochrome P450 Enzymes: A Review. *Int J Mol Sci.*, 19(6), 5 2018.

[20] Karel Berka, Tereza Hendrychová, Pavel Anzenbacher, and Michal Otyepka. Membrane position of ibuprofen agrees with suggested access path entrance to cytochrome P450 2C9 active site. *Journal of Physical Chemistry A*, 115(41):11248–11255, 2011.

[21] Karel Berka, Markéta Paloncýová, Pavel Anzenbacher, and Michal Otyepka. Behavior of human cytochromes P450 on lipid membranes. *Journal of Physical Chemistry B*, 117 (39):11556–11564, 2013.

[22] André Fischer, Charleen G. Don, and Martin Smieško. Molecular Dynamics Simulations Reveal Structural Differences among Allelic Variants of Membrane-Anchored Cytochrome P450 2D6. *Journal of Chemical Information and Modeling*, 58(9):1962–1975, 2018.

[23] Eva Chovancova, Antonin Pavelka, Petr Benes, Ondrej Strnad, Jan Brezovsky, Barbora Kozlikova, Artur Gora, Vilem Sustr, Martin Klvana, Petr Medek, Lada Biedermannova, Jiri Sochor, and Jiri Damborsky. CAVER 3.0: A Tool for the Analysis of Transport Pathways in Dynamic Protein Structures. *PLoS Computational Biology*, 8(10):23–30, 2012.

[24] Petr Jeřábek, Jan Florián, Václav Martínek, P Jerabek, J Florian, and V Martinek. Lipid molecules can induce an opening of membrane-facing tunnels in cytochrome P450 1A2. *Phys. Chem. Chem. Phys.*, 18(44):30344–30356, 2016.

[25] Fatma Goksin Bahar, Kayoko Ohura, Takuo Ogihara, and Teruko Imai. Species difference of esterase expression and hydrolase activity in plasma. *Journal of pharmaceutical sciences*, 101(10):3979–3988, 10 2012.

[26] S Casey Laizure, Vanessa Herring, Zheyi Hu, Kevin Witbrodt, and Robert B Parker. The role of human carboxylesterases in drug metabolism: have we overlooked their importance? *Pharmacotherapy*, 33(2):210–222, 2 2013.

[27] Teruko Imai. Human carboxylesterase isozymes: catalytic properties and rational drug design. *Drug metabolism and pharmacokinetics*, 21(3):173–185, 6 2006.

[28] Peter Ettmayer, Gordon L. Amidon, Bernd Clement, and Bernard Testa. Lessons Learned from Marketed and Investigational Prodrugs. *Journal of Medicinal Chemistry*, 47(10): 2393–2404, 2004.

[29] Jiesi Xu, Yuanyuan Li, Wei-Dong Chen, Yang Xu, Liya Yin, Xuemei Ge, Kavita Jadhav, Luciano Adorini, and Yanqiao Zhang. Hepatic carboxylesterase 1 is essential for both normal and farnesoid X receptor-controlled lipid homeostasis. *Hepatology (Baltimore, Md.)*, 59(5):1761–1771, 5 2014.

[30] Wojciech Schönemann, Simon Kleeb, Philipp Dätwyler, Oliver Schwardt, and Beat Ernst. Prodruggability of carbohydrates — oral FimH antagonists. *Canadian Journal of Chemistry*, 94(11):909–919, 3 2016.

[31] Teruko Imai, Megumi Taketani, Mayumi Shii, Masakiyo Hosokawa, and Kan Chiba. Substrate Specificity of Carboxylesterase Isozymes and Their Contribution to Hydrolase Activity in Human Liver and Small Intestine. *Drug Metabolism and Disposition*, 34(10): 1734 LP – 1741, 10 2006.

[32] Masato Takahashi, Ibuki Hirota, Tomoyuki Nakano, Tomoyuki Kotani, Daisuke Takani, Kana Shiratori, Yura Choi, Masami Haba, and Masakiyo Hosokawa. Effects of steric hindrance and electron density of ester prodrugs on controlling the metabolic activation by human carboxylesterase. *Drug metabolism and pharmacokinetics*, 38:100391, 3 2021.

[33] Li-Wei Zou, Qiang Jin, Dan-Dan Wang, Qing-Kai Qian, Da-Cheng Hao, Guang-Bo Ge, and Ling Yang. Carboxylesterase Inhibitors: An Update. *Current medicinal chemistry*, 25(14):1627–1649, 2018.

[34] Latorya D Hicks, Janice L Hyatt, Shana Stoddard, Lyudmila Tsurkan, Carol C Edwards, Randy M Wadkins, and Philip M Potter. Improved, selective, human intestinal carboxylesterase inhibitors designed to modulate 7-ethyl-10-[4-(1-piperidino)-1-piperidino]carbonyloxycamptothecin (Irinotecan; CPT-11) toxicity. *Journal of medicinal chemistry*, 52(12):3742–3752, 6 2009.

[35] Giulio Vistoli, Alessandro Pedretti, Angelica Mazzolari, and Bernard Testa. In silico prediction of human carboxylesterase-1 (hCES1) metabolism combining docking analyses and MD simulations. *Bioorganic and Medicinal Chemistry*, 18(1):320–329, 2010.

[36] Giulio Vistoli, Alessandro Pedretti, Angelica Mazzolari, and Bernard Testa. Homology modeling and metabolism prediction of human carboxylesterase-2 using docking analyses by GriDock: a parallelized tool based on AutoDock 4.0. *Journal of computer-aided molecular design*, 24(9):771–787, 9 2010.

[37] Vineet K. Dhiman, Michael J. Bolt, and Kevin P. White. Nuclear receptors in cancer - Uncovering new and evolving roles through genomic analysis. *Nature Reviews Genetics*, 19(3):160–174, 2018.

[38] Richard Sever and Christopher K. Glass. Signaling by nuclear receptors. *Cold Spring Harbor Perspectives in Biology*, 5(3):1–4, 2013.

[39] Hollie I Swanson, Taira Wada, Wen Xie, Barbara Renga, Angela Zampella, Eleonora Distrutti, Stefano Fiorucci, Bo Kong, Ann M Thomas, Grace L Guo, Ramesh Narayanan, Muralimohan Yepuru, James T Dalton, and John Y L Chiang. Role of Nuclear Receptors in Lipid Dysfunction and Obesity-Related Diseases. *Drug Metabolism and Disposition*, 41(1):1 LP – 11, 1 2013.

[40] Peter E Lonergan and Donald J Tindall. Androgen receptor signaling in prostate cancer development and progression. *Journal of Carcinogenesis*, 10:20, 8 2011.

[41] S-I Hayashi, H Eguchi, K Tanimoto, T Yoshida, Y Omoto, a Inoue, N Yoshida, and Y Yamaguchi. The expression and function of estrogen receptor alpha and beta in human breast cancer and its clinical application. *Endocrine-related cancer*, 10(2):193–202, 2003.

[42] Kriti Singh, Ravi Shashi Nayana Munuganti, Eric Leblanc, Yu Lun Lin, Euphemia Leung, Nada Lallous, Miriam Butler, Artem Cherkasov, and Paul S Rennie. In silico discovery and validation of potent small-molecule inhibitors targeting the activation function 2 site of human oestrogen receptor alpha. *Breast cancer research : BCR*, 17:27, 2 2015.

[43] Zoran Culig. Molecular Mechanisms of Enzalutamide Resistance in Prostate Cancer. *Current Molecular Biology Reports*, 3(4):230–235, 2017.

[44] P Duesberg, R Stindl, and R Hehlmann. Explaining the high mutation rates of cancer cells to drug and multidrug resistance by chromosome reassortments that are catalyzed by aneuploidy. *Proceedings of the National Academy of Sciences of the United States of America*, 97(26):14295–14300, 12 2000.

[45] Neil Vasan, José Baselga, and David M Hyman. A view on drug resistance in cancer. *Nature*, 575(7782):299–309, 2019.

[46] Shahriar Koochekpour. Androgen receptor signaling and mutations in prostate cancer. *Asian Journal of Andrology*, 12(5):639–657, 2010.

[47] Kush Dalal, Mani Roshan-Moniri, Aishwariya Sharma, Huifang Li, Fuqiang Ban, Mohamed Hessein, Michael Hsing, Kriti Singh, Eric LeBlanc, Scott Dehm, Emma S. Tomlinson Guns, Artem Cherkasov, and Paul S. Rennie. Selectively targeting the DNA-binding domain of the androgen receptor as a prospective therapy for prostate cancer. *Journal of Biological Chemistry*, 289(38):26417–26429, 2014.

[48] E. S. Antonarakis, C. Chandhasin, E. Osbourne, J. Luo, M. D. Sadar, and F. Perabo. Targeting the N-Terminal Domain of the Androgen Receptor: A New Approach for the Treatment of Advanced Prostate Cancer. *The Oncologist*, 21(12):1427–1435, 2016.

[49] André Fischer and Martin Smieško. Ligand Pathways in Nuclear Receptors. *Journal of Chemical Information and Modeling*, 59(7):3100–3109, 7 2019.

[50] Linjie Zhao, Shengtao Zhou, and Jan-Åke Gustafsson. Nuclear Receptors: Recent Drug Discovery for Cancer Therapies. *Endocrine Reviews*, 40(5):1207–1249, 10 2019.

[51] Hamed Ghaffari and Linda R Petzold. Identification of influential proteins in the classical retinoic acid signaling pathway. *Theoretical Biology and Medical Modelling*, 15(1):16, 2018.

[52] Nada Lallous, Kush Dalal, Artem Cherkasov, and Paul S. Rennie. Targeting alternative sites on the androgen receptor to treat Castration-Resistant Prostate Cancer. *International Journal of Molecular Sciences*, 14(6):12496–12519, 2013.

[53] Kristine M Wadosky and Shahriar Koochekpour. Androgen receptor splice variants and prostate cancer: From bench to bedside. *Oncotarget*, 8(11):18550–18576, 2017.

[54] Kyllikki Haapala, Eija R Hyytinen, Mikko Roiha, Marita Laurila, Immo Rantala, Heikki J Helin, and Pasi A Koivisto. Androgen receptor alterations in prostate cancer relapsed during a combined androgen blockade by orchiectomy and bicalutamide. *Laboratory Investigation*, 81(12):1647–1651, 2001.

[55] André Fischer and Martin Smieško. Allosteric binding sites on nuclear receptors: Focus on drug efficacy and selectivity. *International Journal of Molecular Sciences*, 21(2):6–8, 2020.

[56] Fuqiang Ban, Eric Leblanc, Huifang Li, Ravi S N Munuganti, Kate Frewin, Paul S Rennie, and Artem Cherkasov. Discovery of 1 H-indole-2-carboxamides as novel inhibitors of the androgen receptor binding function 3 (BF3). *Journal of Medicinal Chemistry*, 57 (15):6867–6872, 2014.

[57] Ravi Shashi Nayana Munuganti, Eric Leblanc, Peter Axerio-Cilies, Christophe Labriere, Kate Frewin, Kriti Singh, Mohamed D H Hassona, Nathan A Lack, Huifang Li, Fuqiang Ban, Emma Tomlinson Guns, Robert Young, Paul S Rennie, and Artem Cherkasov. Targeting the binding function 3 (BF3) site of the androgen receptor through virtual screening. 2. Development of 2-((2-phenoxyethyl) thio)-1H-benzimidazole derivatives. *Journal of Medicinal Chemistry*, 56(3):1136–1148, 2013.

[58] Eric Biron and François Bédard. Recent progress in the development of protein-protein interaction inhibitors targeting androgen receptor-coactivator binding in prostate cancer, 7 2016.

[59] Víctor Buzón, Laia R. Carbó, Sara B. Estruch, Robert J. Fletterick, and Eva Estébanez-Perpiñ. A conserved surface on the ligand binding domain of nuclear receptors for allosteric control. *Molecular and Cellular Endocrinology*, 348(2):394–402, 2012.

[60] Guillermo Martinez-Ariza and Christopher Hulme. Recent advances in allosteric androgen receptor inhibitors for the potential treatment of castration-resistant prostate cancer. *Pharmaceutical patent analyst*, 4(5):387–402, 2015.

[61] Thomas Engel. Basic overview of chemoinformatics. *Journal of Chemical Information and Modeling*, 46(6):2267–2277, 2006.

[62] Andreas Bender and Nathan Brown. Special Issue: Cheminformatics in Drug Discovery. *ChemMedChem*, 13(6):467–469, 2018.

[63] A R Leach. *Molecular Modelling: Principles and Applications*. Prentice Hall, 2001. ISBN 9780582382107.

[64] Fernando D Prieto-Martínez, Edgar López-López, K Eurídice Juárez-Mercado, and José L Medina-Franco. *Chapter 2 - Computational Drug Design Methods—Current and Future Perspectives*. Academic Press, 2019. ISBN 978-0-12-816125-8.

[65] Lyn H. Jones and Mark E. Bunnage. Applications of chemogenomic library screening in drug discovery. *Nature Reviews Drug Discovery*, 16(4):285–296, 2017.

[66] Leonardo G. Ferreira, Ricardo N. Dos Santos, Glaucius Oliva, and Adriano D. Andricopulo. Molecular docking and structure-based drug design strategies. *Molecules*, 20(7): 13384–13421, 2015.

[67] Colin R. Groom, Ian J. Bruno, Matthew P. Lightfoot, and Suzanna C. Ward. The Cambridge structural database. *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials*, 72(2):171–179, 2016.

[68] RCSB Protein Data Bank, 2021.

[69] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, T N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 1 2000.

[70] Shoemaker Susannah and Nozomi Ando. X-rays in the Cryo-EM Era: Structural Biology's Dynamic Future. *Biochemistry*, 57(3):277–285, 2018.

[71] Hong Wei Wang and Jia Wei Wang. How cryo-electron microscopy and X-ray crystallography complement each other. *Protein Science*, 26(1):32–39, 2017.

[72] Magdalena Woińska, Simon Grabowsky, Paulina M. Dominiak, Krzysztof Woźniak, and Dylan Jayatilaka. Hydrogen atoms can be located accurately and precisely by x-ray crystallography. *Science Advances*, 2(5), 2016.

[73] Phineus R.L. Markwick, Thérèse Malliavin, and Michael Nilges. Structural biology by NMR: Structure, dynamics, and interactions. *PLoS Computational Biology*, 4(9), 2008.

[74] Jeffrey A. Purslow, Balabhadra Khatiwada, Marvin J. Bayro, and Vincenzo Venditti. NMR Methods for Structural Characterization of Protein-Protein Complexes. *Frontiers in Molecular Biosciences*, 7(January):1–8, 2020.

[75] Muhammed Tilahun Muhammed and Esin Aki-Yalcin. Homology modeling in drug discovery: Overview, current applications, and future perspectives. *Chemical Biology and Drug Design*, 93(1):12–20, 2019.

[76] Tareq Hameduh, Yazan Haddad, Vojtech Adam, and Zbynek Heger. Homology modeling in the time of collective and artificial intelligence. *Computational and Structural Biotechnology Journal*, 18:3494–3506, 2020.

[77] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A.A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(August), 2021.

[78] Caterina Bissantz, Bernd Kuhn, and Martin Stahl. A Medicinal Chemist's Guide to Molecular Interactions. *J. Med. Chem.*, 53(16):6241–6241, 2010.

[79] Michael R. Shirts, David L. Mobley, and Scott P. Brown. Free-energy calculations in structure-based drug design. *Drug Design*, pages 61–86, 2010.

[80] André Fischer, Martin Smieško, Manuel Sellner, and Markus A Lill. Decision Making in Structure-Based Drug Discovery: Visual Inspection of Docking Results. *Journal of Medicinal Chemistry*, 64(5):2489–2500, 3 2021.

[81] Veronica Salmaso and Stefano Moro. Bridging molecular docking to molecular dynamics in exploring ligand-protein recognition process: An overview. *Frontiers in Pharmacology*, 9(AUG):1–16, 2018.

[82] E Hadjittofis, S C Das, G G Z Zhang, and J Y Y Heng. Chapter 8 - Interfacial Phenomena. pages 225–252. Academic Press, Boston, 2017. ISBN 978-0-12-802447-8.

[83] Elangannan Arunan, Gautam R Desiraju, Roger A Klein, Joanna Sadlej, Steve Scheiner, Ibon Alkorta, David C Clary, Robert H Crabtree, Joseph J Dannenberg, Pavel Hobza, Henrik G Kjaergaard, Anthony C Legon, Benedetta Mennucci, and David J Nesbitt. Definition of the hydrogen bond (IUPAC Recommendations 2011):. *Pure and Applied Chemistry*, 83(8):1637–1641, 2011.

[84] A. J.Lopes Jesus and J. S. Redinha. Charge-assisted intramolecular hydrogen bonds in disubstituted cyclohexane derivatives. *Journal of Physical Chemistry A*, 115(48):14069–14077, 2011.

[85] Julio Caballero. Considerations for Docking of Selective Angiotensin-Converting Enzyme Inhibitors. *Molecules (Basel, Switzerland)*, 25(2), 1 2020.

[86] Paulo J. Costa. The halogen bond: Nature and applications. *Physical Sciences Reviews*, 2(11):1–16, 2019.

[87] C Rose Kennedy, Song Lin, and Eric N Jacobsen. The Cation–$\pi$ Interaction in Small-Molecule Catalysis. *Angewandte Chemie International Edition*, 55(41):12596–12624, 10 2016.

[88] Dennis A. Dougherty. The cation-$\pi$ interaction. *Accounts of Chemical Research*, 46(4): 885–893, 2013.

[89] Emmanuel A. Meyer, Ronald K. Castellano, and François Diederich. Interactions with aromatic rings in chemical and biological recognition. *Angewandte Chemie - International Edition*, 42(11):1210–1250, 2003.

[90] Kevin E. Riley and Pavel Hobza. On the importance and origin of aromatic interactions in chemistry and biodisciplines. *Accounts of Chemical Research*, 46(4):927–936, 2013.

[91] Ahmadreza Ghanbarpour, Amr H. Mahmoud, and Markus A. Lill. Instantaneous generation of protein hydration properties from static structures. *Communications Chemistry*, 3(1), 2020.

[92] Ying Yang, Bingjie Hu, and Markus A Lill. WATsite2.0 with PyMOL Plugin: Hydration Site Prediction and Visualization BT - Protein Function Prediction: Methods and Protocols. pages 123–134. Springer New York, New York, NY, 2017. ISBN 978-1-4939-7015-5.

[93] Lingle Wang, Jennifer Chambers, and Robert Abel. Protein–Ligand Binding Free Energy Calculations with FEP+ BT - Biomolecular Simulations: Methods and Protocols. pages 201–232. Springer New York, New York, NY, 2019. ISBN 978-1-4939-9608-7.

[94] Johannes Schiebel, Roberto Gaspari, Tobias Wulsdorf, Khang Ngo, Christian Sohn, Tobias E. Schrader, Andrea Cavalli, Andreas Ostermann, Andreas Heine, and Gerhard Klebe. Intriguing role of water in protein-ligand binding studied by neutron crystallography on trypsin complexes. *Nature Communications*, 9(1), 2018.

[95] Jinfang Zheng, Xu Hong, Juan Xie, Xiaoxue Tong, and Shiyong Liu. P3DOCK: a protein–RNA docking webserver based on template-based and template-free docking. *Bioinformatics*, 36(1):96–103, 1 2020.

[96] Sheng You Huang. Comprehensive assessment of flexible-ligand docking algorithms: Current effectiveness and challenges. *Briefings in Bioinformatics*, 19(5):982–994, 2018.

[97] Martin Smieško. DOLINA – Docking Based on a Local Induced-Fit Algorithm: Application toward Small-Molecule Binding to Nuclear Receptors. *Journal of Chemical Information and Modeling*, 53(6):1415–1423, 6 2013.

[98] Guimin Wang and Weiliang Zhu. Molecular docking for drug discovery and development: A widely used approach but far from perfect. *Future Medicinal Chemistry*, 8(14), 2016.

[99] Ruben Abagyan, Manuel Rueda, and Giovanni Bottegoni. Recipes for the selection of experimental protein conformations for virtual screening. *Journal of Chemical Information and Modeling*, 50(1):186–193, 2010.

[100] Jason C Cole, Christopher W Murray, J Willem M Nissink, Richard D Taylor, and Robin Taylor. Comparing protein-ligand docking programs is difficult. *Proteins*, 60(3):325–332, 8 2005.

[101] Kirk E Hevener, Wei Zhao, David M Ball, Kerim Babaoglu, Jianjun Qi, Stephen W White, and Richard E Lee. Validation of Molecular Docking Programs for Virtual Screening against Dihydropteroate Synthase. *Journal of Chemical Information and Modeling*, 49(2):444–460, 2 2009.

[102] Manuel Rueda and Ruben Abagyan. Best Practices in Docking and Activity Prediction. *bioRxiv*, 2 2016.

[103] Maria Kontoyianni, Laura M. McClellan, and Glenn S. Sokol. Evaluation of Docking Performance: Comparative Data on Docking Algorithms. *Journal of Medicinal Chemistry*, 47(3):558–565, 2004.

[104] Rodrigo Quiroga and Marcos A. Villarreal. Vinardo: A scoring function based on autodock vina improves scoring, docking, and virtual screening. *PLoS ONE*, 11(5):1–18, 2016.

[105] Michael M Mysinger, Michael Carchia, John. J Irwin, and Brian K Shoichet. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *Journal of Medicinal Chemistry*, 55(14):6582–6594, 7 2012.

[106] Charly Empereur-Mot, Hélène Guillemain, Aurélien Latouche, Jean François Zagury, Vivian Viallon, and Matthieu Montes. Predictiveness curves in virtual screening. *Journal of Cheminformatics*, 7(1):1–17, 2015.

[107] Amr H. Mahmoud, Matthew R. Masters, Ying Yang, and Markus A. Lill. Elucidating the multiple roles of hydration for accurate protein-ligand binding prediction via deep learning. *Communications Chemistry*, 3(1):1–13, 2020.

[108] David L Mobley and Ken A Dill. Binding of small-molecule ligands to proteins: ”what you see” is not always ”what you get”. *Structure (London, England : 1993)*, 17(4): 489–498, 4 2009.

[109] Daniel E Koshland. The Key-Lock Theory and the Induced Fit Theory Introduction of the Induced Fit Theory. *Angewandte Chemie International Edition*, 33:2375–2378, 1994.

[110] Edward B. Miller, Robert B. Murphy, Daniel Sindhikara, Kenneth W. Borrelli, Matthew J. Grisewood, Fabio Ranalli, Steven L. Dixon, Steven Jerome, Nicholas A. Boyles, Tyler Day, Phani Ghanakota, Sayan Mondal, Salma B. Rafi, Dawn M. Troast, Robert Abel, and Richard A. Friesner. Reliable and Accurate Solution to the Induced Fit Docking Problem for Protein-Ligand Binding. *Journal of Chemical Theory and Computation*, 17 (4):2630–2639, 2021.

[111] Deepak K. Lokwani, Aniket P. Sarkate, Kshipra S. Karnik, Anna Pratima G. Nikalje, and Julio A. Seijas. Structure-based site of metabolism (SOM) prediction of ligand for CYP3A4 enzyme: Comparison of glide XP and induced fit docking (IFD). *Molecules*, 25(7):1–13, 2020.

[112] Juan R. Perilla, Boon Chong Goh, C. Keith Cassidy, Bo Liu, Rafael C. Bernardi, Till Rudack, Hang Yu, Zhe Wu, and Klaus Schulten. Molecular dynamics simulations of large macromolecular complexes. *Current Opinion in Structural Biology*, 31:64–74, 2015.

[113] Jacob D Durrant and J Andrew McCammon. Molecular dynamics simulations and drug discovery. *BMC Biology*, 9(1):71, 2011.

[114] Marco De Vivo, Matteo Masetti, Giovanni Bottegoni, and Andrea Cavalli. Role of Molecular Dynamics and Related Methods in Drug Discovery. *Journal of Medicinal Chemistry*, 59(9):4035–4061, 2016.

[115] Matthew C. Zwier and Lillian T. Chong. Reaching biological timescales with all-atom molecular dynamics simulations. *Current Opinion in Pharmacology*, 10(6):745–752, 2010.

[116] Erik Lindahl. Molecular dynamics simulations. *Methods in molecular biology (Clifton, N.J.)*, 1215:3–26, 2015.

[117] Kleber Carlos Mundim and Dafydd Ellis. Stochastic Classical Molecular Dynamics Coupled to Functional Density Theory: Applications to Large Molecular Systems. *Braz J Phys*, 29, 5 1999.

[118] Hirotaka Ode, Masaaki Nakashima, Shingo Kitamura, Wataru Sugiura, and Hironori Sato. Molecular dynamics simulation in virus research. *Frontiers in Microbiology*, 3 (JUL), 2012.

[119] Vincent Kräutler, Wilfred F. Van Gunsteren, and Philippe H. Hünenberger. A fast SHAKE algorithm to solve distance constraint equations for small molecules in molecular dynamics simulations. *Journal of Computational Chemistry*, 22(5):501–508, 2001.

[120] Igor Omelyan and Andriy Kovalenko. Multiple time step molecular dynamics in the optimized isokinetic ensemble steered with the molecular theory of solvation: Accelerating with advanced extrapolation of effective solvation forces. *Journal of Chemical Physics*, 139(24), 2013.

[121] Cristian Predescu, Adam K. Lerer, Ross A. Lippert, Brian Towles, J. P. Grossman, Robert M. Dirks, and David E. Shaw. The u -series: A separable decomposition for electrostatics computation with improved accuracy. *Journal of Chemical Physics*, 152 (8), 2020.

[122] William L Jorgensen, David S Maxwell, and Julian Tirado-Rives. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *Journal of the American Chemical Society*, 118(45):11225–11236, 1 1996.

[123] Timur Aslyamov and Iskander Akhatov. Zeros of partition functions in the NPT ensemble. *Physical review. E*, 100(5-1):52118, 11 2019.

[124] R. E. Skyner, J. L. McDonagh, C. R. Groom, T. Van Mourik, and J. B.O. Mitchell. A review of methods for the calculation of solution free energies and the modelling of systems in solution. *Physical Chemistry Chemical Physics*, 17(9):6174–6191, 2015.

[125] Kota Kasahara, Shun Sakuraba, and Ikuo Fukuda. Enhanced Sampling of Molecular Dynamics Simulations of a Polyalanine Octapeptide: Effects of the Periodic Boundary Conditions on Peptide Conformation. *Journal of Physical Chemistry B*, 122(9):2495–2503, 2018.

[126] Adam Hospital, Josep Ramón Goñi, Modesto Orozco, and Josep Gelpi. Molecular dynamics simulations: Advances and applications. *Advances and Applications in Bioinformatics and Chemistry*, 8:37–47, 2015.

[127] Matthieu Dreher, Marc Piuzzi, Ahmed Turki, Matthieu Chavent, Marc Baaden, Nicolas Férey, Sébastien Limet, Bruno Raffin, and Sophie Robert. Interactive Molecular Dynamics: Scaling up to Large Systems. *Procedia Computer Science*, 18:20–29, 2013.

[128] Helmut Grubmüller and Paul Tavan. Multiple time step algorithms for molecular dynamics simulations of proteins: How good are they? *Journal of Computational Chemistry*, 19(13):1534–1552, 1998.

[129] Scott A. Hollingsworth and Ron O. Dror. Molecular Dynamics Simulation for All. *Neuron*, 99(6):1129–1143, 2018.

[130] John Russell. Anton 3 Is a 'Fire-Breathing' Molecular Simulation Beast, 2021.

[131] Vittorio Limongelli, Massimiliano Bonomi, and Michele Parrinello. Funnel metadynamics as accurate binding free-energy method. *Proceedings of the National Academy of Sciences of the United States of America*, 110(16):6358–6363, 2013.

[132] Romano T Kroemer. Structure-Based Drug Design: Docking and Scoring. *Current Protein and Peptide Science*, 8:312–328, 2007.

[133] Bingjie Hu and Markus A. Lill. WATsite: Hydration site prediction program with PyMOL interface. *Journal of Computational Chemistry*, 35(16):1255–1260, 2014.

[134] Samuel Genheden and Ulf Ryde. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opinion on Drug Discovery*, 10(5):449–461, 2015.

[135] Ercheng Wang, Huiyong Sun, Junmei Wang, Zhe Wang, Hui Liu, John Z.H. Zhang, and Tingjun Hou. End-Point Binding Free Energy Calculation with MM/PBSA and MM/GBSA: Strategies and Applications in Drug Design. *Chemical Reviews*, 119(16):9478–9508, 2019.

[136] Irene Maffucci, Xiao Hu, Valentina Fumagalli, and Alessandro Contini. An Efficient Implementation of the Nwat-MMGBSA Method to Rescore Docking Results in Medium-Throughput Virtual Screenings. *Frontiers in chemistry*, 6:43, 2018.

[137] Giulio Rastelli and Luca Pinzi. Refinement and Rescoring of Virtual Screening Results. *Frontiers in chemistry*, 7:498, 7 2019.

[138] Steven K. Albanese, John D. Chodera, Andrea Volkamer, Simon Keng, Robert Abel, and Lingle Wang. Is Structure-Based Drug Design Ready for Selectivity Optimization? *Journal of Chemical Information and Modeling*, 60(12):6211–6227, 2020.

[139] Yibing Shan, Eric T. Kim, Michael P. Eastwood, Ron O. Dror, Markus A. Seeliger, and David E. Shaw. How does a drug molecule find its target binding site? *Journal of the American Chemical Society*, 133(24):9181–9183, 2011.

[140] Phuc Chau Do, Eric H. Lee, and Ly Le. Steered Molecular Dynamics Simulation in Rational Drug Design. *Journal of Chemical Information and Modeling*, 58(8):1473–1482, 2018.

[141] Nguyen Quoc Thai, Hoang Linh Nguyen, Huynh Quang Linh, and Mai Suan Li. Protocol for fast screening of multi-target drug candidates: Application to Alzheimer's disease. *Journal of Molecular Graphics and Modelling*, 77:121–129, 2017.

[142] Susanna K. Lüdemann, Valère Lounnas, and Rebecca C. Wade. How do substrates enter and products exit the buried active site of cytochrome P450cam? 2. Steered molecular dynamics and adiabatic mapping of substrate pathways. *Journal of Molecular Biology*, 303(5):813–830, 2000.

[143] César Augusto F. De Oliveira, Donald Hamelberg, and J. Andrew McCammon. Coupling accelerated molecular dynamics methods with thermodynamic integration simulations. *Journal of Chemical Theory and Computation*, 4(9):1516–1525, 2008.

[144] Donald Hamelberg, John Mongan, and J Andrew McCammon. Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules. *Journal of Chemical Physics*, 120(24):11919–11929, 2004.

[145] Giovanni Bussi, Alessandro Laio, and Pratyush Tiwary. Metadynamics: A Unified Framework for Accelerating Rare Events and Sampling Thermodynamics and Kinetics. In *Handbook of Materials Modeling*, pages 565–595. 2020. ISBN 9783319446776.

[146] Alessandro Laio and Francesco L. Gervasio. Metadynamics: A method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Reports on Progress in Physics*, 71(12), 2008.

[147] André Fischer, Gabriela Frehner, Markus A Lill, and Martin Smieško. Conformational Changes of Thyroid Receptors in Response to Antagonists. *Journal of Chemical Information and Modeling*, 2021.

[148] André Fischer, Manuel Sellner, Karolina Mitusińska, Maria Bzówka, Markus A. Lill, Artur Góra, and Martin Smieško. Computational selectivity assessment of protease inhibitors against sars-cov-2. *International Journal of Molecular Sciences*, 22(4):1–17, 2021.

[149] Hitesh Patel, Björn A. Grüning, Stefan Günther, and Irmgard Merfort. PyWATER: a PyMOL plug-in to find conserved water molecules in proteins by clustering. *Bioinformatics (Oxford, England)*, 30(20):2978–2980, 2014.

[150] Robert Abel, Tom Young, Ramy Farid, Bruce J. Berne, and Richard A. Friesner. Role of the active-site solvent in the thermodynamics of factor Xa ligand binding. *Journal of the American Chemical Society*, 130(9):2817–2831, 2008.

[151] Marko Jukič, Janez Konc, Stanislav Gobec, and Dušanka Janežič. Identification of Conserved Water Sites in Protein Structures for Drug Design. *Journal of Chemical Information and Modeling*, 57(12):3094–3103, 2017.

[152] Bingjie Hu and Markus A. Lill. Protein pharmacophore selection using hydration-site analysis. *Journal of Chemical Information and Modeling*, 52(4):1046–1060, 2012.

[153] Peter M U Ung, Phani Ghanakota, Sarah E Graham, Katrina W Lexa, and Heather A Carlson. Identifying binding hot spots on protein surfaces by mixed-solvent molecular dynamics: HIV-1 protease as a test case. *Biopolymers*, 105(1):21–34, 1 2016.

[154] Phani Ghanakota and Heather A. Carlson. Driving Structure-Based Drug Discovery through Cosolvent Molecular Dynamics. *Journal of Medicinal Chemistry*, 59(23):10383–10399, 2016.

[155] Wenbo Yu, Sirish Kaushik Lakkaraju, E. Prabhu Raman, Lei Fang, and Alexander D. Mackerell. Pharmacophore modeling using site-identification by ligand competitive saturation (SILCS) with multiple probe molecules. *Journal of Chemical Information and Modeling*, 55(2):407–420, 2015.

[156] Yaw Sing Tan, David R. Spring, Chris Abell, and Chandra S. Verma. The Application of Ligand-Mapping Molecular Dynamics Simulations to the Rational Design of Peptidic Modulators of Protein-Protein Interactions. *Journal of Chemical Theory and Computation*, 11(7):3199–3210, 2015.

[157] Amr H. Mahmoud, Ying Yang, and Markus A. Lill. Improving Atom-Type Diversity and Sampling in Cosolvent Simulations Using $\lambda$-Dynamics. *Journal of Chemical Theory and Computation*, 15(5):3272–3287, 2019.

[158] S. Roy Kimura, Hai Peng Hu, Anatoly M. Ruvinsky, Woody Sherman, and Angelo D. Favia. Deciphering Cryptic Binding Sites on Proteins by Mixed-Solvent Molecular Dynamics. *Journal of Chemical Information and Modeling*, 57(6):1388–1401, 2017.

[159] Katrina W. Lexa and Heather A. Carlson. Full protein flexibility is essential for proper hot-spot mapping. *Journal of the American Chemical Society*, 133(2):200–202, 2011.

[160] Frank K Brown. Chapter 35 - Chemoinformatics: What is it and How does it Impact Drug Discovery. volume 33, pages 375–384. Academic Press, 1998. ISBN 0065-7743.

[161] Yu Chen Lo, Stefano E. Rensi, Wen Torng, and Russ B. Altman. Machine learning in chemoinformatics and drug discovery. *Drug Discovery Today*, 23(8):1538–1546, 2018.

[162] Stefano A Bini. Artificial Intelligence, Machine Learning, Deep Learning, and Cognitive Computing: What Do These Terms Mean and How Will They Impact Health Care? *The Journal of Arthroplasty*, 33(8):2358–2361, 2018.

[163] Vladimir Svetnik, Andy Liaw, Christopher Tong, J Christopher Culberson, Robert P Sheridan, and Bradley P Feuston. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *Journal of Chemical Information and Computer Sciences*, 43(6):1947–1958, 11 2003.

[164] Man Luo, Xiang Simon Wang, Bryan L Roth, Alexander Golbraikh, and Alexander Tropsha. Application of Quantitative Structure–Activity Relationship Models of 5-HT1A Receptor Binding to Virtual Screening Identifies Novel and Potent 5-HT1A Ligands. *Journal of Chemical Information and Modeling*, 54(2):634–647, 2 2014.

[165] Matías F. Andrada, Esteban G. Vega-Hissi, Mario R. Estrada, and Juan C. Garro Martinez. Application of k-means clustering, linear discriminant analysis and multivariate linear regression for the development of a predictive QSAR model on 5-lipoxygenase inhibitors. *Chemometrics and Intelligent Laboratory Systems*, 143:122–129, 2015.

[166] Vaibhaw, Jay Sarraf, and P K Pattnaik. Chapter 2 - Brain–computer interfaces and their applications. pages 31–54. Academic Press, 2020. ISBN 978-0-12-821326-1.

[167] Alexander Tropsha. Best Practices for QSAR Model Development, Validation, and Exploitation. *Molecular Informatics*, 29(6-7):476–488, 7 2010.

[168] Pavel Polishchuk. Interpretation of Quantitative Structure-Activity Relationship Models: Past, Present, and Future. *Journal of Chemical Information and Modeling*, 57(11):2618–2639, 2017.

[169] Artem Cherkasov, Eugene N Muratov, Denis Fourches, Alexandre Varnek, Igor I Baskin, Mark Cronin, John Dearden, Paola Gramatica, Yvonne C Martin, Roberto Todeschini, Viviana Consonni, Victor E Kuz'min, Richard Cramer, Romualdo Benigni, Chihae Yang, James Rathman, Lothar Terfloth, Johann Gasteiger, Ann Richard, and Alexander Tropsha. QSAR Modeling: Where Have You Been? Where Are You Going To? *Journal of Medicinal Chemistry*, 57(12):4977–5010, 6 2014.

[170] Daylight Chemical Information Systems. SMILES Notation, 2019.

[171] Hans Christian Ehrlich and Matthias Rarey. Systematic benchmark of substructure search in molecular graphs - From Ullmann to VF2. *Journal of Cheminformatics*, 4(7):1–17, 2012.

[172] Dejun Jiang, Zhenxing Wu, Chang Yu Hsieh, Guangyong Chen, Ben Liao, Zhe Wang, Chao Shen, Dongsheng Cao, Jian Wu, and Tingjun Hou. Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. *Journal of Cheminformatics*, 13(1):1–23, 2021.

[173] Laurianne David, Amol Thakkar, Rocío Mercado, and Ola Engkvist. Molecular representations in AI-driven drug discovery: a review and practical guide. *Journal of Cheminformatics*, 12(1):1–22, 2020.

[174] Oya Gürsoy and Martin Smieško. Searching for bioactive conformations of drug-like ligands with current force fields: How good are we? *Journal of Cheminformatics*, 9(1): 1–13, 2017.

[175] M R Rajesh Kanna and R Jagadeesh. Mathematics And its Applications Distinct Energies of Paracetamol Energy of a Graph. *Int. J. Math. And Appl.*, 6:781–789, 2018.

[176] Noel M O'Boyle, Michael Banck, Craig A James, Chris Morley, Tim Vandermeersch, and Geoffrey R Hutchison. Open Babel. *Journal of Cheminformatics*, 3(33):1–14, 2011.

[177] Rajarshi Guha, Kevin Gilbert, Geoffrey Fox, Marlon Pierce, David Wild, and Huapeng Yuan. Advances in cheminformatics methodologies and infrastructure to support the data mining of large, heterogeneous chemical datasets. *Current computer-aided drug design*, 6(1):50–67, 3 2010.

[178] Ingo Muegge and Prasenjit Mukherjee. An overview of molecular fingerprint similarity search in virtual screening. *Expert opinion on drug discovery*, 11(2):137–148, 2016.

## CHAPTER 2

# Spontaneous Ligand Access Events to Membrane-Bound CYP2D6 Sampled at Atomic Resolution

This study revealed, for the first time, the complete unbiased ligand recognition process of small-molecule substrates to the buried active site of wild-type CYP2D6 as well as an allelic variant with increased metabolic activity. Hence, it also introduces the family of drug-metabolizing enzymes which, along with NRs, are the main proteins addressed in this thesis. This work advances our understanding of the complex ligand recognition behavior of CYPs and gives insight into the structural consequences of amino acid mutations resulting from genetic polymorphism.

---

**Author contributions:** Conceptualization, A.F. and M.S.; methodology, A.F.; formal analysis, A.F.; writing and original draft preparation, A.F.; writing, review and editing, A.F., M.S.; visualization, A.F.; programming, A.F., M.S.; supervision, M.S.

---

*Based on the published research article:*

Fischer, A.; Smieško, M. Spontaneous Ligand Access Events to Membrane-Bound Cytochrome P450 2D6 Sampled at Atomic Resolution. *Sci. Rep.* 2019, 9, 16411.

## Abstract

The membrane-anchored enzyme Cytochrome P450 2D6 (CYP2D6) is involved in the metabolism of around 25% of marketed drugs and its metabolic performance shows a high interindividual variation. While it was suggested that ligands access the buried active site of the enzyme from the membrane, no proof from unbiased simulations has been provided to support this hypothesis. Laboratory experiments fail to capture the access process which is suspected to influence binding kinetics. Here, we applied unbiased molecular dynamics (MD) simulations to investigate the access of ligands to wild-type CYP2D6, as well as the allelic variant CYP2D6*53. In multiple simulations, substrates accessed the active site of the enzyme from the protein-membrane interface to ultimately adopt a conformation that would allow a metabolic reaction. We propose the necessary steps for ligand access and the results suggest that the increased metabolic activity of CYP2D6*53 might be caused by a facilitated ligand uptake.

## Introduction

Cytochrome P450 enzymes (CYPs) are essential proteins involved in the detoxification of foreign compounds reaching the human body. CYP2D6 accounts for the oxidative metabolism of roughly 25% of all marketed drugs and therefore belongs to the most relevant enzymes involved in phase I biotransformation. In addition, the enzyme is subject to a high interindividual variation in metabolic performance due to a genetic polymorphism. In drug therapy, this can ultimately lead to either severe adverse effects or the suppression of a therapeutic effect [1]. The allelic variant CYP2D6*53, which harbors the two amino acid mutations F120I and A122S, shows an increased metabolic rate towards several substrates in experiments indicating a pending designation as ultrarapid metabolizer (UM) phenotype [2, 3, 4, 5, 6, 7].

The active site of CYPs is located in a buried cavity inside the enzyme that is connected to the surrounding environment by tunnels [3, 8, 9, 10, 11, 12, 13, 14]. These tunnels are believed to influence both the poorly understood substrate specificity and binding kinetics of CYPs [1, 11, 13, 15]. As the prediction of CYP metabolism is of major importance for drug development [1], the influence of enzyme tunnels on small molecule binding has been intensively investigated [2, 10, 13, 16, 17]. Available experimental methods have only limited applicability for the determination and characterization of enzyme tunnels or complex transport phenomena. While crystal structures only provide a static view of the protein and fail to capture dynamic events, techniques such as NMR spectroscopy can produce dynamic information on the conformations of a protein [16, 8, 9]. Nonetheless, atomic details on the dynamic uptake of CYP2D6 ligands has not yet been produced by any experimental method [16]. Computer simulations, on the other hand, have provided fundamental insight into the transport process of ligands in CYPs. Various molecular dynamics (MD) simulation techniques have been applied to study the dynamic tunnels and their capability to transport ligands in CYPs [16, 15, 10, 11, 18, 19]. While most groups focused on the egress routes of ligands from the active site, only a handful of studies were focused on access routes [10, 18, 19] which are likely to be different [20, 13]. Due to the long timescale of such molecular processes [21], biasing potentials have been applied in nearly all studies to increase the likeliness of a successful translocation. Although two studies applied unbiased MD protocols to study the access to a CYP, they were focused on a soluble, bacterial enzyme [12]. Studies with CYP2D6 were limited to the determination and characterization of enzyme tunnels independent of a particular ligand [2, 4, 17, 22]. Overall, the access of ligands to mammalian CYPs is poorly understood and could not yet be observed in its full complexity in unbiased simulations.

In contrast to prokaryotes, mammalian drug-metabolizing CYPs are membrane-anchored and their globular domain is partially embedded in the membrane [14]. Based on the spatial location of several tunnels at the protein-membrane interface and the rather lipophilic character of many CYP ligands, it was suggested that ligands may access the active site from the membrane compartment and leave it efficiently through solvent-facing tunnels [2, 8, 10, 13, 14, 15, 18, 20]. For example, it was shown that the preferred position of ibuprofen relative to a membrane agrees with superficial entry points of access tunnels in CYP2C9 [13]. In another study, the spontaneous, non-reproducible insertion of a membrane lipid in an enzyme tunnel was observed [10]. No unbiased MD protocol was applied to confirm this hypothesis in a mammalian CYP, let alone in CYP2D6.

In this study, we performed over 20 $\mu$s of unbiased MD simulations with the aim to study the access of ligands to the buried active site of CYP2D6 in a model of the full-length structure of the enzyme anchored and partly embedded in a biological membrane [2]. In eight simulations, we observed substrates accessing the buried active site cavity of the enzyme via specific tunnels located at the protein-membrane interface. We propose the key steps governing a successful access of the ligand. Further, the results support the pending designation of the allelic variant CYP2D6*53 as a cause for ultra-rapid metabolizer phenotype based on a more efficient ligand uptake compared to the wild-type.

**Table 1** For each access event, the simulation identifier, the used protein structure, the accessing ligand, the time it took to be recognized at the tunnel entrance ($T_R$), the time it took for translocation to the active site ($T_T$), and the tunnel it translocated through, is shown.

| *Simulation* | *Structure* | *Ligand* | $T_R$ *(μs)* | $T_T$ *(μs)* | *Tunnel* | *SOM* |
|---|---|---|---|---|---|---|
| #3 | CYP2D6*53 | APAP-18 | 0.04 | 0.28 | 2f | yes |
| #4 | CYP2D6*53 | APAP-7 | 0.35 | 0.19 | 2f | yes |
|  |  | APAP-18 | n/a | n/a | 2f | no |
| #5 | CYP2D6*53 | APAP-18 | 0.15 | 0.82 | 4 | yes |
| #6 | CYP2D6*53 | APAP-6 | 0.61 | 0.08 | 2b | yes |
| #7 | CYP2D6*53 | APAP-3 | 0.27 | 0.19 | 2b | yes |
|  |  | APAP-8 | 0.61 | 0.41 | 2b | yes |
| #8 | wild-type | APAP-20 | 1.42 | n/a | 2f | no |
| #13 | CYP2D6*53 | BTD-11 | 0.03 | 0.03 | 2c | yes |
| #14 | CYP2D6*53 | BTD-3 | 0.01 | 0.03 | 2c | yes |

## Results and Discussion

**Access of CYP2D6 ligands from the protein-membrane interface.** Due to recent advances in computational capabilities, researchers are able to observe rare molecular events, such as intramolecular diffusion, inaccessible to laboratory experiments, using computer simulations [16, 20, 21, 23]. Simulations of events such as ligand binding can not only improve our understanding of fundamental molecular processes, but can also be used to estimate binding affinities and residence times of drug candidates [24]. Specifically for CYPs it was shown that distinct tunnels, predominantly located at the protein-membrane interface, connect the buried active site to its surrounding environment. Together with the general hydrophobicity of CYP substrates, this led to the widely discussed hypothesis of ligand access from the membrane [2, 8, 10, 13, 14, 18, 20]. Although, this hypothesis could not yet be proven based on a complete, unbiased trajectory of a ligand accessing the active site, it was supported by a study applying accelerated MD simulations to CYP3A4. However, the used technique might not have accounted for the exact dynamics of the system due to the biasing potential [25]. Further, the accordance of other simulation techniques involving biasing potentials to conventional MD is not inherently given as it was shown for adaptive sampling methods [26]. Therefore, unbiased simulations could serve as a blueprint to validate accelerated simulations and other biased simulation techniques that can then be used to tackle complex issues such as the determination of drug binding affinities. Previously, unbiased simulations were limited to the soluble, bacterial Camphor 5-monooxygenase (CYP101A1) meaning that the involvement of the membrane could not be considered [12, 27]. Here, we conducted unbiased MD simulations to investigate the access of ligands to CYP2D6 in a membrane-anchored model. We were able, for the first time, to observe the complete translocation of a ligand from the solvent to the buried active site cavity of a mammalian CYP in multiple simulations (Table 1 and S3). We randomly distributed 20 ligand molecules of either acetaminophen (APAP), butadiene (BTD), chlorzoxazone (CZX), debrisoquine (DEB), or propofol (PPF) inside the aqueous phase of the periodic boundary systems in an average distance of 13.8 Å (ranging between 2.7 and 49.2 Å) to the next protein heavy atom (Figure S1 and Table S4). In two exploratory simulations, a smaller number of ligands was used. From the solvent, the accessing ligands APAP and BTD sampled the simulation system, adhered to the tunnel entrance, and translocated to the active site of the allelic variant CYP2D6*53 through membrane-facing tunnels (Figures 1b,b and S3). Remarkably, our simulation setup did not require prior knowledge of the binding path or the location of the active site and produced ten access events in a total of 24 simulations.

In two simulations, we observed two ligands entering the active site consecutively. Notably, we observed only one, with 1.42 $\mu$s needed for recognition comparatively slow, access event with the wild-type structure despite substantial simulation efforts. Interestingly, the accessing BTD molecules were quickly recognized at the tunnel entrance without prior contact to the membrane. The other simulations performed in the context of this study did not result in the successful access of a ligand to the active site. Poten-

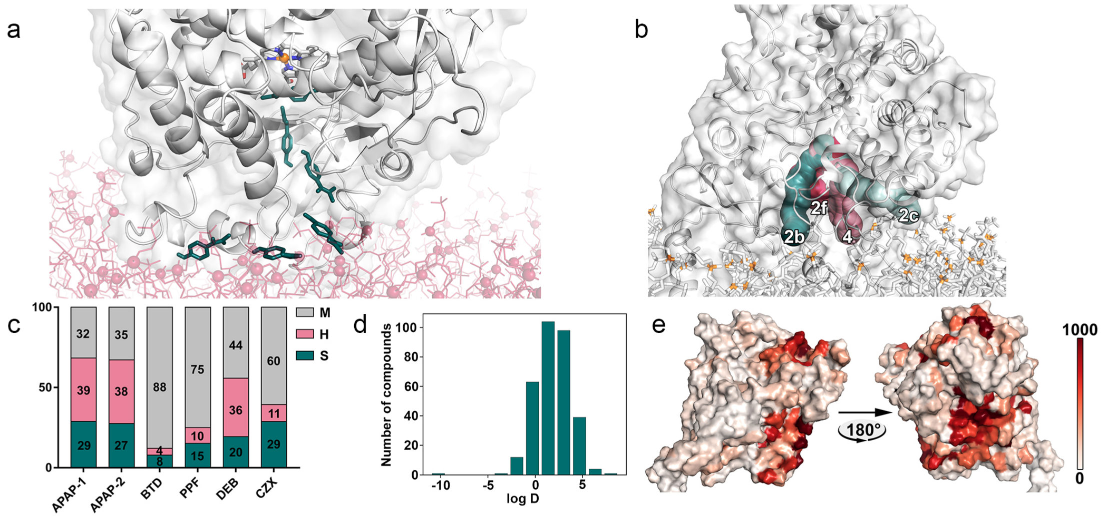


**Figure 12** Access tunnels and the spatial preference of ligands. (a) APAP accessing CYP2D6 from the protein-membrane interface in simulation #3. The ligand, shown in pine green, is starting outside the enzyme (bottom left) to access the active site. The membrane is colored red and phosphorus atoms are shown in sphere representation. Four helices are indicated for better orientation. (b) The four largest access tunnels are shown with the structure of CYP2D6*53. For orientation, the membrane phosphorus atoms are shown in orange. Other tunnels facing the solvent are not shown for simplicity. (c) The averaged distribution of ligands relative to three compartments consisting of membrane (M), head groups (H), and the remaining space (S) is shown. The two different measurements for APAP resulted from simulations at different temperatures. (d) Plot of the log D values predicted for a database of CYP2D6 ligands. The plot was generated using Matplotlib. (e) A visualization of the hotspots of APAP on the surface of CYP2D6 is shown in different shades of red. The scale from 0–1000 describes the cumulative number of ligand heavy atoms in a 5 Å radius of the CB atom (CA atom for glycine) of the protein amino acids. Two large hotspots were denoted as H1 and H2. For better orientation, the position of the FG loop is indicated.

tially, the simulations with DEB, PPF, and CZX could have led to ligand binding events if prolonged appropriately, but due to the high computational cost of the simulations we focused on the most promising candidate to reproducibly simulate binding in the given timescale, which was APAP based on the analysis of the first 480 ns of the simulations. The in-depth analysis of the access events revealed a considerable heterogeneity among them regarding the time taken to access the enzyme, the favored tunnel, as well as the adopted conformation in the active site (Table 1). After a slow recognition phase, the ligands adhered to one of the tunnel entrances located close to the protein-membrane interface. In two simulations (#3 and #5), the starting position was relatively close to the tunnel entrance leading to a fast recognition phase and limited sampling of the complete simulation system (Table 1). Further on, simulation #4 was a replica simulation meaning that one of the two accessing ligands (APAP-18) already started at the tunnel entrance, while the second accessing ligand (APAP-7) started in over 50 Å distance from the next protein heavy atom. In the remaining simulations, we could observe the recognition phase to be the rate-limiting step of the process with extensive sampling of the simulation system. An initial phase to recognize the tunnel entrance followed by a temporary superficial association, as we observed it, stands in agreement with a model describing a two-step binding process allowing kinetically efficient ligand uptake. This would enable the ligand to efficiently minimize the, otherwise even longer, recognition phase and is in accordance with recent observations regarding proteins with similarly buried active sites such as CYP101A1 and nuclear receptors [11, 21, 27, 28, 29, 30]. In the active site, eight out of ten accessing ligands adopted a pose which would allow an oxidation reaction to proceed at a site of metabolism (SOM) that would ultimately result in a metabolite in agreement with experiment [31]. In simulation #4, two ligands occupied the binding site simultaneously and one molecule was accommodated distant (>10 Å) from the heme. Further, simulation #8 with the wild-type structure did not result in a pose that was in agreement with metabolism of APAP.

**Validation of the simulations.** The models used in this study originated from our previously built, characterized, and validated full-length models of wild-type CYP2D6 as well as the allelic variant CYP2D6*53 [2]. Here, we evaluated key parameters and proved their accordance to our previous observations and experimentally derived literature values (Figure S3 and Table S5). These parameters included the burying depth of the enzyme, the heme tilt angle that describes the orientation of the enzyme to the membrane, as well as the root mean square deviation (RMSD) and root mean square fluctuation (RMSF). As mentioned above, we distributed multiple ligands into one simulation system. In order to determine to which degree ligands interacted with each other, we examined the presence of ligands in the proximity of the accessing molecule. Overall, we detected few interactions to other ligands, with the clear exception in the case of a double access event, where contacts were expected (Table S6). To rule out that large agglomerates formed during the simulations, we measured the distances of all ligands in the simulations and averaged them for each MD frame. While the data indicates that there was no large formation of agglomerates, the fluctuations of the average distance indicated the formation of transient small agglomerates (Figure S4). In comparison, the simulations with PPF showed a slightly increased trend for agglomeration, which might have contributed to the fact, that we did not observe any PPF molecules entering a tunnel. For a more detailed description of the validation, please refer to SI Results and Discussion.

**Preference of ligands for protein, tunnels, and membrane.** Regarding the difference in the preferred tunnels for translocation, it was suggested that multiple tunnels might serve as an access route to CYPs, specifically to govern the substrate specificity of the enzyme [8, 20]. In particular, tunnel entrances differing in burying depth within the membrane would allow the uptake of ligands varying in lipophilicity and therefore in their favored position relative to the membrane [13, 14, 18]. Indeed, the environment around the entrances of tunnels that were favored by the ligands varied as it can be seen at the example of tunnel 2c (Figure 1b). Furthermore, our analysis of the favored position relative to the membrane revealed significant differences among CYP2D6 ligands (Figure 1c), supporting this presumption. We logically divided the simulation box into three zones consisting of the membrane core (M), the head group region (H), and the remaining space made up of protein and solvent (S). BTD, CZX as well as PPF mainly partitioned towards the membrane core, while APAP preferred the head group region. The two slightly different temperatures in the simulations with APAP only had minor impact on the distribution. Despite the relatively similar behavior of DEB and APAP, we did not observe any DEB molecules accessing the enzyme. This might have been caused by the bulkier character of DEB requiring larger conformational changes for uptake as well as its slightly greater preference for the membrane core. During MD simulations, long residence times in the membrane core potentially reduce the probability to observe ligand access in the microsecond timescale. This is supported by the fact that both accessing lipophilic BTD molecules quickly entered the enzyme through the mostly solvated entrance of tunnel 2c near the protein-membrane interface without thorough sampling of the membrane core in our simulations. Likely, the hydrophobic milieu inside tunnel 2c [2] offered a favorable environment for the accessing BTD molecules. The calculated distribution coefficients (log D), describing the general preference of ligands towards a hydrophobic environment, revealed a peak around 2.5 for CYP2D6 ligands (Figure 1d). The clear difference in lipophilicity between APAP and BTD potentially influenced the selected tunnel for translocation to the active site. While APAP did not show a clear preference for a specific tunnel, BTD translocated through tunnel 2c in both access events (Table 1), which is likely associated with the relatively high lipophilicity of the amino acid residues lining this tunnel [2]. This, together with their different positions relative to the membrane compartment, underlines the importance of tunnels and their constitution for substrate specificity, since distinct chemical and geometrical features allow selective uptake of substrates [11, 13, 14].

Only recently, ligand-dependent long-range motions have been detected in an allosteric mechanism for CYP101A1, in which the occupancy of a peripheral site on the enzyme surface induces the opening of an access tunnel [12]. We identified two main sites (denoted as H1 and H2) during the nine simulations with APAP included in this analysis. Site H1 corresponded to a pocket around helices C, E, and H similar to the described allosteric site in CYP101A1, while site H2 highlighted a broad surface around helices F and A as well as the $\beta$ sheet 4 close to the entrance of tunnel 2f (Figure 1e). The H1 site was distant from the opening of any of the major tunnels. We determined these sites on the protein surface according to the number of ligand heavy atoms that were present in a 5 Å sphere around the amino acid residues in the respective simulations. Although we frequently observed the occupancy of the described H1 site in our simulations with APAP, the data indicated a secondary role of the above-mentioned allosteric mechanism for CYP2D6, since H1 occupancy was not mandatory for a successful translocation (Table S7). However, the data indicated that the H1 site might be involved in the opening of tunnel 2f. In this context, it was shown that the association of redox partners and dioxygen binding might additionally influence the conformational state of the enzyme [32]. Since the H2 site corresponds to a surface near the entrance of tunnel 2f, the data additionally supports the above-mentioned two-step binding mechanism. The BTD molecules did not sample the protein surface as extensively as APAP (Figure S5).

Structural adaptation of the protein. Since crystal structures do not provide a comprehensive explanation on how ligands access or leave the buried active site of CYPs, the protein has to undergo structural fluctuations to allow ligand access [8, 9, 15, 28]. In-

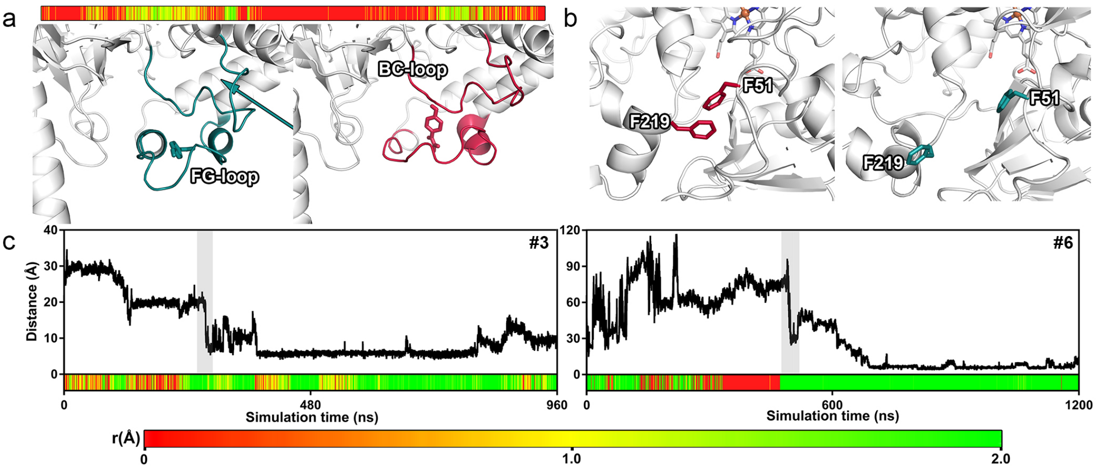


**Figure 13** Opening and adaptation of ligand tunnels. (a) APAP is shown at the entrance of tunnel 4 (defined by the FG loop) in simulation #5 at two different time points. On the left side the FG loop is clearly extended, while it presented a different conformation after the ligand advanced (right). The simultaneous movement of the BC loop lead to the reversible narrowing of tunnel 2c (arrow, left side), as it is indicated by the time-evolved bottleneck radius (shown above) of the simulation. The respective frames are marked on the color bar. (b) Gate between F51 and F219 shown in two different states. While the gate is closed at the beginning of simulation #3 (left), it adopted an open state after ligand translocation in simulation #4, forming a so-called wing gate [9]. (c) The distance between the SOM and the heme iron is plotted against the simulation time as well as the time-evolved bottleneck radius. The simulation identifier is shown at the top right of the plots, while gray bars indicate the period of tunnel translocation. The legend at the bottom indicates the coloring scheme for the bottleneck radii.

deed, we detected several dynamic adaptations of the protein that were, in certain cases, directly related to the accessing ligand. Similar to a recent study [27], the rearrangements did not alter the overall structural composition of the enzyme. The conformational changes of the secondary structure were mostly located on the protein surface, predominantly in regions with increased flexibility, and sometimes even in great distance from the ligand. In simulation #6 for example, the ligand induced a reversible conformational change of the FG loop, forming the entrance of tunnel 4, in order to propagate. This rearrangement additionally impacted the nearby BC loop leading to the tightening of tunnel 2c formed by this loop, as it was reflected by its temporarily decreased bottleneck radius (Figure 2a). This points towards an induced-fit mechanism as opposed to conformational selection [33]. Only recently the latter was proposed to be the main mechanism for multiple CYPs including CYP2D6 [30], even though for several other CYPs induced-fit scenarios were not ruled out. Besides movements of the

FG loop, we observed helix A, the BC loop, the HI loop, and helix B to be involved in conformational changes (Table S8). Since the mentioned structural adaptations often occurred in tunnels during the translocation of the ligand, it is likely that those structural elements are involved in gating the active site, as it was shown for other enzymes [9, 32]. On the level of amino acids, gates frequently consist of aromatic residues [2, 9, 11, 32]. We found individual residues to be involved in the gating of tunnel 2f, where F51 and F219 showed different conformations before and after ligand translocation (Figure 2b) without direct involvement of the ligand. In contrast to the above-mentioned conformational changes at tunnel 4, the opening of this gate can be best described as a conformational selection mechanism since the conformational change took place independent of the ligand molecule [33]. This suggests that depending on the tunnel used for translocation both induced-fit and conformational selection can describe the observed conformational changes. F51 and F219, among several other residues, functioned as bottleneck residues (Figure S6) which are often involved in gating tunnels [9]. Gates regulating enzyme tunnels are typically located at their most narrow part, which is determined by the bottleneck radius, and can be formed by secondary structural elements or individual residues. To determine the opening degree of the tunnels used for translocation, we monitored their bottleneck radii in simulations with access events (Figures 2c and S7). Based on the bottleneck radius, we discovered the favored tunnels to be open during the translocation of the ligand. Especially in simulation #6, it was clearly visible that the tunnel was closed when the ligand was approaching and opened shortly before its translocation (Figure 2c). Even though simulation #13 with BTD presented a similar opening pattern, the tunnel closed after translocation, implying conformational adaptations on the side of the protein. The conformational changes in relation to the movement of the ligand are in accordance with recent findings on an induced-fit driven mechanism of ligand binding to CYP101A1 [27]. In contrast to commonly described induced-fit mechanisms in active sites [34, 35], the described motions occurred at peripheral sites of the protein, such as the FG loop that is involved in the formation of multiple enzyme tunnels. Interestingly, we also observed motions of secondary structural elements in a significant distance (over 10 Å) from the ligand during the exploration of the active site such as the movements of the HI loop. In general, structural adaptations and protein flexibility are not only important to improve our understanding of the structural mechanism behind ligand access [8], but are also crucial to be considered in molecular docking calculations [35]. Even though MD-simulations are regularly used in a supporting role to post-process and refine poses obtained from docking, it was recently suggested that docking might even be replaced by MD-based techniques [35, 36]. Since our simulations lead from an unbound state to a bound state in the active site, this further supports these suggestions. Additionally, the results from docking APAP and BTD indicated, that the poses generated by flexible docking were strongly dependent on the used receptor structure (Table S9). When we compared poses obtained from docking and MD, we identified a similar (RMSD $< 2$ Å) pose in three out of eight access simulations (Figure 3a and Table S10). Our results show that the poses obtained from MD can closely resemble the ones obtained from docking, but they additionally allow to get insight into the dynamic interplay of the protein and the ligand and offer more potential for interpretation. Due to high computational expense that comes with conventional MD simulations as we used them, simulation techniques employing biasing potentials would offer a higher throughput for pose prediction from a completely unbound state [25, 36]. Together with the above-mentioned significant structural adaptations of the protein backbone, we conclude a rather limited applicability of traditional docking methods to such flexible proteins and support the use of MD-based methods.

**The driving forces for translocation.** Although the opening of gates was crucial for a successful ligand translocation, additional forces are required for the ligand to propagate to the active site in order for metabolism to occur steadily and reproducibly. We analyzed the nature of the interaction between the ligand and the protein in each simulation with an access event to identify the driving force for ligand translocation through the tunnels. Therefore, we considered contributions from electrostatics, hydrophobic contacts as implemented in the VSGB 2.0 model [37], and hydrogen bonds based on a term that accounts for their directionality [38]. While the ligands showed favorable hydrophobic interactions towards the membrane lipids at first, they generally decreased upon contact to the protein surface and increased again as the ligand advanced through the tunnel to enter the active site (Figure 3b). The correlation was especially evident

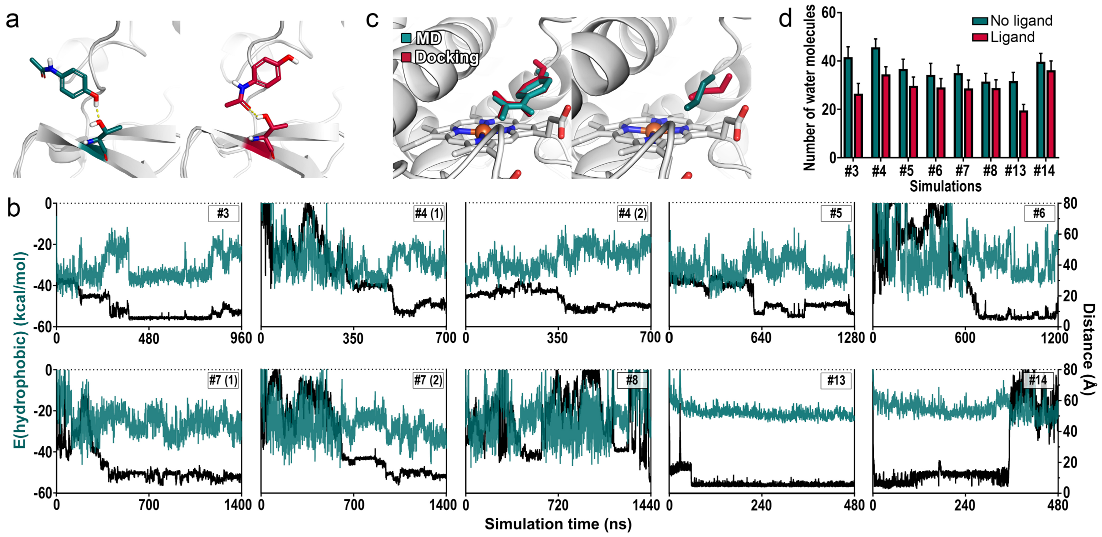


**Figure 14** Driving force for translocation, poses in the active site, and its desolvation. (a) T394 acting as a guiding rail for the ligand by hydrogen bonding. (b) The hydrophobic energy was plotted against the distance between the ligand SOM and the heme iron as well as the simulation time for all simulations with a successful access event. (c) Comparison of best matching poses of APAP (left) and BTD (right) obtained from MD simulations and molecular docking. (d) The number of water molecules in the active site in presence and absence of a ligand.

at the example of simulation #3, where a slight displacement of the ligand from its favored pose in the active site directly caused a substantial weakening of the hydrophobic energy. Even though not all simulations showed a clear correlation (#4 and #5), most of them presented a trend for a gain in hydrophobic energy during the translocation from the enzyme surface to the buried active site, where the known hydrophobic environment [1, 15] seemed to offer a favorable milieu for the ligands. The relatively fast access of both non-polar BTD molecules added additional evidence for the relevance of hydrophobicity. Contributions from electrostatics and hydrogen bonds were constantly present, but remained steady whether the ligand was in the solvent, membrane, or in the active site (Figure S8). However, polar contacts allowed ligands to adhere to the tunnel entrance and we observed distinct hydrogen bonds to guide APAP toward the active site by consecutively interacting with different heteroatoms (Figures 3a and S9a,b). This supports the role of polar contacts as a secondary driving force for the access of APAP, while BTD obviously could not form polar contacts due to the lack of heteroatoms. Together with gates, distinct polar interactions in enzyme tunnels have to play a relevant role in regulating the substrate specificity of the enzyme since hydrophobicity is a general property shared by many CYP substrates. The residues that interacted with the ligands during recognition, translocation, and the phase in the active site are shown in Tables S11-S13.

Other factors potentially influencing ligand uptake include the desolvation of the active site and the ligand. The displacement of water molecules in a binding site is a common strategy to optimize the binding affinity of compounds in the field of medicinal chemistry. Depending on the environment of the water molecule, the displacement can be both favorable or unfavorable [39, 40]. We identified a trend for a decreased number of water molecules in the active site when a ligand occupied it (Figure 3d) indicating a modest desolvation effect. However, the absolute numbers of displaced water molecules did not converge, likely due to the comparably small size of the ligands, the known enlarged active site cavity of CYP2D6*53 [2], and the overall heterogeneity of the individual access events. Similarly, the number of water molecules forming the hydration shell around the ligand did show great variation with no clear trend for APAP (Figure S9c,d). On the other hand, the number of water molecules accompanying the hydrophobic BTD molecules decreased to a significant amount on their journey from the bulk solvent to the active site. The desolvation of a drug-like ligand, associated with its binding to hydrophobic active sites [41], is generally a penalizing contribution toward affinity. In the case of BTD however, the solvation energy presents a positive value (0.61 kcal/mol) [42], leading to a favorable contribution for its desolvation. This suggests the partial desolvation of BTD as a favorable contribution toward its translocation. The, in this case, negligible influence of the conformational flexibility (Figure S10) on the access process is discussed in the SI Results and Discussion.

**Increased metabolic activity of CYP2D6*53.** Measurements of the enzymatic activity of allelic variant CYP2D6*53 have revealed increased metabolic rates towards bufuralol, dextromethorphan, and N-desmethyltamoxifen [3, 6, 7]. In contrast, a recent study reported a decrease in the clearance of primaquine [43]. It is suspected that the mostly increased metabolic rates of CYP2D6*53 are caused by an enlargement of enzyme tunnels allowing efficient ligand uptake to the enzyme [2, 4]. Altogether, this resulted in the speculation of an ultrarapid metabolizer (UM) phenotype for this allelic variant, which is usually only granted to phenotypes resulting from gene duplication [3, 4, 7]. Similar to our previous observations, tunnel 2b had a wider average bottleneck radius in CYP2D6*53 compared to the wild-type, likely due to the F120I mutation located at the entrance to the active site [2]. Further, the ligand access was faster in the CYP2D6*53 variant compared to the wild-type (Table 1). Therefore, our results moderately support the potential designation of CYP2D6*53 as UM phenotype based on a more efficient ligand uptake of the analyzed substrates.

## Conclusion

The results presented here revealed the atomic mechanism of ligand uptake to the buried active site of membrane-anchored CYP2D6 from the protein-membrane interface. The ligands APAP and BTD accessed the enzyme via different enzyme tunnels, which supports the notion of multiple functional tunnels within a single protein system. The tunnels varied in their burying depth in the membrane which would allow ligands differing in lipophilicity to access the active site. However, presumably due to the relative bulkiness of DEB, CZX and PPF and their increased partitioning towards the membrane core, the simulations with these ligands did not result in any binding events in this timescale. We show that the access process is linked to conformational adaptations of the protein backbone that can occur either in close proximity or in significant distance from the ligand molecule. While the conformational change at tunnel 4 followed an induced-fit mechanism, we also observed motions of residues that could be better described by a conformational selection model suggesting that both processes can occur in CYP2D6 depending on the tunnel. Together with the fact, that our simulations lead from an unbound to a bound state in a fully flexible unbiased manner, we support the use of MD-based techniques as opposed to docking, which stands in accordance with recent suggestions in the literature. Further, we show that the uptake process is mainly driven by hydrophobic interactions with a secondary role for polar contacts during recognition and translocation of the ligand molecules. In addition, the binding process was potentially facilitated by a modest desolvation of the active site. The difference in burying depth, physicochemical properties, and geometrical features of the tunnels influence their capability to transport certain ligands and therefore likely influence the specificity of the enzyme. Similarly, our results indicate that the increased metabolic rates of the allelic variant CYP2D6*53 might be caused by an efficient uptake of ligands compared to the wild-type enzyme. Our study could serve as a blueprint for simulations employing biasing potentials and it proves the capability of unbiased MD simulations to study ligand transport processes.

## Methods

As a starting point for our simulations we used our previously validated full-length model of wild-type CYP2D6 and CYP2D6*53 anchored to a membrane [2]. After randomly distributing multiple substrates in the solvent space around the enzyme, we performed over 20 $\mu$s of total unbiased MD simulations with multiple replica systems and various ligands with the aim to study ligand partitioning and to observe a translocation from the bulk solvent to the buried active site of the enzyme. All simulations were performed using the Desmond engine [44]. To determine the enzyme tunnels, we used CAVER 3.0 [45]. For the subsequent calculations, we either used workflows included in the Schrodinger Small-Molecule Drug Discovery Suite [46] or in-house routines. For a complete set of detailed materials and methods, please refer to SI Methods.

## References

[1] Charleen G. Don and M Smieško. Out-compute drug side effects: Focus on cytochrome P450 2D6 modeling. *WIREs Comput Mol Sci.*, 2018.

[2] André Fischer, Charleen G. Don, and Martin Smieško. Molecular Dynamics Simulations Reveal Structural Differences among Allelic Variants of Membrane-Anchored Cytochrome P450 2D6. *Journal of Chemical Information and Modeling*, 58(9):1962–1975, 2018.

[3] Sarah M Glass, Cydney M Martell, Alexandria K Oswalt, Victoria Osorio-Vasquez, Christi Cho, Michael J Hicks, Jacqueline M Mills, Rina Fujiwara, Michael J Glista, Sharat S Kamath, and Laura Lowe Furge. CYP2D6 Allelic Variants *34, *17-2, *17-3, and *53 and a Thr309Ala Mutant Display Altered Kinetics and NADPH Coupling in Metabolism of Bufuralol and Dextromethorphan and Altered Susceptibility to Inactivation by SCH 66712. *Drug Metab. Dispos.*, 46(8):1106–1117, 8 2018.

[4] Parker W. De Waal, Kyle F. Sunden, and Laura Lowe Furge. Molecular dynamics of CYP2D6 polymorphisms in the absence and presence of a mechanism-based inactivator reveals changes in local flexibility and dominant substrate access channels. *PLoS ONE*, 9 (10), 2014.

[5] Andrea Gaedigk, Magnus Ingelman-Sundberg, Neil A. Miller, J. Steven Leeder, Michelle Whirl-Carrillo, and Teri E. Klein. The Pharmacogene Variation (PharmVar) Consortium: Incorporation of the Human Cytochrome P450 (CYP) Allele Nomenclature Database. *Clinical Pharmacology and Therapeutics*, 00(00):4–6, 2017.

[6] Yuka Muroi, Takahiro Saito, Masamitsu Takahashi, Kanako Sakuyama, Yui Niinuma, Miyabi Ito, Chiharu Tsukada, Kiminori Ohta, Yasuyuki Endo, Akifumi Oda, Noriyasu Hirasawa, and Masahiro Hiratsuka. Functional Characterization of Wild-type and 49 CYP2D6 Allelic Variants for N-Desmethyltamoxifen 4-Hydroxylation Activity. *Drug Metabolism and Pharmacokinetics*, 29(5):360–366, 2014.

[7] Kanako Sakuyama, Takamitsu Sasaki, Shuta Ujiie, Kanako Obata, Michinao Mizugaki, Masaaki Ishikawa, and Masahiro Hiratsuka. Functional Characterization of 17 CYP2D6 Allelic Variants ( CYP2D6 . 2 , 10 , 14A – B , 18 , 27 , 36 , 39 , 47 – 51 , 53 – 55 , and 57 ). *Pharmacology*, 36(12):2460–2467, 2008.

[8] Vlad Cojocaru, Peter J. Winn, and Rebecca C. Wade. The ins and outs of cytochrome P450s. *Biochim Biophys Acta.*, 1770(3):390–401, 2007.

[9] Artur Gora, Jan Brezovsky, and Jiri Damborsky. Gates of enzymes. *Chemical Reviews*, 113(8):5871–5923, 2013.

[10] Petr Jeřábek, Jan Florián, Václav Martínek, P Jerabek, J Florian, and V Martinek. Lipid molecules can induce an opening of membrane-facing tunnels in cytochrome P450 1A2. *Phys. Chem. Chem. Phys.*, 18(44):30344–30356, 2016.

[11] Philippe Urban, Thomas Lautier, Denis Pompon, and Gilles Truan. Ligand Access Channels in Cytochrome P450 Enzymes: A Review. *Int J Mol Sci.*, 19(6), 5 2018.

[12] Alec H Follmer, Mavish Mahomed, David B Goodin, and Thomas L Poulos. Substrate-Dependent Allosteric Regulation in Cytochrome P450cam (CYP101A1). *Journal of the American Chemical Society*, 140:16222–16228, 2018.

[13] Karel Berka, Tereza Hendrychová, Pavel Anzenbacher, and Michal Otyepka. Membrane position of ibuprofen agrees with suggested access path entrance to cytochrome P450 2C9 active site. *Journal of Physical Chemistry A*, 115(41):11248–11255, 2011.

[14] Karel Berka, Markéta Paloncýová, Pavel Anzenbacher, and Michal Otyepka. Behavior of human cytochromes P450 on lipid membranes. *Journal of Physical Chemistry B*, 117(39): 11556–11564, 2013.

[15] Peter J Winn, Susanna K Lüdemann, Ralph Gauges, Valère Lounnas, and Rebecca C Wade. Comparison of the dynamics of substrate access channels in three cytochrome P450s reveals different opening mechanisms and a novel functional role for a buried arginine. *Proceedings of the National Academy of Sciences of the United States of America*, 99(8): 5361–5366, 2002.

[16] Markéta Paloncýova, Veronika Navrátilova, Karel Berka, Alessandro Laio, and Michal Otyepka. Role of Enzyme Flexibility in Ligand Access and Egress to Active Site: Bias-Exchange Metadynamics Study of 1,3,7-Trimethyluric Acid in Cytochrome P450 3A4. *Journal of Chemical Theory and Computation*, 12(4):2101–2109, 2016.

[17] Charleen G. Don and Martin Smieško. Microsecond MD simulations of human CYP2D6 wild-type and five allelic variants reveal mechanistic insights on the function. *PLoS ONE*, 13(8):1–21, 2018.

[18] John C. Hackett. Membrane-embedded substrate recognition by cytochrome P450 3A4. *Journal of Biological Chemistry*, 293(11):4037–4046, 2018.

[19] Shabana Vohra, Maria Musgaard, Stephen G. Bell, Luet Lok Wong, Weihong Zhou, and Philip C. Biggin. The dynamics of camphor in the cytochrome P450 CYP101D2. *Protein Science*, 22(9):1218–1229, 2013.

[20] Karin Schleinkofer, Sudarko, Peter J. Winn, Susanne K. Lüdemann, and Rebecca C. Wade. Do mammalian cytochrome P450s show multiple ligand access pathways and ligand channelling? *EMBO Reports*, 6(6):584–589, 2005.

[21] R. O. Dror, A. C. Pan, D. H. Arlow, D. W. Borhani, P. Maragakis, Y. Shan, H. Xu, and D. E. Shaw. Pathway and mechanism of drug binding to G-protein-coupled receptors. *Proceedings of the National Academy of Sciences*, 108(32):13118–13123, 2011.

[22] Yeimy Viviana Ariza Márquez, Ignacio Briceño, Fabio Aristizábal, Luis Fernando Niño, and Juvenal Yosa Reyes. Dynamic Effects of CYP2D6 Genetic Variants in a Set of Poor Metaboliser Patients with Infiltrating Ductal Cancer Under Treatment with Tamoxifen. *Scientific Reports*, 9(1):2521, 2019.

[23] J. Rydzewski and W. Nowak. Ligand diffusion in proteins via enhanced sampling in molecular dynamics. *Physics of Life Reviews*, 22-23:58–74, 12 2017.

[24] Yibing Shan, Eric T. Kim, Michael P. Eastwood, Ron O. Dror, Markus A. Seeliger, and David E. Shaw. How does a drug molecule find its target binding site? *Journal of the American Chemical Society*, 133(24):9181–9183, 2011.

[25] Levi C T Pierce, Romelia Salomon-Ferrer, Cesar Augusto F de Oliveira, J Andrew McCammon, and Ross C Walker. Routine Access to Millisecond Time Scale Events with Accelerated Molecular Dynamics. *Journal of Chemical Theory and Computation*, 8(9): 2997–3002, 9 2012.

[26] Robin M Betz and Ron O Dror. How Effectively Can Adaptive Sampling Methods Capture Spontaneous Ligand Binding? *Journal of Chemical Theory and Computation*, 15(3): 2053–2063, 3 2019.

[27] Navjeet Ahalawat and Jagannath Mondal. Mapping the Substrate Recognition Pathway in Cytochrome P450. *Journal of the American Chemical Society*, 140(50):17743–17752, 2018.

[28] Yao Huili, McCullough Christopher R., Costache Aurora D., Pullela Phani Kumar, and Sem Daniel S. Structural evidence for a functionally relevant second camphor binding site in P450cam: Model for substrate entry into a P450 active site. *Proteins.*, 69(1):125–138, 6 2007.

[29] Karl Edman, Ali Hosseini, Magnus K Bjursell, Anna Aagaard, Lisa Wissler, Anders Gunnarsson, Tim Kaminski, Christian Köhler, Stefan Bäckström, Tina J Jensen, Anders Cavallin, Ulla Karlsson, Ewa Nilsson, Daniel Lecina, Ryoji Takahashi, Christoph Grebner, Stefan Geschwindner, Matti Lepistö, Anders C Hogner, and Victor Guallar. Ligand Binding Mechanism in Steroid Receptors: From Conserved Plasticity to Differential Evolutionary Constraints. *Structure*, 23(12):2280–2291, 2015.

[30] F. Peter Guengerich, Clayton J. Wilkey, and Thanh T.N. Phan. Human cytochrome P450 enzymes bind drugs and other substrates mainly through conformational-selection modes. *Journal of Biological Chemistry*, 294(28):10928–10941, 2019.

[31] Chris De Graaf, Chris Oostenbrink, Peter H.J. J Keizers, Tushar Van Der Wijst, Aldo Jongejan, and Nico P.E. E Vermeulen. Catalytic site prediction and virtual screening of cytochrome P450 2D6 substrates by consideration of water and rescoring in automated docking. *Journal of Medicinal Chemistry*, 49(8):2417–2430, 2006.

[32] Alec H Follmer, Sarvind Tripathi, and Thomas L Poulos. Ligand and Redox Partner Binding Generates a New Conformational State in Cytochrome P450cam (CYP101A1). *Journal of the American Chemical Society*, 141:2678–2683, 2019.

[33] Fabian Paul and Thomas R Weikl. How to Distinguish Conformational Selection and Induced Fit Based on Chemical Relaxation Rates. *PLOS Computational Biology*, 12(9): e1005067, 9 2016.

[34] Daniel E. Koshland. The Key–Lock Theory and the Induced Fit Theory. *Angew. Chem.*, 33(23-24):2375–2378, 1995.

[35] Veronica Salmaso and Stefano Moro. Bridging molecular docking to molecular dynamics in exploring ligand-protein recognition process: An overview. *Frontiers in Pharmacology*, 9(AUG):1–16, 2018.

[36] Dario Gioia, Martina Bertazzo, Maurizio Recanatini, Matteo Masetti, and Andrea Cavalli. Dynamic docking: A paradigm shift in computational drug discovery. *Molecules*, 22(11): 1–21, 2017.

[37] Jianing Li, Robert Abel, Kai Zhu, Yixiang Cao, Suwen Zhao, and Richard A. Friesner. The VSGB 2.0 model: A next generation energy model for high resolution protein structure modeling. *Proteins.*, 79(10):2794–2812, 2011.

[38] Angelo Vedani and David W Huhta. A new force field for modeling metalloproteins. *Journal of the American Chemical Society*, 112(12):4759–4767, 6 1990.

[39] Joel Wahl and Martin Smieško. Thermodynamic Insight into the Effects of Water Displacement and Rearrangement upon Ligand Modifications using Molecular Dynamics Simulations. *ChemMedChem*, 13(13):1325–1335, 2018.

[40] Caterina Bissantz, Bernd Kuhn, and Martin Stahl. A Medicinal Chemist's Guide to Molecular Interactions. *J. Med. Chem.*, 53(16):6241–6241, 2010.

[41] Michael M Mysinger and Brian K Shoichet. Rapid Context-Dependent Ligand Desolvation in Molecular Docking. *J. Chem. Inf. Model.*, pages 1561–1573, 2010.

[42] A. V. Marenich, C. P. Kelly, J. D. Thompson, G. D. Hawkins, C. C. Chambers, D. J. Giesen, P. Winget, C. J. Cramer, and D. G. Truhlar. Minnesota Solvation Database—version 2012. *University of Minnesota*, 2012.

[43] Takahiro Saito, Evelyn Marie Gutiérrez Rico, Aoi Kikuchi, Akira Kaneko, Masaki Kumondai, Fumika Akai, Daisuke Saigusa, Akifumi Oda, Noriyasu Hirasawa, and Masahiro Hiratsuka. Functional characterization of 50 CYP2D6 allelic variants by assessing primaquine 5-hydroxylation. *Drug Metabolism and Pharmacokinetics*, 33:250–257, 2018.

[44] Kevin Bowers, Edmond Chow, Huafeng Xu, Ron Dror, Michael Eastwood, Brent Gregersen, John Klepeis, Istvan Kolossvary, Mark Moraes, Federico Sacerdoti, John Salmon, Yibing Shan, and David Shaw. Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters. *ACM/IEEE SC 2006 Conference (SC'06)*, (November):43, 2006.

[45] Eva Chovancova, Antonin Pavelka, Petr Benes, Ondrej Strnad, Jan Brezovsky, Barbora Kozlikova, Artur Gora, Vilem Sustr, Martin Klvana, Petr Medek, Lada Biedermannova, Jiri Sochor, and Jiri Damborsky. CAVER 3.0: A Tool for the Analysis of Transport Pathways in Dynamic Protein Structures. *PLoS Computational Biology*, 8(10):23–30, 2012.

[46] New York NY Schrödinger, LLC and New York N Y Schrödinger LLC. Small-Molecule Drug Discovery Suite 2017-2, 2017.

## 2.1 Supporting Information

## Supporting Materials and Methods

### Computational setup and general simulation conditions

The molecular dynamics (MD) simulations were performed on consumer-grade desktop computers equipped with graphics processing units (GPUs) or a dedicated, rack-mounted GPU server. On all used machines in this study, the Desmond simulation engine (v2016-4) was installed in a Linux environment [1]. Prior to the MD simulations, the default relaxation protocol of Desmond (Table S1) was conducted.

**Table S 1** Relaxation protocol prior to MD simulation.

| *Desmond stage* | *Procedure* |
|---|---|
| 1 | Task (reading files, initializing parameters) |
| 2 | Simulate, Brownian Dynamics, NVT, T = 10 K, small time steps, and restraints on solute heavy atoms, 100 ps |
| 3 | Simulate, NVT, T = 10 K, small time steps, and restraints on solute heavy atoms, 12 ps |
| 4 | Simulate, NPT, T = 10 K, and restraints on solute heavy atoms, 12 ps |
| 5 | Solvate pocket |
| 6 | Simulate, NPT and restraints on solute heavy atoms, 12 ps |
| 7 | Simulate, NPT and no restraints, 24 ps |

We chose the OPLS_2005 force field in an NPT ensemble and combined the Martyna-Tobias-Klein barostat with a relaxation time of 2.0 ps at 300 K with the Nose-Hoover thermostat at a relaxation time of 1.0 ps. We used the u-series [2] method to treat long-range interactions combined with a cutoff of 9 Å for short range interactions. By default, the M-SHAKE algorithm was used to constrain bonds to hydrogen atoms and no hydrogen mass partitioning was applied. The orthorhombic periodic boundary boxes were solvated with TIP3P water molecules, just as in our previous work [3]. The time step of the RESPA integrator was set to 2.0 fs and frames with atomic coordinates were written every 48 ps in all simulations. If it is not indicated otherwise, figures of molecules were generated using PyMol [4] and plots were generated using Prism GraphPad.

### Ligand preparation

To perform the ligand access simulations, we selected five ligands, including acetaminophen (APAP), 1,3-butadiene (BTD), chlorzoxazone (CZX), debrisoquine (DEB), and propofol (PPF) from a database of CYP2D6 ligands [5]. We retrieved the two-dimensional (2D) ligand structures from the PubChem structure database (Table S2)

[6]. Maestro from the Schrodinger Small-Molecule Drug Discovery Suite8 provides the Epik [7] environment to predict the protonation state of ligands. After pipelining all ligands through Epik at physiological pH (7.4), with water as solvent, and the inclusion of tautomers, the highest scored output structure was used as an input in the Conformational Search panel in Maestro [8]. Thereat, we selected the OPLS3 force field since it previously showed to deliver reliable results in the determination of ligand conformations [9]. We chose the Mixed torsional/Low-mode sampling algorithm with enhanced torsional sampling and a maximal number of 5000 Monte Carlo steps. The conformational search was carried out with water as solvent. For minimization after the conformational search, we selected the Truncated Newton Conjugate Gradient (TNCG) method with the maximal number of iterations set to 500. We retained the default convergence threshold of 0.05. The highest ranked structures were selected for the following simulations. Further, we retrieved the 2D structures of 323 CYP2D6 ligands from the PubChem structure database [6], according to the list published by Rendic and colleagues [5]. We used the cxcalc module provided by ChemAxon [10] to compute the log D values at physiological pH for all the structures.

**Table S 2** Ligands used in this study, abbreviations, and PubChem ID codes.

| *Ligand* | *Abbreviation* | *PubChem ID code* |
|---|---|---|
| acetaminophen | APAP | 1983 |
| 1,3-butadiene | BTD | 7845 |
| chlorzoxazone | CZX | 2733 |
| debrisoquine | DEB | 2966 |
| propofol | PPF | 4943 |

## Ligand preparation

The preparation of the protein structures of CYP2D6 as well as the placement of the membrane for the simulations in this study is extensively described elsewhere [3]. In brief, we used a covalently linked combination of the globular domain of CYP2D6 and its corresponding membrane anchor, both preequilibrated in the membrane environment, as a starting point for this study. The globular domain of the protein originally derived from a crystal structure (PDB ID code 3TDA). For simulations with the allelic variant CYP2D6*53, we introduced the mutations according to the PharmVar database [11] in the Maestro graphical user interface (GUI). In 22 simulations, 20 ligands were randomly distributed around the enzyme in the aqueous phase. Additionally, we performed two exploratory simulations with two or six ligands respectively (Table S3). The ligands were randomly translated and rotated relative to the simulation system to obtain unique starting positions. Two exceptions are as follows: simulation #4 was a replica simulation based on simulation #3 started from frame 3500 and simulation #5 was conducted at a different temperature, ensuring that the trajectories were set for a unique course. As the membrane constituent, we chose 1-palmitoyl-2-oleoylphosphatidylcholine (POPC) molecules and built a simulation system using the Desmond System Builder. Next, we used the Desmond Minimization routine to relax the system with 10000 as a maximal number of steps and a convergence threshold of 0.5 kcal/mol/Å$^3$. The simulations were set up to run for different durations and several of them were continued with all settings of the prior simulation being retained. For the two initial simulations, the temperature was left at the default value of 300.00 K (26.85 °C or 80.33 °F), while we selected the temperature to be either 310.00 K (36.85 °C or 98.33 °F) or 313.15 K (40 °C or 104 °F) for the following simulations. The increased temperature compared to the physiological state represents a patient with fever which is one of the main indications for the pharmacotherapy with APAP [12]. We determined the RMSD as well as the RMSF of the simulations using the Simulation Interaction Diagram panel within Maestro. For these calculations, the residues forming the flexible membrane anchor (residue numbers 1-31) were excluded due to their large movements compared to the rest of the protein. To assess the heme tilt angle during our simulations, we used an in-house script looping over MD frames in the PDB format extracted every 960 ps of the simulations. The heme tilt angle is defined as the angle between the heme plane, defined by the porphyrin nitrogen atoms, and the z-axis of the system representing the membrane normal. Likewise, we used the same frames to calculate the burying depth of the enzyme in the membrane according to the method established by Ducassou and colleagues [13], who defined the distances between the mass centers of the protein $\alpha$-carbons and the C1 atoms of the membrane molecules as burying depth. As before, we used an in-house routine to pipeline the MD frames through this calculation and determine average values. The contacts between the ligands were also determined using an in-house python routine that evaluated every frame of the respective MD simulation. Thereat, we determined the number of frames, in which the 5 Å zone around accessing ligand molecule included a heavy atom of another ligand based on individual MD frames exported from the Maestro GUI. We divided the results into three phases according to the progress of the access event. Residues involved in the translocation of the ligand were determined using the Simulation Interaction Diagram panel in Maestro. Simultaneously, the torsion angles of the ligands were monitored. The adaptation of secondary structure elements was determined based on the RMSD and RMSF diagrams as well as the careful visual examination of the MD trajectories. A ligand was considered to be in a pose which would allow oxidation reaction to proceed at a site of metabolism (SOM) when the distance between the SOM and the heme iron was between 5 and 7 Å.

**Table S 3** Overview of all simulations conducted throughout this study.

| *Simulation* | *Structure* | *Ligands* | *Temperature (K)* | *Duration ($\mu s$)* |
|---|---|---|---|---|
| #1 | CYP2D6 WT | 2x CZX | 300.00 | 1.00 |
| #2 | CYP2D6 WT | 6x BTD | 300.00 | 1.44 |
| #3 | CYP2D6*53 | 20x APAP | 313.15 | 0.96 |
| #4 | CYP2D6*53 | 20x APAP | 313.15 | 0.70 |
| #5 | CYP2D6*53 | 20x APAP | 310.00 | 1.28 |
| #6 | CYP2D6*53 | 20x APAP | 313.15 | 1.20 |
| #7 | CYP2D6*53 | 20x APAP | 310.00 | 1.44 |
| #8 | CYP2D6 WT | 20x APAP | 313.15 | 1.44 |
| #9 | CYP2D6WT | 20x APAP | 313.15 | 1.92 |
| #10 | CYP2D6WT | 20x APAP | 313.15 | 1.92 |
| #11 | CYP2D6WT | 20x APAP | 310.00 | 0.72 |
| #12 | CYP2D6*53 | 20x BTD | 310.00 | 0.48 |
| #13 | CYP2D6*53 | 20x BTD | 310.00 | 0.48 |
| #14 | CYP2D6*53 | 20x BTD | 310.00 | 0.48 |
| #15 | CYP2D6*53 | 20x CZX | 310.00 | 0.48 |
| #16 | CYP2D6*53 | 20x CZX | 310.00 | 0.48 |
| #17 | CYP2D6*53 | 20x CZX | 310.00 | 0.48 |
| #18 | CYP2D6*53 | 20x DEB | 310.00 | 0.48 |
| #19 | CYP2D6*53 | 20x DEB | 310.00 | 0.48 |
| #20 | CYP2D6*53 | 20x DEB | 310.00 | 0.48 |
| #21 | CYP2D6*53 | 20x PPF | 310.00 | 0.48 |
| #22 | CYP2D6*53 | 20x PPF | 310.00 | 0.48 |
| #23 | CYP2D6*53 | 20x PPF | 310.00 | 0.48 |
| #24 | CYP2D6*53 | 20x APAP | 313.15 | 1.08 |

### Preference of ligands for protein, tunnels, and membrane

To determine hotspots of ligands on the protein surface, we developed a python script detecting the presence of ligand heavy atoms in the vicinity of the respective amino acid in a range of 5 Å. For glycine, we used the $\alpha$-carbon atom, while we chose the $\beta$-carbon atom for the remaining amino acids. For this calculation we used superimposed frames collected every 960 ps of the corresponding simulation. For APAP, simulations #3 to #11 were included, while simulations #13 and #14 were considered for BTD. The data was averaged for APAP and BTD and visualized on the surface of CYP2D6. The occupancy of the H1 site was determined for the phase of tunnel passage. Further, we divided the simulation box into three logical compartments to measure the preference of all ligands in the system for any of them. The first compartment consisted of the space not covered by the membrane (denoted as S), while the other two zones divided the membrane into head groups (H) and membrane core (M). We defined the head group region to be located between the mass center of the nitrogen atoms and the mass center of the C2 atoms the POPC molecules. Accordingly, we defined the membrane core to be located between the C2 atoms of the upper and lower POPC leaflets. An in-house python routine determined the location of the mass center of the ligand in z-direction and compared it to the boundaries of the three mentioned compartments. This analysis was performed for simulations #3 to #7 and #11 to #23. To normalize the compared time spans of the simulations, the interval between 200 and 480 ns of each simulation was considered for the analysis with frames being collected every 528 ps. Average values were calculated for every ligand. Since the included simulations of APAP were conducted at two different temperatures, the average results for APAP were divided in two groups (denoted as APAP-1 and APAP-2).

### Tunnel analysis

We used CAVER 3.0 to detect and characterize the tunnels in all simulations with a successful access event [14]. For that, we collected MD frames every 960 ps of the simulations, aligned them in the Protein Structure Alignment panel in Maestro, and determined the starting point for the tunnel computation using CAVER Analyst 1.0 [15]. We defined the starting point based on the residues E216, D301, and the heme for every simulation. We used a clustering threshold of 4.5, as it was determined to deliver good results in a previous study [3], while the rest of the settings were left on default. The nomenclature of the enzyme tunnels was adapted from Cojocaru and colleagues [16]. Average bottleneck radii, time-evolved bottleneck radii, and bottleneck residues were derived from the output of the tunnel computation.

### Ligand-protein and ligand-membrane energies

In the case of a successful access event, we determined the energy between the ligand and the protein with an in-house routine programmed in C++ language in MD frames in the MacroModel file format extracted at an interval of 480 ps. Interaction energies were calculated with a 12 Å cutoff from the ligand.

$$E_{ele} = \left(1 - \left(\frac{r}{12}\right)^2\right)^2 \frac{Q_1 Q_2}{4\pi\varepsilon_0 r} \; (I) \qquad E_{H-bond} = \left(\frac{C}{r_{ij}^{12}} - \frac{D}{r_{ij}^{10}}\right) \cdot cos^2(\theta_{Don-H \cdots Acc}) \; (II) \quad E_{hydrophobic} = \sum_{ij} E_{hydrophobic}^{ij} \; (III)$$

$$E_{hydrophobic}^{ij} = \begin{cases} 0.0 \;\; (1 \leq scale) \\ 0.25 \cdot scale^3 - 0.75 \cdot scale \\ 0.5 \;\; (-1.0 < scale < 1.0 \\ 1.0 \;\; (scale \leq -1.0) \end{cases} \quad (IV) \qquad scale = 2.0 \cdot \left(r_{ij} - r_i^{vdw} - r_j^{vdw} - 2.0\right)/3.0 \quad (V)$$

We evaluated partial contributions from electrostatics and hydrogen bonds according to the Yeti force field terms (Equation I and II) [17, 18]. The term for hydrogen bonds accounts for their directionality. The energies for hydrophobic contributions were calculated according to a term adapted from the VSGB 2.0 model (Equations III-V) [19, 20]. Membrane molecules were included in the analysis.

### Hydration analyses

We used an in-house routine to determine the number of the first-shell water molecules around the ligands in every frame of the simulations with a successful access event. The routine determined the number of water molecules in a given MD frame within a distance of 3.5 Å from any ligand atom. To determine the degree of active site desolvation in response to ligand binding, we extracted MD frames in a frame step of 480 ps covering a spherical zone of 15 Å around active site residues and the ligand of interest. Thereat, the residues 110, 112, 120, 121, 209, 212, 213, 216, 244, 247, 248, 297, 300, 301, 304, 305, 308, 309, 370, 443, 483, and 484 were included due to their proximity to the co-crystallized ligand in the underlying crystal structure (PDB ID code 3TDA). Frames with no ligand atoms detected in the binding site were included for calculating of the average number of waters in the empty state (unliganded). On the other hand, only frames with all ligand atoms present within the binding site were included for calculating of the average number of waters in the occupied state (liganded).

### Docking and pose comparison

The ligands for which an access event could be observed in the MD simulations were docked into the active site of CYP2D6. The 3D ligand structures of APAP and BTD were available from previous steps. Prior to docking, we used MGL Tools (v.1.5.6) [21] to prepare the receptor and ligand structures. We defined the search space to be cubic with a side length of 45 Å, manually changed the charge of the heme iron to $Fe^{2+}$ in the PDBQT file, and removed sodium atoms interfering with the calculation. Due to the high flexibility of CYP enzymes [22], including the potential structural adaptations related to ligand binding [16, 23, 24, 25], we considered several residues to be flexible for our docking calculations. In particular, we selected 112, 120, 211, 216, 221, 244, 296, 297, 301, 304, and 483 to be flexible. We used two different protein structures derived from simulation #3, differing in the orientation of APAP, as input structures. To enrich the results, docking runs were performed in the presence and absence of structural water molecules. Additionally, we performed docking runs with the amide bond of APAP regarded as flexible and rigid respectively. The docking was performed using AutoDock Vina (v.1.1.2) [26] with an exhaustiveness of 8. After docking, the obtained poses were filtered according to three criteria: (i) site of metabolism (SOM) in known range for metabolic reaction; (ii) ligand located within known binding site region; and (iii) docking score higher than -5.5 kcal/mol for the poses of APAP. In the case of BTD, only one pose derived from docking met the criteria.

To compare the poses from docking with ligand poses obtained from the MD simulations presenting successful ligand access, we chose the highest scored docking pose complying with the above-mentioned criteria to determine its similarity to MD poses. Therefore, we extracted all frames of the respective simulations, in which the ligand occupied the binding site, and aligned them to the corresponding docking pose using the Protein Structure Alignment tab within Maestro. Next, we removed all atoms from these frames except for the ligand of interest to ultimately compare the poses using the rmsd.py script provided by Schrodinger. Hydrogen atoms were excluded to compare the similarity between the poses.

## Supporting Results and Discussion

### Access of CYP2D6 ligands from the protein-membrane interface

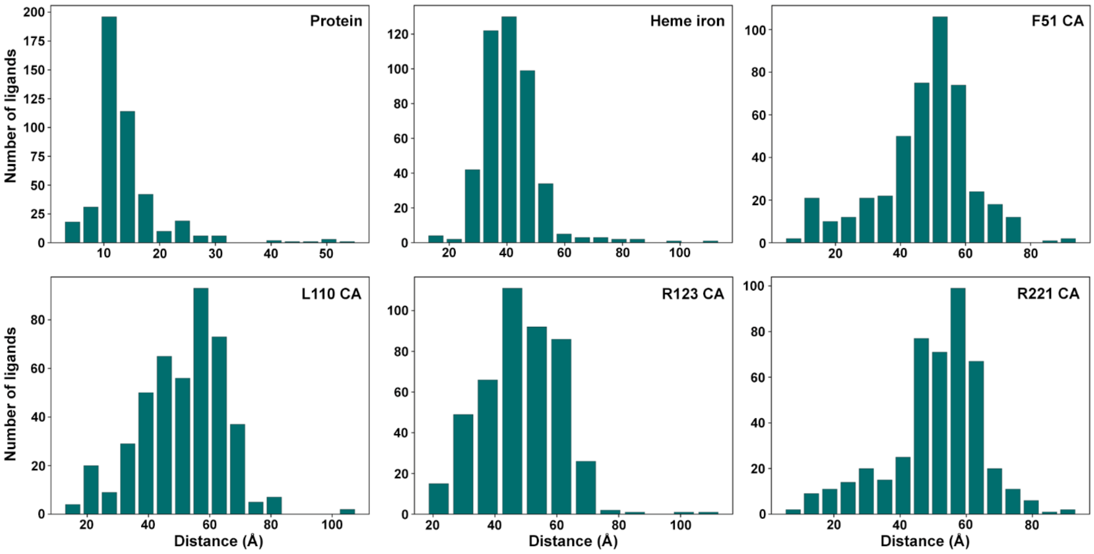


**Figure S 1** The distances from the ligand starting positions to selected protein atoms are shown. Only heavy atoms were considered. The plots were created in Matplotlib [27].

### Model validation

The root mean square deviation (RMSD) of the protein backbone indicated a good convergence of the systems besides minor drifts (Figure S2a). Except for one simulation, the values mostly remained between 2 and 3 Å. The RMSD diagram of simulation #3 presented several spikes that were caused by the movement of P267 as it was indicated by the high root mean square fluctuation (RMSF) of this particular residue located in the flexible GH loop on the protein surface (Figure S2b). A visual examination of the simulation revealed a reversible contraction of the loop after the ligand reached the active site, explaining the increased RMSD and RMSF values. However, the last frame of the simulation showed a value of 2.9 Å similar to the other simulations confirming convergence. In general, the RMSF diagrams indicated similar regions of local flexibility among the simulations that were in agreement with our previously published data [3]. The burying depth of the globular domain is used to validate and compare membrane-anchored models of CYPs [13]. The averages (Table S5) as well as the time-evolved values (Figure S2c) presented a narrow range around 38.5 Å comparable

**Table S 4** Minimal distances from the starting positions of accessing ligands to protein atoms. The minimal distances from the heavy atoms of the accessing ligand molecules to the heme iron and the next protein heavy atom are shown.

| *Simulation* | *Ligand* | *Ligand-iron (Å)* | *Ligand-Protein (Å)* |
|---|---|---|---|
| #3 | APAP-18 | 29.1 | 7.3 |
| #4 | APAP-7 | 84.0 | 55.3 |
| | APAP-18 | 18.2 | 2.9 |
| #5 | APAP-18 | 29.1 | 7.3 |
| #6 | APAP-6 | 36.2 | 12.0 |
| #7 | APAP-3 | 55.7 | 23.8 |
| | APAP-8 | 46.1 | 13.7 |
| #8 | APAP-20 | 36.3 | 13.0 |
| #13 | BTD-11 | 42.4 | 8.3 |
| #14 | BTD-3 | 35.4 | 12.4 |

[a] Since this was a replica simulation, this ligand started near the entrance of tunnel 2f.
[b] Simulation #5 was started from the same coordinates as simulation #3 at a different temperature.

to the literature value of 35±9 Å and our previous observations [3, 28]. On the other hand, the heme tilt angle, describing the angle between the z-axis and the plane of the porphyrin nitrogens of the heme, showed stronger fluctuations (Table S5 and Figure S2c). Nevertheless, the fluctuations were within the boundaries of 38-78° reported in the literature [29]. The placement of multiple ligands in a simulation system to study rare molecular events, as it was used in previous studies [30, 31], comes with advantages as well as disadvantages. Obviously, an advantage is the increased likeliness of observing a ligand accessing the enzyme, while a disadvantage is the potential influence of the ligands on each other. Even though molecules regularly contact each other in the crowded cellular environment [32], the comparably limited size of a simulation box (e.g. 104.6 x 128.5 x 191.5 $Å^3$ in simulation #3) could have potentially intensified such phenomena. Therefore, we determined the degree to which the accessing ligands contacted other ligand molecules during the different phases of the uptake process (Table S6). The percentage of frames, in which a heavy atom of the accessing ligand was within a radius of 5 Å from another heavy atom of a different ligand, was generally low for the BTD molecules. In the case of APAP, the values were scattered between 2.3 and 36.3%. High values were observed in the case of a dual ligand access, where the concurrent occupation of the active site naturally led to contacts between the two ligands. In simulations #5 and #8 however, we observed increased values despite only a single molecule accessing the enzyme. In the case of simulation #5, the contacts were low during the recognition and translocation phases. The high values, when the accessing ligand occupied the active site, were caused by an additional APAP molecule located on the surface of the enzyme in around 4.5 Å distance among their heavy atoms. In simulation #8, the contacts occurred during the recognition phase, when multiple ligand molecules formed transient agglomerates before APAP-20 initiated its translocation to the active site. In summary, the uptake process was not influenced by ligand contacts in our simulations with the exception of dual access events.

**Table S 5** Validation parameter average values and standard deviation.

| *Simulation* | *Heme tilt angle (°)* | *Burying depth (Å)* |
|---|---|---|
| #3 | 57.0 ± 8.5 | 38.9 ± 2.3 |
| #4 | 54.8 ± 6.3 | 38.8 ± 1.8 |
| #5 | 48.3 ± 8.0 | 38.1 ± 1.7 |
| #6 | 57.4 ± 8.0 | 38.4 ± 1.6 |
| #7 | 57.0 ± 5.8 | 38.5 ± 1.6 |
| #8 | 38.8 ± 7.0 | 38.0 ± 1.6 |
| #13 | 44.0 ± 7.0 | 38.6 ± 1.5 |
| #14 | 53.6 ± 5.4 | 39.4 ± 1.4 |

**Table S 6** Contacts of accessing molecules to other ligands.

| *Simulation* | *Ligand* | *Recognition* | *Translocation* | *Active site* |
|---|---|---|---|---|
| #3 | APAP-18 | 0<br>(n=84) | 7.8<br>(n=5771) | 0<br>(n=14145) |
| #4 | APAP-7 | 3.9<br>(n=7211) | 78.7<br>(n=3896) | 56.1<br>(n=3841) |
|  | APAP-18 | n/a [a] | 27.1<br>(n=14588) | n/a [a] |
| #5 | APAP-18 | 0.6<br>(n=313) | 7.7<br>(n=16792) | 35.1<br>(n=9562) |
| #6 | APAP-6 | 4.5<br>(n=12731) | 0<br>(n=1583) | 0<br>(n=10689) |
| #7 | APAP-3 | 14.4<br>(n=5646) | 0<br>(n=3895) | 47.5<br>(n=19627) |
|  | APAP-8 | 5.7<br>(n=12751) | 19.5<br>(n=8501) | 96.9<br>(n=7914) |
| #8 | APAP-20 | 36.8<br>(n=29573) | 0<br>(n=428) | n/a [b] |
| #13 | BTD-11 | 1.0<br>(n=605) | 0<br>(n=583) | 0<br>(n=8814) |
| #14 | BTD-3 | 13.6<br>(n=22) | 0<br>(n=62) | 0<br>(n=7271) |

The percentages of frames, in which the accessing molecules contacted surrounding ligands is divided into three phases. Contacts between heavy atoms in the range of 5 Å were considered. Together with the percentage, the number of frames in the respective interval is given.

[a] Since simulation #4 was a replica starting in the entrance of tunnel 2f, but did not reach the active site in a conformation that would allow a metabolic reaction.

[b] The ligand molecule did not reach a conformation in agreement with a metabolic reaction.

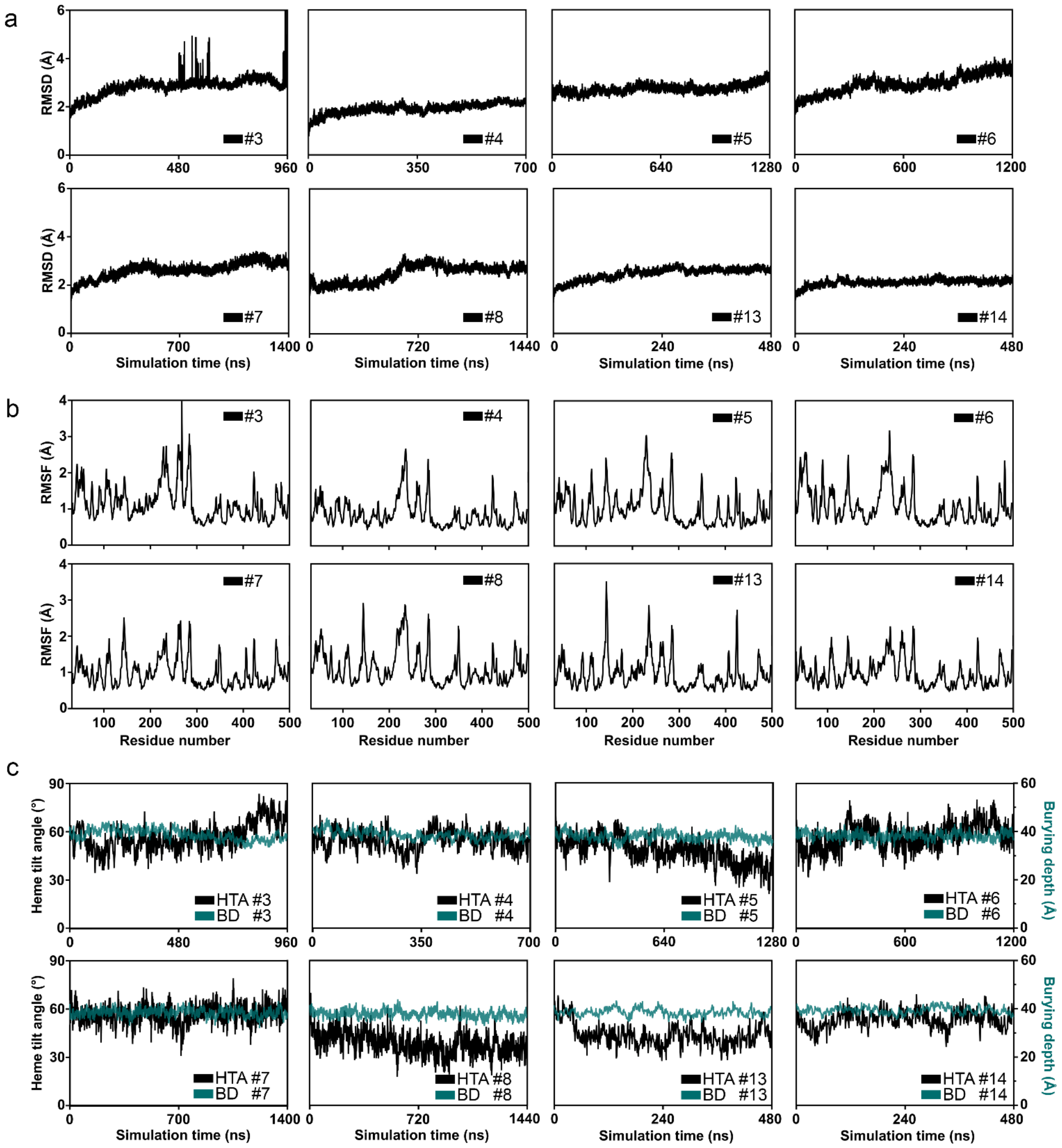


**Figure S 2** Validation of simulations with a successful access event. (a) The RMSD of all eight simulations presenting a successful access event is shown. The corresponding simulation identifier is indicated at the bottom right of the plots. (b) The RMSF of the simulations presenting a successful access event is shown. The corresponding simulation identifier is indicated at the top right of the plots. (c) The heme tilt angle (HTA) for simulations presenting a successful access event is shown together with the burying depth (BD) of the enzyme. The heme tilt angle is shown in black, while the burying depth is colored pine green.

## Preference of ligands for the protein, tunnels, and the membrane

**Table S 7** Occupancy of H1 allosteric site during ligand access.

| *Simulation* | *Ligand* | *Tunnel* | *Occupancy* |
|---|---|---|---|
| #3 | APAP-18 | 2f | yes |
| #4 | APAP-7 | 2f | yes |
| | APAP-18 | 2f | yes |
| #5 | APAP-18 | 4 | no |
| #6 | APAP-6 | 2b | no |
| #7 | APAP-3 | 2b | no |
| | APAP-8 | 2b | no |
| #8 | APAP-20 | 2f | yes |
| #13 | BTD-11 | 2c | no |
| #14 | BTD-3 | 2c | no |

Occupancy of H1 allosteric site during ligand access.The determined occupancy of the potential allosteric site H1 is shown together with the simulation identifier, the accessing ligand and the preferred tunnel.

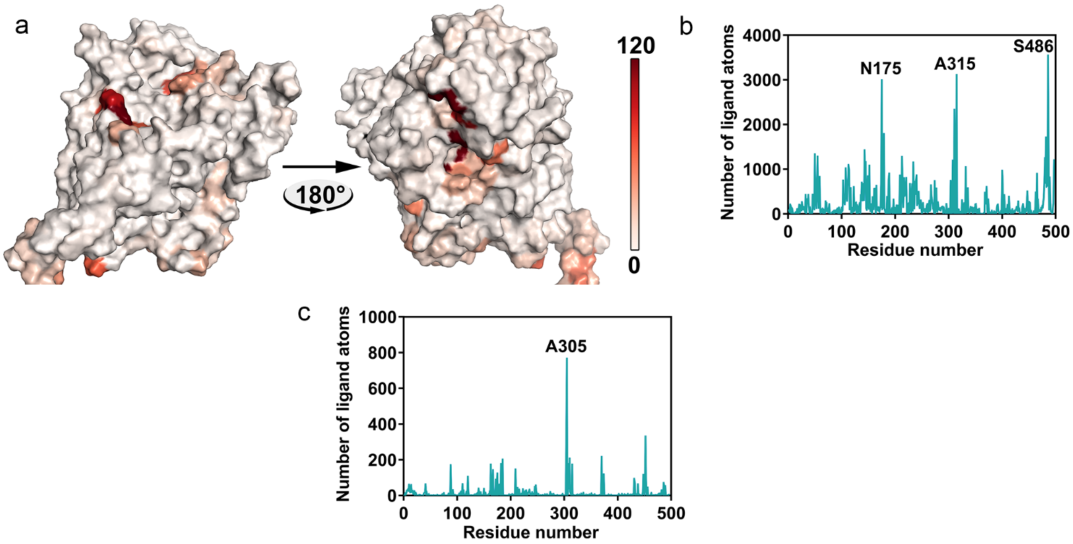


**Figure S 3** Ligand hotspots on the protein surface. (a) The visualization of ligand hotspots on the surface of CYP2D6 determined for BTD is shown. The scale from 0-120 describes the cumulative number of ligand heavy atoms in a 5 Å radius of the C$\beta$ atom (C$\alpha$ atom for glycine) of the protein amino acids. (b) A plot of the ligand hotspots of APAP for CYP2D6. The comparably intensive peaks of N175, A315, and S486 are indicated. (c) A plot of the ligand hotspots of BTD for CYP2D6. The comparably intensive peak of A305 is indicated.

## Structural adaptation of the protein

**Table S 8** Adaptation of the secondary structure of CYP2D6.

| *Simulation* | *Residues* | *Secondary structure* | *Ligand involvement* |
|---|---|---|---|
| #3 | V49-F58 | $\alpha$A' and $\alpha$A | directly |
| #4 | H48-L61 | $\alpha$A' and $\alpha$A | directly |
| #5 | V229-L236 | FG loop | directly |
|  | S288-N291 | HI loop | distant |
| #6 | H48-L61 | $\alpha$A' and $\alpha$A | distant |
| #7 | n/a | n/a | n/a |
| #8 | n/a | n/a | n/a |
| #13 | L110-S116 | BC loop | directly |
|  | K283-K288 | HI loop | distant |
| #14 | E280-S289 | BC loop and $\alpha$B | distant |

The structural adaptations of the protein secondary structure during ligand access are shown. The involvement of the ligand was visually determined and was classified to be either directly or distant. For simulations #7 and #8, no adaptations were observed.

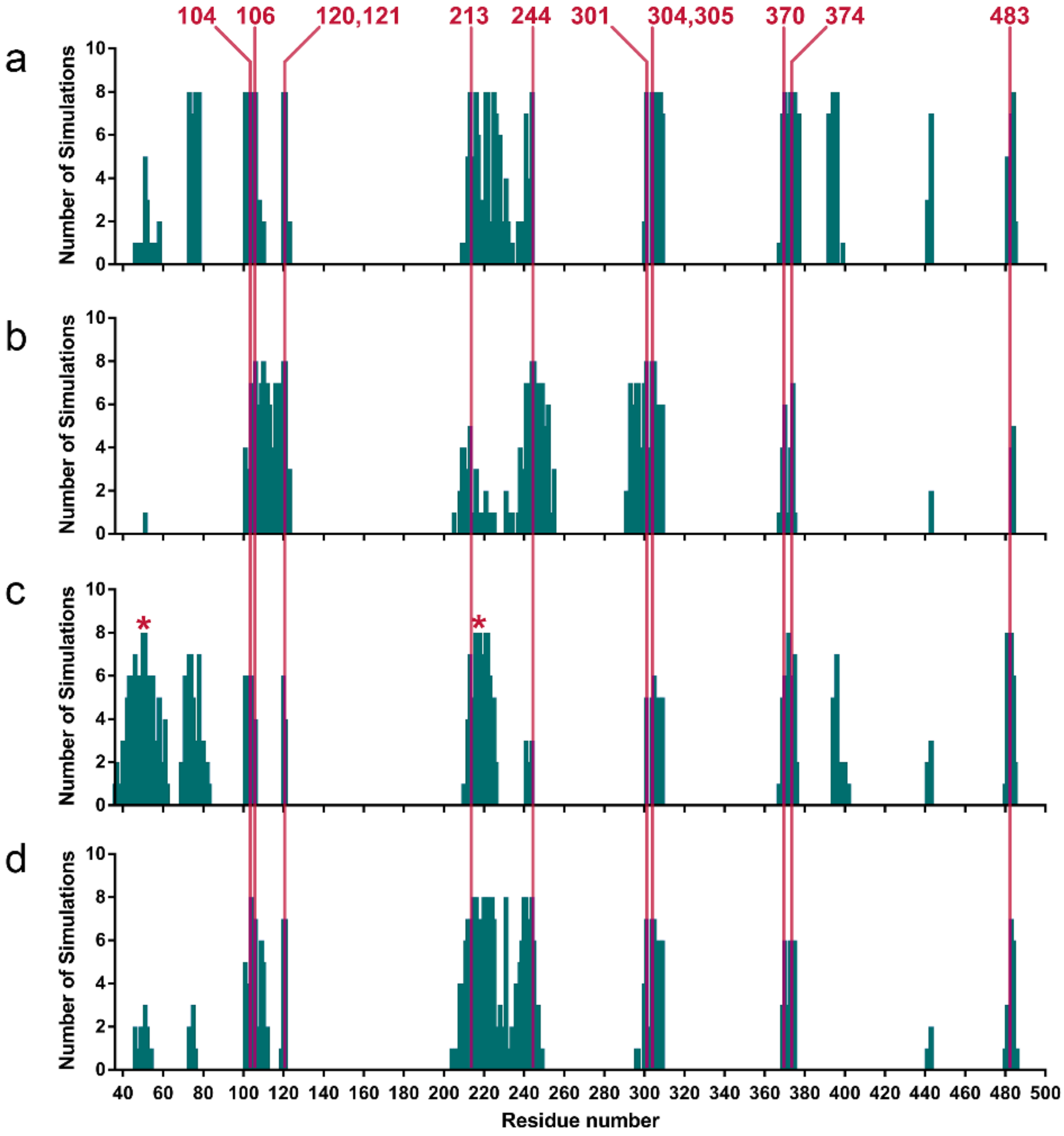


**Figure S 4** Most prominent bottleneck residues in CYP2D6.The major bottleneck residues for all simulations presenting a successful access event are shown with the number of simulations, in which the residues participated in bottlenecking the respective tunnel. Residues with high scores are indicated, while the gating residues F51 and F219 are pinpointed by red asterisks. The results are shown for (a) tunnel 2b, (b) tunnel 2c, (c) tunnel 2f, and (d) tunnel 4.

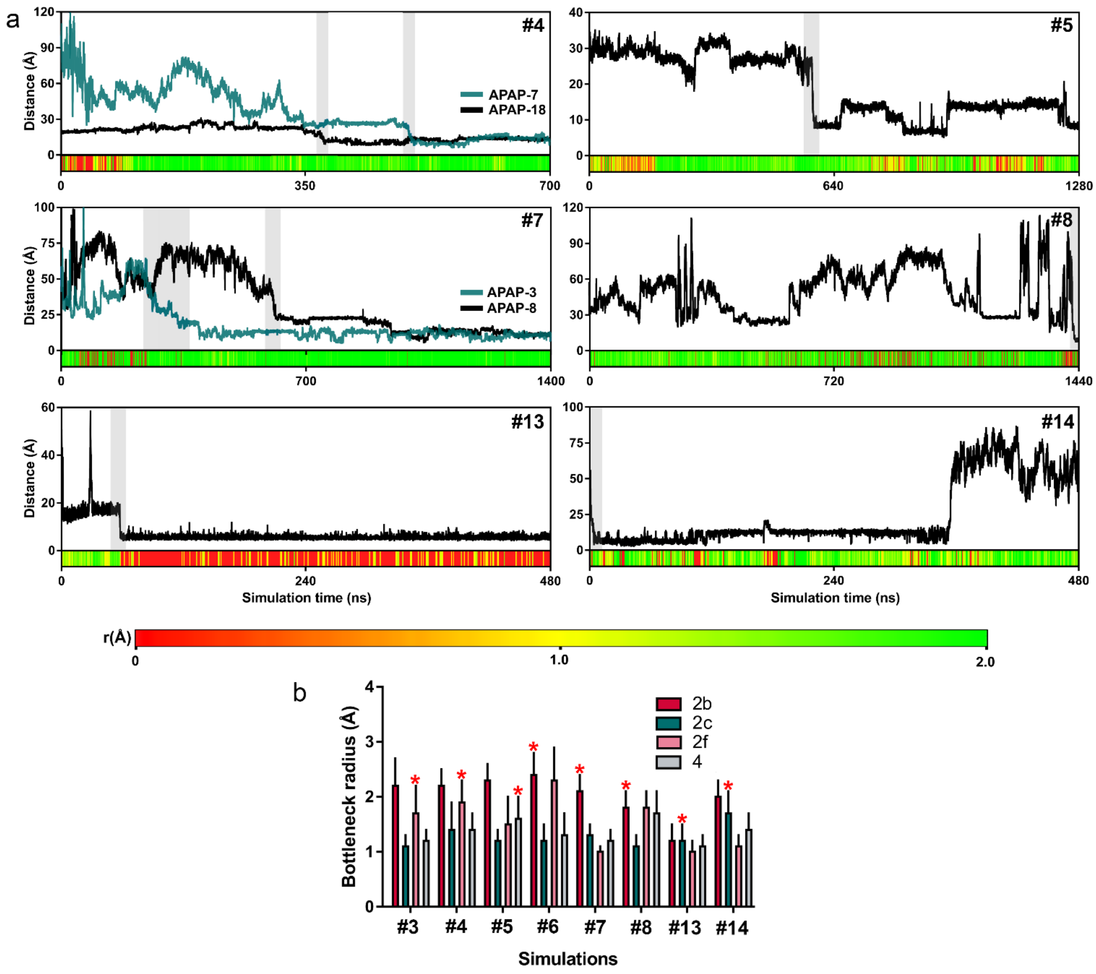


**Figure S 5** Bottleneck radii of enzyme tunnels.(a) The distance of the ligand SOM is plotted against the simulation time and the time-evolved bottleneck radius. The simulation identifiers are shown at the top right of the plots. Gray bars indicate the period of tunnel passage. The legend below indicates the coloring scheme for the bottleneck radii. (b) The average bottleneck radii for the simulations presenting a successful access event are shown. The tunnels, which were used by the ligand in the respective simulation are indicated by red asterisks. The values are shown with standard deviation.

**Table S 9** Characteristics of docking poses meeting the selection criteria.

| *Pose* | *Ligand* | *Protein* | *Score (kcal/mol)* | *Ligand flexibility* | *Water* |
|---|---|---|---|---|---|
| A1 | APAP | a | -8.1 | rigid | yes |
| A2 | APAP | a | -7.4 | flexible | yes |
| A3 | APAP | a | -6.8 | flexible | no |
| A4 | APAP | a | -6.7 | rigid | no |
| A5 | APAP | a | -6.8 | flexible | yes |
| A6 | APAP | a | -6.7 | rigid | no |
| A7 | APAP | b | -6.7 | flexible | yes |
| A8 | APAP | a | -6.5 | rigid | yes |
| A9 | APAP | a | -6.3 | rigid | yes |
| A10 | APAP | a | -6.2 | rigid | no |
| B1 | BTD | a | -4.4 | n/a | yes |

Overview of the predicted binding free energies (Score) for docking poses obtained for APAP and BTD with two different protein structures and different parameters regarding ligand flexibility and the presence of water molecules. Only poses matching the inclusion criteria (see SI Computational Methods) are shown. Note that only one pose of the docking calculations with BTD fulfilled the selection criteria and no special restraints were applied to the bonds of BTD.

**Table S 10** Comparison of poses from docking and MD simulations.

| *Simulation* | *Ligand* | *RMSD (Å)* |
|---|---|---|
| #3 | APAP-18 | 0.28 |
| #4 | APAP-7 | 4.96 |
|  | APAP-18 | 4.60 |
| #5 | APAP-18 | 3.57 |
| #6 | APAP-6 | 0.42 |
| #7 | APAP-3 | 4.49 |
|  | APAP-8 | 5.93 |
| #8 | APAP-20 | 4.07 |
| #13 | BTD-11 | 2.89 |
| #14 | BTD-3 | 1.79 |

The heavy atom RMSD between the selected docking pose and the closest resembling pose from MD simulations is shown.

### The driving forces for translocation

During the binding process to an active site, ligands generally experience some degree of strain [33] associated with a penalty toward binding affinity on the target. Our results did not reveal a clear trend for a reduced conformational freedom inside the enzyme based on the ligand torsion angles. Especially for BTD, such a result was to be expected since it is not able to fill the volume of enzyme tunnels or the active site to a similar degree as APAP due to its smaller size. In this case, our calculations were hampered by hardly comparable time intervals. Even when the number of frames inside and outside the enzyme were similar (Figure S6 D, G, I), the results were still inconclusive. While APAP-8 in simulation #7 showed a clear restriction in its conformational freedom inside the enzyme despite a higher number of frames for this period, the other ligands did not behave similarly. Frequently, the diversity of visited torsion angle values was rather altered than restricted, indicating a limited constraint on the ligand. Furthermore, the number of torsions in APAP and BTD is limited, potentially reducing the impact of ligand strain compared to a larger molecule such as the cocrystallized ligand prinomastat. The number of water molecules measured in the active site cavity was similar to observations other CYPs [34].

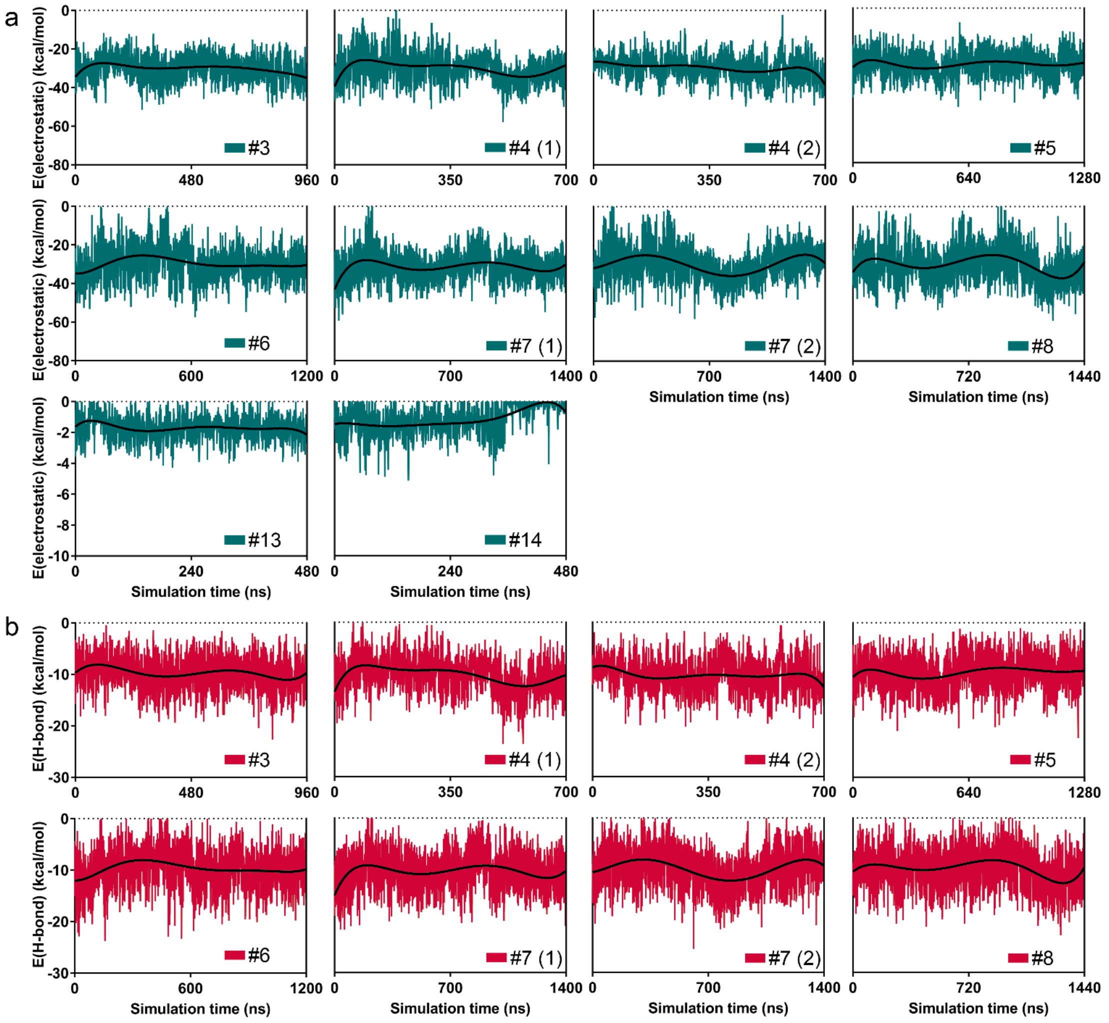


**Figure S 6** Energetic contributions from electrostatics and hydrogen bonds between the ligand and the protein. The energetic contributions of (a) electrostatics and (b) hydrogen bonds in simulations presenting a successful access event are shown. The simulation identifier is indicated at the bottom right. The regression lines show the centered sixth order polynomial fitted to the values. In the case of a double access event, the results were numbered sequentially after Table 1 in the main article.

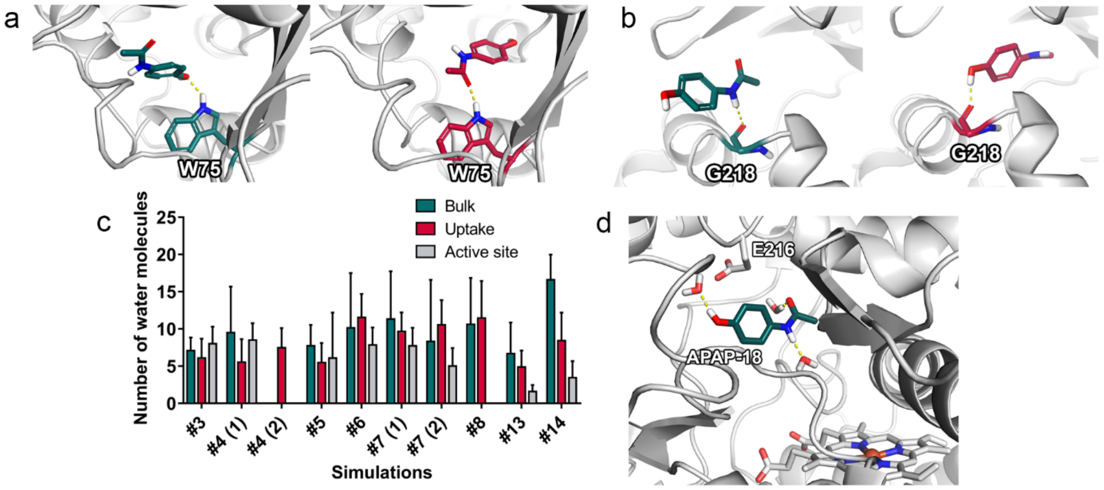


**Figure S 7** Polar interactions and hydration shell of the ligands. (a) The interaction between APAP and W75 shown at two different time points of simulation #4. (b) The interaction between APAP and G218 in simulation #4 shown at two different time points. (c) The number of water molecules in a 3.5 Å radius around the ligand are shown for the three phases of ligand uptake. In the case of a double access event, the data is sequentially shown for the ligands according to the order in Table 1. The values are presented with standard deviation. (d) The ligand is shown surrounded by water molecules in simulation #3. Note the hydrogen bonds between the ligand and the water molecules. For orientation, the location of E216 is shown at the top of the figure.

**Table S 11** Residues interacting with the ligand during the recognition process.

| *Simulation* | *Tunnel* | *Residues recognition* |
|---|---|---|
| #3 | 2f | F51, L46, V49, G218, F219 |
| #4 (1) | 2f | N45, L46, V49, D50, F51, F219 |
| #4 (2) | 2f | N45, L46, V49, F51, T54, L73, S217, G218, L372 |
| #5 | 4 | E215, L220, R221, R242, K245 |
| #6 | 2b | R25, R26, R123, K391 |
| #7 (1) | 2b | Q108, N225, L231, H232 |
| #7 (2) | 2b | R101, F112, Q117, L121, R123 |
| #8 | 2f | Q52, N53, K214, F481, A482 |
| #13 | 2c | L110, F112 |
| #14 | 2c | I106, I109, L110, L241 |

The residues that were determined to interact with the respective ligand are shown for the corresponding phase of the access event. In the case of a double access event, the data is sequentially shown for the ligands according to the order in Table 1.

**Table S 12** Residues interacting with the ligand during the recognition process.

| *Simulation* | *Tunnel* | *Residues translocation* |
|---|---|---|
| #3 | 2f | V49, F51, L73, W75, F219, V370, L372, T375, T394 |
| #4 (1) | 2f | V49, F51, Q52, T54, L73, W75, L213, E216, G218, R221, E222, V370, T375 |
| #4 (2) | 2f | L46, V49, F51, L73, W75, G218, F219, F243, F247, S304 |
| #5 | 4 | E216, L224, N225, V227, Q244, K245, F247, T375, F483 |
| #6 | 2b | H48, E216, E222, V370, T375, T394, F481 |
| #7 (1) | 2b | F51, T54, W75, P103, Q108, A209, E216, R221, E222, G373, T394, F483 |
| #7 (2) | 2b | F51, P103, R123, E216, E222, N225, E244, F247, D301, T394, F483 |
| #8 | 2f | F120, L213, E216, A305, A308, V370, T375 |
| #13 | 2c | L110, F112, I120, L121, I297 |
| #14 | 2c | L121, L241, A305, F483, L484 |

The residues that were determined to interact with the respective ligand are shown for the corresponding phase of the access event. In the case of a double access event, the data is sequentially shown for the ligands according to the order in Table 1.

**Table S 13** Residues interacting with the ligand during the recognition process.

| *Simulation* | *Tunnel* | *Residues active site* |
|---|---|---|
| #3 | 2f | R101, G367, I369, V370, V374, T375, F483 |
| #4 (1) | 2f | R101, L213, E216, S304, P371, V370, V374, F483 |
| #4 (2) | 2f | n/a |
| #5 | 4 | E216, F247, S304, F483 |
| #6 | 2b | I120, E216, S304, V370, V374, T375, F483 |
| #7 (1) | 2b | F247, D301, S304, T375 |
| #7 (2) | 2b | E216, F247, D301, S304, F483 |
| #8 | 2f | n/a |
| #13 | 2c | I120, L213, A305, V370, V374 |
| #14 | 2c | I120, L213, F243, F247, A305, V370, V374 |

The residues that were determined to interact with the respective ligand are shown for the corresponding phase of the access event. APAP-18 in simulation #4 as well as APAP-20 in simulation #8 did not adopt pose in the active site that is in accordance with a metabolic reaction. In the case of a double access event, the data is sequentially shown for the ligands according to the order in Table 1.

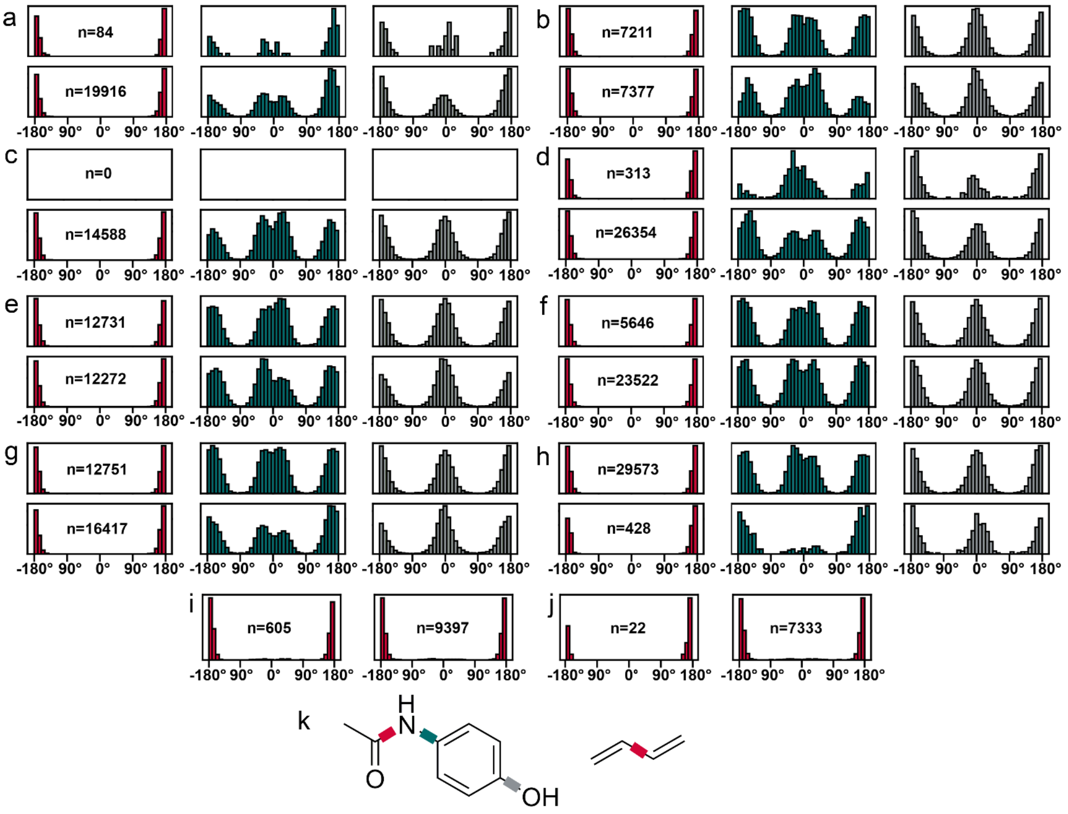


**Figure S 8** Distribution of torsion angle values of the accessing ligands inside and outside the enzyme. (a) For all simulations the values outside the enzyme (top) are compared with the ones where the ligand was inside the enzyme (bottom). The number of frames for the respective interval are described by n. The plots were generated using Matplotlib [27]. Here, the torsion angles of APAP-18 in simulation #3 are shown. (b) Torsion angles of APAP-7 in simulation #4. (c) Torsion angles of APAP-18 in simulation #4. (d) Torsion angles of APAP-18 in simulation #5. (e) Torsion angles of APAP-6 in simulation #6. (f) Torsion angles of APAP-3 in simulation #7. (g) Torsion angles of APAP-8 in simulation #7. (h) Torsion angle of APAP-20 in simulation #8. (i) Torsion angles of BTD-11 in simulation #13. (j) Torsion angles of BTD-3 in simulation #14. (k) The torsion angles of both ligands are shown with the corresponding color used to plot. While APAP is shown on the left side, BTD is shown on the right side. The molecular structures were created in ChemDraw13.

# References

[1] Kevin Bowers, Edmond Chow, Huafeng Xu, Ron Dror, Michael Eastwood, Brent Gregersen, John Klepeis, Istvan Kolossvary, Mark Moraes, Federico Sacerdoti, John Salmon, Yibing Shan, and David Shaw. Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters. *ACM/IEEE SC 2006 Conference (SC'06)*, (November):43, 2006.

[2] David E. Shaw, J. P. Grossman, Joseph A. Bank, Brannon Batson, J. Adam Butts, Jack C. Chao, Martin M. Deneroff, Ron O. Dror, Amos Even, Christopher H. Fenton, Anthony

Forte, Joseph Gagliardo, Gennette Gill, Brian Greskamp, C. Richard Ho, Douglas J. Ierardi, Lev Iserovich, Jeffrey S. Kuskin, Richard H. Larson, Timothy Layman, Li Siang Lee, Adam K. Lerer, Chester Li, Daniel Killebrew, Kenneth M. Mackenzie, Shark Yeuk Hai Mok, Mark A. Moraes, Rolf Mueller, Lawrence J. Nociolo, Jon L. Peticolas, Terry Quan, Daniel Ramot, John K. Salmon, Daniele P. Scarpazza, U. Ben Schafer, Naseer Siddique, Christopher W. Snyder, Jochen Spengler, Ping Tak Peter Tang, Michael Theobald, Horia Toma, Brian Towles, Benjamin Vitale, Stanley C. Wang, and Cliff Young. Anton 2: Raising the Bar for Performance and Programmability in a Special-Purpose Molecular Dynamics Supercomputer. *International Conference for High Performance Computing, Networking, Storage and Analysis, SC*, 2015-Janua(January):41–53, 2014.

[3] André Fischer, Charleen G. Don, and Martin Smieško. Molecular Dynamics Simulations Reveal Structural Differences among Allelic Variants of Membrane-Anchored Cytochrome P450 2D6. *Journal of Chemical Information and Modeling*, 58(9):1962–1975, 2018.

[4] Schrodinger LLC. The PyMOL Molecular Graphics System, Version 2.1.1. 2018.

[5] Slobodan Rendic. Summary of information on human CYP enzymes: Human P450 metabolism data. *Drug Metabolism Reviews*, 34(1-2):83–448, 2002.

[6] Sunghwan Kim, Paul A Thiessen, Evan E Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A Shoemaker, Jiyao Wang, Bo Yu, Jian Zhang, and Stephen H Bryant. PubChem Substance and Compound databases. *Nucleic acids research*, 44(D1):1202–13, 1 2016.

[7] Jeremy R. Greenwood, David Calkins, Arron P. Sullivan, and John C. Shelley. Towards the comprehensive, rapid, and accurate prediction of the favorable tautomeric states of drug-like molecules in aqueous solution. *Journal of Computer-Aided Molecular Design*, 24(6-7):591–604, 2010.

[8] New York NY Schrödinger, LLC and New York N Y Schrödinger LLC. Small-Molecule Drug Discovery Suite 2017-2, 2017.

[9] Oya Gürsoy and Martin Smieško. Searching for bioactive conformations of drug-like ligands with current force fields: How good are we? *Journal of Cheminformatics*, 9(1): 1–13, 2017.

[10] ChemAxon. Marvin (v.20.4.0), 2020.

[11] Andrea Gaedigk, Magnus Ingelman-Sundberg, Neil A. Miller, J. Steven Leeder, Michelle Whirl-Carrillo, and Teri E. Klein. The Pharmacogene Variation (PharmVar) Consortium: Incorporation of the Human Cytochrome P450 (CYP) Allele Nomenclature Database. *Clinical Pharmacology and Therapeutics*, 00(00):4–6, 2017.

[12] Shay McGuinness, Manoj Saxena, John Myburgh, Naomi Hammond, Steve Webb, Mark Holliday, Rinaldo Bellomo, Paul Young, Richard Beasley, Mark Weatherall, Colin McArthur, Diane Mackle, Frank van Haren, Ross Freebairn, and Seton Henderson. Acetaminophen for Fever in Critically Ill Patients with Suspected Infection. *New England Journal of Medicine*, 373(23):2215–2224, 2015.

[13] Lionel Ducassou, Laura Dhers, Gabriella Jonasson, Nicolas Pietrancosta, Jean Luc Boucher, Daniel Mansuy, and François André. Membrane-bound human orphan cytochrome P450 2U1: Sequence singularities, construction of a full 3D model, and substrate docking. *Biochimie*, 140:166–175, 2017.

[14] Eva Chovancova, Antonin Pavelka, Petr Benes, Ondrej Strnad, Jan Brezovsky, Barbora Kozlikova, Artur Gora, Vilem Sustr, Martin Klvana, Petr Medek, Lada Biedermannova, Jiri Sochor, and Jiri Damborsky. CAVER 3.0: A Tool for the Analysis of Transport Pathways in Dynamic Protein Structures. *PLoS Computational Biology*, 8(10):23–30, 2012.

[15] Barbora Kozlikova, Eva Sebestova, Vilem Sustr, Jan Brezovsky, Ondrej Strnad, Lukas Daniel, David Bednar, Antonin Pavelka, Martin Manak, Martin Bezdeka, Petr Benes, Matus Kotry, Artur Gora, Jiri Damborsky, and Jiri Sochor. CAVER Analyst 1.0: Graphic tool for interactive visualization and analysis of tunnels and channels in protein structures. *Bioinformatics*, 30(18):2684–2685, 2014.

[16] Vlad Cojocaru, Peter J. Winn, and Rebecca C. Wade. The ins and outs of cytochrome P450s. *Biochim Biophys Acta.*, 1770(3):390–401, 2007.

[17] Angelo Vedani, Max Dobler, Zhenquan Hu, and Martin Smieško. OpenVirtualToxLab-A platform for generating and exchanging in silico toxicity data. *Toxicology Letters*, 232(2): 519–532, 2015.

[18] Angelo Vedani and David W Huhta. A new force field for modeling metalloproteins. *Journal of the American Chemical Society*, 112(12):4759–4767, 6 1990.

[19] Jianing Li, Robert Abel, Kai Zhu, Yixiang Cao, Suwen Zhao, and Richard A. Friesner. The VSGB 2.0 model: A next generation energy model for high resolution protein structure modeling. *Proteins.*, 79(10):2794–2812, 2011.

[20] Marcel L Verdonk, Jason C Cole, Michael J Hartshorn, Christopher W Murray, and Richard D Taylor. Improved protein–ligand docking using GOLD. *Proteins: Structure, Function, and Bioinformatics*, 52(4):609–623, 9 2003.

[21] Gm Morris and Ruth Huey. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J Comput Chem.*, 30(16):2785–2791, 2009.

[22] Tereza Hendrychová, Eva Anzenbacherová, Jiří Hudeček, Josef Skopalík, Reinhard Lange, Peter Hildebrandt, Michal Otyepka, and Pavel Anzenbacher. Flexibility of human cytochrome P450 enzymes: Molecular dynamics and spectroscopy reveal important

function-related variations. *Biochimica et Biophysica Acta - Proteins and Proteomics*, 1814(1):58–68, 2011.

[23] T. L. Poulos. Cytochrome P450 flexibility. *Proceedings of the National Academy of Sciences*, 100(23):13121–13122, 2003.

[24] Peter J Winn, Susanna K Lüdemann, Ralph Gauges, Valère Lounnas, and Rebecca C Wade. Comparison of the dynamics of substrate access channels in three cytochrome P450s reveals different opening mechanisms and a novel functional role for a buried arginine. *Proceedings of the National Academy of Sciences of the United States of America*, 99(8): 5361–5366, 2002.

[25] Artur Gora, Jan Brezovsky, and Jiri Damborsky. Gates of enzymes. *Chemical Reviews*, 113(8):5871–5923, 2013.

[26] B. R. Brooks, C. L. Brooks III, Jr. A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M.W. I W I Hodoscek, and M. Karplus. AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *Journal of computational chemistry*, 30(10):1545–1614, 2009.

[27] J D Hunter. Matplotlib: A 2D Graphics Environment. *Computing in Science and Engineering*, 9(3):90–95, 2007.

[28] T. H. Bayburt and S. G. Sligar. Single-molecule height measurements on microsomal cytochrome P450 in nanometer-scale phospholipid bilayer disks. *Proceedings of the National Academy of Sciences*, 99(10):6725–6730, 2002.

[29] Yoshihiro Ohta, Suguru Kawato, Hiroko Tagashira, Shigeki Takemori, and Shiro Kominami. Dynamic Structures of Adrenocortical Cytochrome P-450 in Proteoliposomes and Microsomes: Protein Rotation Study. *Biochemistry*, 31(50):12680–12687, 1992.

[30] Yibing Shan, Eric T. Kim, Michael P. Eastwood, Ron O. Dror, Markus A. Seeliger, and David E. Shaw. How does a drug molecule find its target binding site? *Journal of the American Chemical Society*, 133(24):9181–9183, 2011.

[31] R. O. Dror, A. C. Pan, D. H. Arlow, D. W. Borhani, P. Maragakis, Y. Shan, H. Xu, and D. E. Shaw. Pathway and mechanism of drug binding to G-protein-coupled receptors. *Proceedings of the National Academy of Sciences*, 108(32):13118–13123, 2011.

[32] Stephen Smith, Claudia Cianci, and Ramon Grima. Macromolecular crowding directs the motion of small molecules inside cells. *Journal of the Royal Society, Interface*, 14(131): 20170047, 6 2017.

[33] David L Mobley and Ken A Dill. Binding of small-molecule ligands to proteins: ”what you see” is not always ”what you get”. *Structure (London, England : 1993)*, 17(4):489–498, 4 2009.

[34] Patrik Rydberg, Thomas H Rod, Lars Olsen, and Ulf Ryde. Dynamics of water molecules in the active-site cavity of human cytochromes P450. *Journal of Physical Chemistry B*, 111(19):5445–5457, 2007.

# CHAPTER 3

# A Conserved Allosteric Site on Drug-Metabolizing CYPs: A Systematic Computational Assessment

In the course of the study highlighted in Chapter 2, the potential influence of a superficial allosteric site in CYP2D6 on the ligand access process was raised in accordance to work published on a bacterial CYP. The study presented in this chapter systematically examines this allosteric site in the nine most relevant drug-metabolizing enzymes with a multi-scale computational modeling approach. The allosteric regulation of protein function is an integral component of molecular recognition.

**Author contributions:** Conceptualization, A.F. and M.S.; methodology, A.F.; formal analysis, A.F.; writing and original draft preparation, A.F.; writing, review and editing, A.F., M.S.; visualization, A.F.; supervision, M.S.

*Based on a manuscript submitted to Int. J. Mol. Sci.:*

Fischer, A.; Smieško, M. A Conserved Allosteric Site on Drug-Metabolizing CYPs: A Systematic Computational Assessment

## Abstract

Cytochrome P450 enzymes (CYPs) are the largest group of enzymes involved in human drug metabolism. Ligand tunnels connect their active site buried at the core of the membrane-anchored protein to the surrounding solvent environment. Recently, evidence of a superficial allosteric site, here denoted as hotspot 1 (H1), involved in the regulation of ligand access in a soluble prokaryotic CYP emerged. Here, we applied multi-scale computational modeling techniques to study the conservation and functionality of this allosteric site in the nine most relevant mammalian CYPs responsible for approximately 70% of drug metabolism. In total, we systematically analyzed over 44 $\mu$s of trajectories from conventional MD, cosolvent MD, and metadynamics simulations. Our bioinformatics analysis and simulations with organic probe molecules revealed the site to be well conserved in the CYP2 family with the exception of CYP2E1. In the presence of a ligand bound to the H1 site, we could observe an enlargement of a ligand tunnel in several members of the CYP2 family. Further, we could detect the facilitation of ligand translocation by H1 interactions with statistical significance in CYP2C8 and CYP2D6, even though all other enzymes except for CYP2C19, CYP2E1, and CYP3A4 presented a similar trend. As the detailed comprehension of ligand access and egress phenomena remains one of the most relevant challenges in the field, this work contributes to its elucidation, and ultimately, helps in estimating the selectivity of metabolic transformations using computational techniques.

## Introduction

Cytochrome P450 enzymes (CYPs) are the most relevant class of enzymes responsible for the biotransformation of approximately 70-80% marketed of drugs. CYPs can catalyze a range of oxidative and reductive reactions including hydroxylation, heteroatom oxygenation, dealkylation, and epoxidation with a distinct substrate specificity [1, 2]. Clinical complications related to metabolism can occur due to potential drug-drug interactions and interindividual differences resulting from genetic polymorphism, both altering drug elimination [1, 3, 4]. The prediction of CYP metabolism, both in regard to ligand selectivity and catalytic efficiency, is of pivotal importance for rational design as a large share of drug attrition is caused by a poor pharmacokinetic profile [5, 6].

Even though differences in active sites residues partially allow to explain the complex substrate specificity of CYPs, other structural mechanisms such as ligand access and egress have been evidenced to influence it [7, 8, 9, 10]. In particular, the active site of CYPs is buried within the core of the protein and is connected to its surrounding environment by dynamic tunnels. The most narrow region of such a tunnel is referred to as bottleneck, where gating residues act as molecular filters for compounds accessing the enzyme. The opening of these tunnels depends on conformational changes of the protein. As mammalian CYPs are membrane-anchored proteins (Figure 1A), it is thought that lipophilic ligands access the binding site through membrane-facing tunnels, while hydrophilic compounds, including the metabolic products, prefer tunnels reaching the bulk solvent [10, 11, 8, 12]. In general, experimental methods can only provide limited insight into such ligand translocation processes [8] as, for example, the tunnels are often not apparent in static crystal structures. However, due to their dynamic nature, computational methods such as molecular dynamics (MD) simulations introducing a natural degree of structural flexibility have been widely applied to study ligand tunnels in atomic detail [8, 13, 10, 14, 15].

Allosteric regulation of enzymatic activity is a powerful mechanism for cells to adapt to their cellular environment by propagating information between two distinct sites of the protein [16, 17]. Only recently, a superficial allosteric site among helices C, E, and H (Figure 1B) was proposed to be involved in the regulation of substrate access to the soluble prokaryotic enzyme CYP101A1. Upon ligand association at this site, the enzyme was described to shift from a closed to an open conformation, facilitating the access of other ligands through a tunnel located between the F-G loop and the B-C loop [14]. Especially in the presence of a high substrate concentration, allosteric regulation may allow adaption to the environment by enhancing metabolic activity, thus functioning as a protective mechanism to facilitate the excretion of large amounts of potentially toxic xenobiotics. In our recent work focused on CYP2D6 ligand access, we could verify the association of small-molecules at this site, which we denoted as hotspot 1 (H1). Similar to the above-mentioned study by Follmer and colleagues, we examined ligand access and observed a considerable correlation between the occupancy of the H1 site and complete ligand translocation to the active site [8].

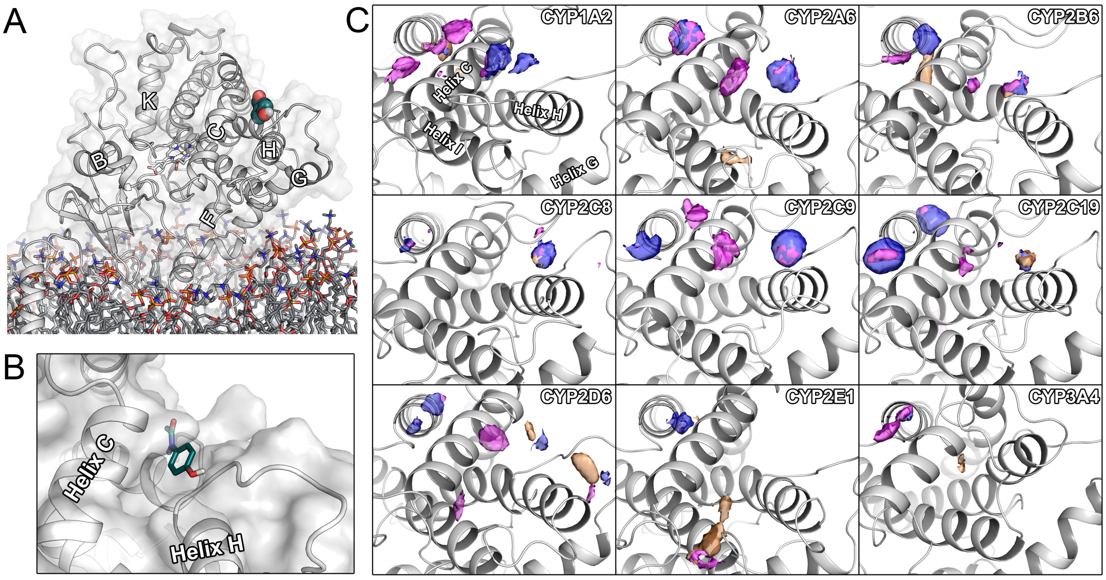


**Figure 1** Structural overview and cosolvent densities. (A) Structure of membrane-anchored CYP2D6 with a ligand bound to the H1 site (shown as spheres). For orientation, helices have been denoted with letters. (B) Close-up view of the H1 allosteric site in CYP2D6 with bound ligand acetaminophen. (C) Probe densities obtained from cosolvent MD simulations. Pink densities correspond to pyridine, blue ones to isopropanol, and orange ones to acetonitrile.

Intrigued from previous findings regarding the H1 site, this work was focused on elucidating its functionality in a panel of nine mammalian drug-metabolizing CYPs covering the majority of phase I metabolism. We systematically applied bioinformatic tools, cosolvent simulations, conventional MD (cMD) simulations, and metadynamics simulations to study in detail CYP1A2, CYP2A6, CYP2B6, CYP2C8, CYP2C9, CYP2C19, CYP2D6, CYP2E1, and CYP3A4 enzymes. We examined conservation, small-molecule association, conformational adaptation, and the facilitation of ligand translocation. Our results indicate that the site is mostly conserved among the CYP2 family. Additionally, the association of organic probes and small molecules revealed a nearby site separated from H1 by helix C. Ultimately, metadynamics simulations indicated the occupation of H1 site to facilitate ligand egress in most systems, with statistically significant differences in CYP2C8 and CYP2D6. Our work improves the understanding of the allosteric regulation of ligand access to drug-metabolizing CYPs, a process closely related to their substrate specificity and relevant for the accurate prediction of metabolic outcomes.

## Results and Discussion

**Simulation techniques and model validation.** By introducing flexibility into molecular systems, MD simulations find many different applications such as studying time-evolved ligand-protein interactions, conformational changes in protein structures, assessing the role of solvent molecules, or identifying putative binding sites [18, 19, 20]. Here, we applied various MD techniques including cMD simulations, metadynamics, and cosolvent simulations to study the most relevant drug-metabolizing CYPs. By introducing molecular probes to a simulation system, one can observe their association with the protein of interest and highlight potential binding sites [18]. Based on these simulations, we could deduce different pharmacophores of the H1 site. Due to the short timescale of these simulations (Table 1) preventing from significant structural changes, we modeled the soluble protein without the membrane. Similar to the cosolvent simulations with small organic probes, we evaluated the association of small drug-like compounds with the proteins. Next, in order to investigate the effect of allosteric ligands bound to the H1 site proposed by Follmer and colleagues [14], we placed 20 ligands around the protein and observed their association with the H1 site. The root mean-square deviation (RMSD) indicated good convergence of the respective simulations (Figure S1). As mentioned above, mammalian drug-metabolizing CYPs are membrane anchored enzymes. The structural regions of CYPs involved in forming several well-characterized access tunnels are in direct contact with head groups of the membrane molecules [8, 21, 10, 11]. Thus, to accurately study the opening and closing of ligand tunnels, we constructed membrane models for each enzyme by adding membrane anchors (as detailed in the Materials and Method section) and embedding the protein in a preequilibrated 1-palmitoyl-2-oleoylphosphatidylcholine (POPC) membrane. The orientation and embedding of the protein was predicted using the well-established PPM server, similar to our previous work [8, 21]. The resulting systems were subjected to a 300 ns equilibration simulation, during which we computed two metrics commonly used to characterize and validate models of membrane-anchored CYPs. The heme tilt angle is defined as the angle of the plane of the porphyrin nitrogens of the heme moiety and the membrane normal corresponding to the z-axis of the system [21, 11]. Based on rotational diffusion measurements, the heme tilt angle for CYPs was determined to be in the range of 38-78° [22]. The average values determined in the equilibration simulations ranged between 53.1 and 73.3° in agreement with the experimentally determined boundaries (Table S1). Further, the angle remained relatively stable around the starting value, indicating that there were no large changes in respect to the input structures (Figure S2). When compared to previous results by Berka and colleagues [11], the values were highly similar for CYP1A2, CYP2C9, and CYP2E1, while there were considerable differences for CYP2A6, CYP2D6, and CYP3A4. However, when we compared the results of CYP2D6 to our previous work [8], the observed angles were again highly similar. Another parameter we monitored, was the burying depth of the globular domain of each enzyme, which was experimentally determined to be 35±9 Å for CYP2B4 by atomic force microscopy [23]. In our equilibration simulations, we observed burying depths between 35.7 and 38.9 Å in agreement with experimental observations. In addition to the heme tilt angle and the burying depth, we assessed the backbone RMSD of the simulations, which indicated acceptable convergence (Figure S3).

| *Type* | *Membrane* | *Duration (ns)* | *Replicas*[a] | $Lig_{Ortho}{}^{b}$ | $Lig_{Allo}{}^{b}$ |
|---|---|---|---|---|---|
| Cosolvent | no | 60 | 10 | no | no |
| Association | no | 500 | 1 | no | yes[c] |
| Equilibration | yes | 300 | 1 | yes | no |
| Sampling | yes | 1005 | 3 | no | yes |
| Metadynamics | yes | 50 | 10 | yes | no |
| Metadynamics | yes | 50 | 10 | yes | yes |

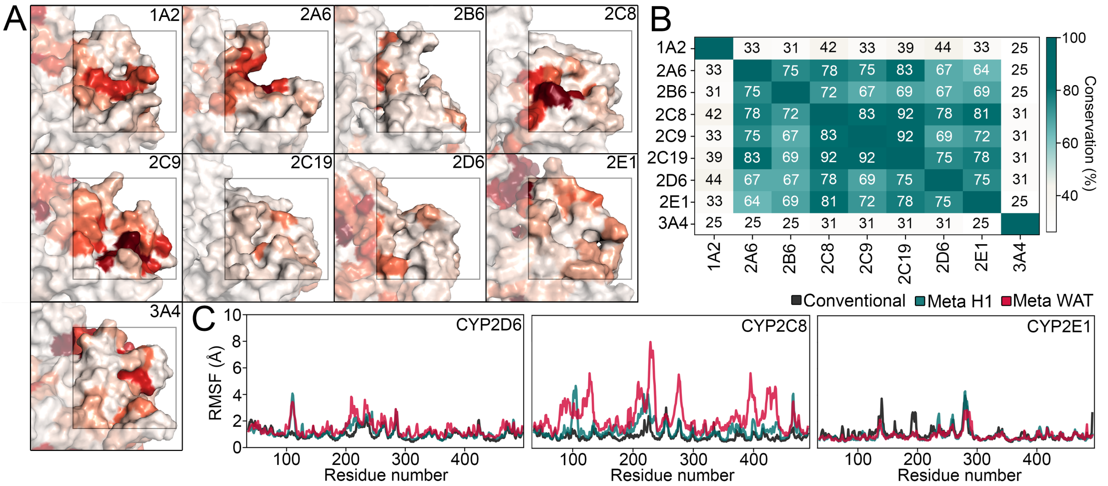
**Table 1** Simulations conducted in this work. [a] Number of replicas per enzyme. [b] Indicator if a ligand was present at the respective site. [c] Transient presence of ligand.

Using the validated membrane models, we conducted microsecond simulations with an allosteric ligand bound and, after 5 ns of unrestrained simulations, restrained to the H1 site. In these simulations, we observed high RMSD values for CYP1A2, CYP2A6, CYP2C19, and CYP2D6 (Figure S4). Upon inspection of the trajectories, we could observe large structural changes of the membrane anchor, while the globular domain and the overall fold of the protein remained stable (Figure S5). In a last step, we conducted metadynamics simulations focused on studying ligand egress from the buried binding site with and without the presence of an allosteric ligand bound to H1. Due to the conformational changes imposed by the translocating ligand, one replica simulation of

CYP2C8 as well as multiple simulations of CYP1A2 presented increased values compared to the remaining enzymes, which showed acceptable RMSD values below 4 Å (Figures S6 and S7). However, high values were to be expected due to the application of biasing potentials to the systems and the propagating ligand.

**Small molecules regularly associate with the H1 site of several CYPs.** As mentioned in the previous section, cosolvent MD simulations can be applied to map binding sites of a protein by introducing small organic probes. Here, we conducted simulations with acetonitrile, isopropanol, and pyridine to compare the obtained densities among the panel of CYPs (Figure 1C). We detected considerable density of isopropanol and pyridine probes at the H1 site of all studied CYPs besides CYP2E1 and CYP3A4. In CYP2B6, CYP2C8, CYP2C19, CYP2D6, and CYP3A4, there were densities of acetonitrile, even though they were less pronounced. Interestingly, only the densities of pyridine presented an overlap between the two highly similar (regarding sequence) enzymes CYP2C9 and CYP2C19. Instead, CYP2C9 was more similar to CYP2A6 with an isolated density of of pyridine close to the center of helix C as well as a shared density of pyridine and isopropanol between the center of helix C and the C-terminal of helix H. In all enzymes besides CYP2E1 and CYP3A4, this distribution of densities was present. Besides the densities at the H1 site, we could detect the association of probes on a neighboring site, which is separated from H1 by helix C. At the given isovalue of 15, every CYP studied here presented a density of at least one probe in this region. However, this adjacent density was comparatively small in CYP2C8 and CYP2E1. In CYP2C19, the density of the neighboring site was even larger than the ones detected at the H1 site.

Previously, the association of the small-molecules camphor and acetaminophen were reported in the literature as well as our previous work [8, 14]. Additional evidence for the association of small-molecules at the H1 site was provided by X-ray crystallography of CYP101A1 cocrystallized in excess of camphor. The resulting structure revealed a slight electron density of camphor at the H1 site [14, 24]. Thus, we decided to study the association of drug-like molecules with the H1 site in addition to the small organic probes in cosolvent simulations. CYP1A2, CYP2A6, CYP2C8, CYP2C9 presented a high degree of ligand association at the H1 site (Figure 2A). In analogy to the cosolvent simulations, only minor association of ligands at the H1 site was seen at CYP2E1 and CYP3A4, indicating a missing functionality of this site in the context of regulation by small molecules. In contrast to our previous findings [8], CYP2D6 only showed limited association of ligands with the H1 site, but rather with the neighboring site separated from H1 by helix C. Generally, the hotspots roughly overlapped with the regions showing increased probe densities in cosolvent simulations.



| | 1A2 | 2A6 | 2B6 | 2C8 | 2C9 | 2C19 | 2D6 | 2E1 | 3A4 |
|---|---|---|---|---|---|---|---|---|---|
| 1A2 | | 33 | 31 | 42 | 33 | 39 | 44 | 33 | 25 |
| 2A6 | 33 | | 75 | 78 | 75 | 83 | 67 | 64 | 25 |
| 2B6 | 31 | 75 | | 72 | 67 | 69 | 67 | 69 | 25 |
| 2C8 | 42 | 78 | 72 | | 83 | 92 | 78 | 81 | 31 |
| 2C9 | 33 | 75 | 67 | 83 | | 92 | 69 | 72 | 31 |
| 2C19 | 39 | 83 | 69 | 92 | 92 | | 75 | 78 | 31 |
| 2D6 | 44 | 67 | 67 | 78 | 69 | 75 | | 75 | 31 |
| 2E1 | 33 | 64 | 69 | 81 | 72 | 78 | 75 | | 25 |
| 3A4 | 25 | 25 | 25 | 31 | 31 | 31 | 31 | 25 | |

**Figure 2** (A) Small-molecule binding hotspots on the surface of the nine enzymes studied in this work. Different shades of red indicate the degree of ligand association per residues as described in the Materials and Methods section. (B) Heatmap depicting the conservation of the H1 site within the panel of CYPs. (C) Analysis of structural changes observed in metdynamics simulations.

To assess the conservation of a protein region, sequence alignments followed by the comparison of the respective residues is a commonly used technique [25]. We observed a high conservation of the site in the CYP2 family, especially among CYP2A6, CYP2B6, CYP2C8, and CYP2C19 (Figure 2B). Especially in the CYP2C subfamily, the identity among the studied members was above 83%. As it is implied by the nomenclature [2], CYP3A4 shared less sequence identity with the CYP2 family. Again, this stands in accordance with the results from the MD simulations discussed above.

**The effect of allosteric ligands bound to the H1 site is isoform-dependent.** As mentioned above, Follmer and colleagues observed a shift to an open conformation in response to a ligand bound to the H1 site of CYP101A1. To translate those results to mammalian CYPs, we used our membrane-anchored CYP models and restrained a ligand to the H1 site to ensure constant occupancy of the allosteric site. In the microsecond simulations, we monitored the conformational state of the enzyme, as well as the three most relevant ligand tunnels described in the literature [8, 14, 10, 26, 11, 15]. These tunnels included tunnel 2b located among the B-C loop, the F-G loop, and the $\beta 4$ sheet, tunnel 2f located between helix A and the F-G loop, as well as tunnel 2c between the N-terminus of helix I and the B-C loop [8, 15, 21]. After processing MD frames from the simulations using CAVER [27], we determined the average bottleneck radii of the three tunnels and statistically compared them to the values obtained from the final frames of the membrane equilibration simulations (Tables 2, S2 and S3). Whereas there was no clear trend for an enlargement of tunnels 2b and 2c, enzymes of the CYP2 family including CYP2B6, CYP2C8, CYP2C9, and CYP2C19 presented a statistically significant increase in the bottleneck radius of tunnel 2f (Figures S8-S10). In the work on CYP101A1, the F-G loop was described as a key regulatory structure for the opening of tunnels [14]. As this loop separates tunnel 2f and 2b, a slight movement away from the tunnel entrances can lead to the merging of these tunnels. In CYP2B6, we could observe an enlargement of both tunnel 2b and 2f, potentially connected to the increase in the root mean-square fluctuation (RMSF) of two replica simulations between residues 220 and 250 (Figure S11). Indeed, the other enzymes of the CYP2 family sharing an increased opening of tunnel 2f if an allosteric ligand was present exhibited a similar behavior in the RMSF diagrams. Thus, rearrangements of the helices F and G, as well as the loop connecting them, are likely responsible for the observed behavior. CYP1A2, CYP2C19, and CYP2D6 presented larger bottleneck radii for tunnel 2c, indicating isoform-dependent behavior of the gating mechanisms, as described in the literature [15]. In CYP3A4, we could not detect any enlargement of these tunnels in response to a ligand bound to the H1 site.

The radius of gyration can be used to evaluate the compactness of a protein structure, and thus, to analyze the conformational state of protein [28, 29]. To monitor the conformational state of the enzymes during these simulations and detect a potential shift towards an open conformation, we computed the radius of gyration for each replica (Figure S12). However, the values remained stable throughout all simulations, indicating no large deviations from the closed, compact conformation observed in crystal structures. Thus, an allosteric modulation by ligand bound to the surface hotspot seems to trigger dynamical conformation effects affecting only the tunnel microenvironment, with no apparent reach on the overall protein fold.

In a next step, we aimed to study the translocation of ligands through the identified pathways under the influence of a ligand bound to the H1 site and compared the results to systems with no such additional ligand. Generally, products of CYP-mediated metabolic transformations are more hydrophilic and, thus, it was proposed that products might egress from solvent-exposed tunnels as opposed to substrates, which prefer tunnels directed toward the membrane to reach the binding pocket [10, 15]. Thus, the selection process between access and egress tunnels is thought to be dictated by the physicochemical properties of the ligand. To avoid biasing the route selected by the ligand by placing it near an entrance at the outer surface of the enzyme, we decided to study the ligand dissociation from the binding pocket, which constitutes the inverse process based on the current rationale. Ligand unbinding is a comparatively slow process that requires substantial simulation times and represents a so-called rare molecular event. There are several methods available that enhance the sampling of such rare events by applying various biasing potentials on top of the regular force field. These techniques include steered MD, random-accelerated MD, umbrella sampling, protein energy landscape exploration, accelerated MD, as well as metadynamics simulations [9, 26, 30, 31, 32, 33]. In the metadynamics protocol, Gaussian potentials are applied toward collective variables (CVs) defining the reaction coordinate of such a rare event [31]. Here, we selected the distance of the ligand center of mass to the heme iron atom within the active site as CV. We conducted twenty replica simulations per enzyme with half of them having a ligand bound to the H1 site (Table 1). Except two simulations without allosteric ligand in CYP2C8, the ligands completely egressed from the binding pocket in all simulations. In analogy to the above-described cMD simulations, we compared the RMSF values of the metadynamics simulations to the ones obtained from the membrane equilibration (Figures 2C and S13). Generally, the values were higher in the metadynamics simulations, especially if no allosteric ligand was bound to the H1 site. The most significant changes could be detected in the region of the F-G loop, as well as the C-terminus of helix B.

To relate the results from the metadynamics simulations to the bottleneck radii obtained

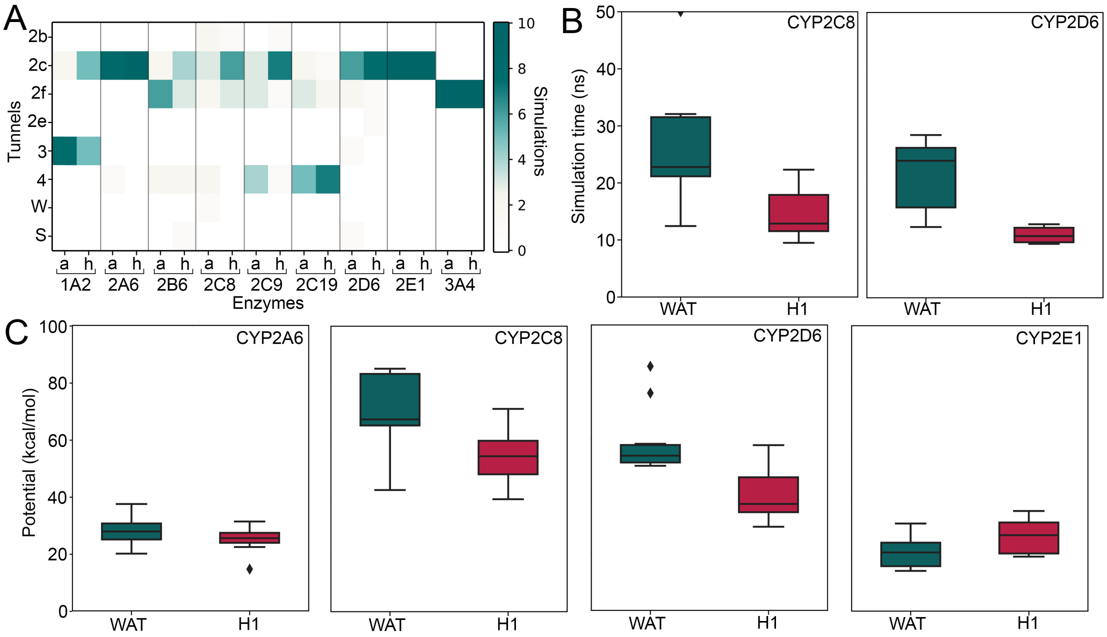


**Figure 3** Results from metadynamics simulations. (A) Heat map of selected pathways with (h) and without (a) allosteric ligand. (B) Boxplots of simulation times ($\Delta T$) until the ligand completely dissociated from the active site. (C) Boxplots of maximal potential ($P_{max}$) registered until the ligand completely dissociated from the active site.

in the cMD simulations, we monitored the tunnels selected by the ligands to egress from the binding site (Figure 3A). In accordance to the bottleneck radii, we observed more trajectories during which the ligand selected tunnel 2c when an allosteric ligand was present in CYP1A2. In contrast to the cMD simulations, which indicated tunnel 2b to be enlarged if an allosteric ligand was bound to H1, the ligands preferred different tunnels in CYP2A6 and CYP2B6. In analogy to the results from the cMD simulations, there was a slight preference for tunnel 2f in CYP2C8 in response to H1 interactions. While there was no consensus between both techniques in the closely related CYP2C9 and CYP2C19, there was a preference for tunnel 2c as selected tunnel in CYP2D6 in the presence of an allosteric ligand in accordance with the observed bottleneck radii. We observed only a single pathway to be of relevance independent of interactions with the H1 site in CYP2E1 and CYP3A4, indicating a missing functionality for the H1 site in these enzymes in addition to the above-mentioned small-molecule and organic probe densities.

| | | | | $r_B$ [b] | | |
|---|---|---|---|---|---|---|
| **Enzyme** | **Outcome $\Delta T$[a]** | **Outcome $P_{max}$ [a]** | **Significance** | **2b** | **2c** | **2f** |
| CYP1A2 | H1 (0.386) | H1 (0.400) | no | = | + | = |
| CYP2A6 | H1 (0.237) | H1 (0.179) | no | + | = | = |
| CYP2B6 | H1 (0.204) | H1 (0.522) | no | + | n/a[c] | + |
| CYP2C8 | H1 (0.010) | H1 (0.025) | yes | − | − | + |
| CYP2C9 | WAT (0.675) | WAT (0.559) | no | = | − | + |
| CYP2C19 | H1 (0.454) | H1 (0.862) | no | − | + | + |
| CYP2D6 | H1 (0.001) | H1 (0.001) | yes | − | + | − |
| CYP2E1 | WAT (0.232) | WAT (0.085) | yes | + | − | = |
| CYP3A4 | H1 (0.571) | WAT (0.713) | no | n/a[c] | = | − |

**Table 2** Statistical analysis of metadynamics simulations. [a] Outcome (H1 for allosteric, WAT for no allosteric) of the metadynamics simulations for the simulation time ($\Delta T$) and maximal potential ($P_{max}$). [b] Statistically outcome of potential changes in bottleneck radius ($r_B$). [c] Tunnel not present.

Next, we computed the maximal biasing potential ($P_{max}$) deposited during the dissociation, as well as the simulation time ($\Delta T$) elapsed until the ligand completely dissociated (Tables 2, S4, and S5). Both parameters have been correlated with residence times of ligands in previous work [34, 35] and, thus, they can be used to study the influence of the H1 interactions on ligand translocation. The significance of the differences was assessed using a two-sided t-test for both readouts. While only CYP2C8 and CYP2D6 presented significant improvements in $\Delta T$ or $P_{max}$ in the presence of an allosteric ligand (Figures 3B, S14, and S15), all enzymes besides CYP2C9, CYP2E1, and CYP3A4 presented an lower values of the two metrics with H1 interactions. In accordance to all our previous results, the association of small-molecules at the H1 seemed to have no effect on the functionality of ligand tunnels in CYP2E1 or CYP3A4, with the former even presenting more favorable values if there was no allosteric ligand present (Figure 3C).

## Conclusions

Previous work on the prokaryotic CYP101A1 indicated an allosteric site involved in the regulation of ligand access to its buried binding pocket by shifting the enzyme toward an open conformation upon small-molecule interaction. Here, we conducted a multi-scale modeling approach to follow up on this hypothesis focusing on the nine most relevant drug-metabolizing CYP enzymes in humans. Using MD simulations, we quantified the interaction of small-molecules and organic solvents with this allosteric site, here denoted as H1 site, which is located among helices C, E, and H. These interactions indicated that all studied enzymes besides CYP2E1 and CYP3A4 have the potential to bind ligands at the H1 site, although the observed interactions in CYP2C19 were comparatively limited. When we analyzed the conservation of residues in the H1 site, we discovered the whole CYP2 family to share a high degree of similarity as opposed to CYP1A2 and CYP3A4. To potentially reproduce previous findings on CYP101A1, we restrained a ligand to the H1 site and analyzed the three most relevant tunnels, which may be used by ligands to access the active site, in a triplicate of microsecond MD simulations. In CYP2B6, CYP2C8, CYP2C9, and CYP2C19, we could observe an enlargement of tunnel 2f, which is separated from tunnel 2b by the F-G loop. The structural adaptation of this loop was shown to mediate the switch to an open conformation, and thus, an enlargement of either tunnel 2b as it took place in CYP2A6, CYP2B6, and CYP2E1, or tunnel 2f, points towards such an adaptation. Interestingly, all members of the CYP2 family, with the exception of CYP2D6, presented such a change in accordance to previous results. As in the previous analyses, we could not detect any enlargement of tunnels in CYP3A4. Despite the results regarding the changes of bottleneck radii, we could not detect any significant changes in the radius of gyration during the microsecond MD simulations. Hence, we could not detect a transition to an open conformation in our simulations. To study ligand translocation directly, we conducted metadynamics simulations by selecting the distance of the ligand to the heme iron atom as CV. Although it was only significant in CYP2C8 and CYP2D6, all studied enzymes except for CYP2C9, CYP2E1, and CYP3A4 presented either shorter residence in the active site or a lower maximal potential to induce ligand dissociation. Thus, interactions at the allosteric site indeed facilitated the translocation of ligands. In contrast, CYP2E1 presented a statistically significant increase in both maximal potential and simulation time for the ligand to dissociate. In conclusion, the results from different computational techniques suggested that the findings in CYP101A1 can be translated to several mammalian enzymes of the CYP2 family.

## Materials and Methods

**Bioinformatics analysis.** All protein sequences format were obtained from the UniProt database [36] in FASTA format according to the accession codes listed in Table S6. Based on our previous work with CYP2D6 [8], we selected the 18 residues S137, T138, L139, R140, N141, L142, G143, L144, G145, K146, L149, L189, P268, R269, D270, L271, A274, and A277 to define the H1 allosteric site. Based on a multiple sequence alignment of the catalytic domains, we identified the corresponding residues in the remaining enzymes. For the sequence alignment we selected the Clustal W algorithm [37] within the UGENE suite of tools (v34.0) [38]. The conservation was determined according to the amino acid groups provided in our previous work [39]. If a residue was in the same group, we assigned it a conservation value of 0.5, while a value of 1.0 was given for identical residues and 0 was assigned if none of the above-mentioned conditions was fulfilled.

**Model building.** The crystal structures of all studied CYPs were obtained from the Protein Data Bank [40] according to the accession codes provided in Table S6. For proteins deposited as multimers, only chain A was retained without organic solvents, histidine tags, or ions. From the obtained PDB files, FASTA sequences were extracted using an in-house python routine. The generated sequences were aligned to the wild-type protein with the ClustalW algorithm [37] in the UGENE toolkit [38]. Mismatched amino acids were mutated to the ones corresponding to the wild-type enzyme in the 3D Builder panel within the Maestro Small-Molecule Drug Discovery suite (v2019-3) [41]. Next, the structures were processed with the Protein Preparation Wizard [42] by assigning bond orders, adding hydrogen atoms, predicting protonation states with Epik, reorienting the hydrogen bonding network with PROPKA at pH 7.4. Next, the structures were subjected to a restrained minimization with the OPLS3e force field to a convergence threshold of 0.3 Å for protein heavy atoms. Missing residues and side chains were added using Prime. The heme moieties of the protein were modeled with six coordination partners to the $Fe^{3+}$ ion including cysteine, the four porphyrin nitrogens, as well as an uncharged oxygen atom. If there was an inhibitor present in the active site, it was replaced with a structurally similar substrate that, however, is a substrate of the respective enzyme according to the review article by Rendic and colleagues [43]. These replacement and superposition procedures are detailed in Figure S16. To obtain systems covering the complete protein sequence, the transmembrane anchor was modeled as ideal alpha helix in the 3D Builder panel within Maestro by considering residues missing in the crystal structures. The anchors were placed perpendicular to the membrane plane along the z-axis and were aligned the C-terminus of the anchor to the N-terminus of the globular domain based on their backbone carbonyl atoms. Ultimately, we modeleted a covalent bond linking the two parts. To avoid steric clashes, torsion angles of the membrane anchor residues were adjusted if it was necessary. A prequilibrated POPC membrane leaflet was placed on the predicted position by the Orientations of Proteins in Membranes (OPM) protocol accessible at the PPM server, which estimates the membrane position based on energetic contributions [44], as previously described [21]. All used ligand structures in this work were pre-processed using the LigPrep routine at a pH of 7.4 for ionizable functional groups and the OPLS3e force field to obtain energy-minimized conformers. Based on the Ligprep output, we retained the most favorable protomer for each ligand.

**MD simulations.** All classical MD simulations were performed with the Desmond (v2019-1) simulation engine [45]. Five different sets of MD simulations were conducted: (i.) simulations with 20 ligand molecules arbitrarily placed around the enzymes, (ii). cosolvent simulations, (iii.) equilibration simulations of membrane-anchored CYPs, (iv.) production simulations with allosteric ligand restrained to the protein surface, and (v.) metadynamics simulations. All simulations were treated with the default equilibration protocol of Desmond, before the production phase was conducted in an NPT ensemble at atmospheric pressure regulated by the Martyna-Tobias-Klein barostat barostat and a temperature of 310 K maintained by the Nose-Hoover thermostat. The orthorhombic periodic boundary systems were solvated with TIP3P water molecules, counterions were used to neutralize the systems, and the OPLS_2005 force field was selected. In all simulations, the orthorhombic simulation box was defined with a distance cut-off of at least 10 Å to the nearest atom in all three cartesian directions. The time step of the RESPA integrator was set to 2 fs, long-range interactions were treated with the u-series algorithm [46], and bonds to hydrogen atoms controlled with M-SHAKE. For replica simulations, the random seed for initial velocities was modified.

For the first set of simulations, the membrane anchor was not included and the cocrystallized ligands were removed from the prepared structures. For each CYP, 20 substrate molecules were randomly placed around the enzyme to monitor the association of drug-like ligands on the surface of the enzyme (Table S7). A duration of 500 ns per enzyme was selected with atomic coordinates recorded at an interval of 50 ps. Cosolvent MD simulations were conducted with the Mixed Solvent MD workflow within the Desmond (v2019-1) simulation engine [45]. As probe molecules, we selected isopropanol, acetonitrile, and pyridine at a concentration of 5% (by volume). Again, the orthosteric ligand was removed from the prepared protein structures without membrane anchor. The simulations were conducted with 10 replicas per cosolvent resulting in 30 individual simulations per enzyme. After an equilibration phase of 15 ns, the association of the probe molecules was sampled for 5 ns leading to a total simulation time of 600 ns per system.

The third set of simulations was conducted on the complete protein-membrane systems in order to equilibrate them. The simulation time was set to 300 ns and atomic coordinates were recorded at an interval of 30 ps.

The fourth set of simulations were started from the last MD frame of the prequilibrated protein-membrane systems. Using the rebuild_cms.py script that comes with Maestro, new systems were generated with a ligand bound to the H1 site. To obtain a starting conformation of the selected ligand bound to the H1 allosteric site, we used the Glide standard-precision docking protocol [47]. The selected ligands are given in Table S8 and Figure S17. We defined the search space for docking considering the position of acetaminophen bound to CYP2D6 H1 from our previous work. Water molecules overlapping with ligand atoms were removed assuming the probe size of 1.65 Å. After initial 5 ns of simulation to allow the ligand to freely accommodate within the allosteric site, we applied harmonic distance restraints with a force constant of 2.5 kcal*mol$^{-1}$*Å$^{-1}$ between the central atom of the ligand and $\alpha$-carbons of three residues of the protein (Table S9) located in the comparatively rigid helices C, E, and I. These simulations were conducted in triplicates with a duration of 1 $\mu$s and atomic coordinates stored at an interval of 100 ps.

The metadynamics simulations were also conducted with the Desmond simulation engine. The simulation systems were retained from the above-mentioned production simulations with a ligand restrained to the H1 site combined with an orthosteric ligand. As CV, we selected the distance between the centroid of the ligand and the heme iron atom with a wall of 45 Å. We retained the height of the Gaussian at 0.03 kcal/mol as well as the width of 0.05 Å according to the default specification.

**Evaluation of the MD trajectories.** For all simulations, except for the comparatively short cosolvent MD simulations, we computed RMSD and RMSF values using the Simulation Interaction Diagram panel in Maestro. For the metadynamics simulation, we truncated the trajectories beforehand to only represent the dissociation process using the trajectory_extract_subsystem.py python routine that comes with Maestro.

Initially, the first set of simulations was conducted to obtain a stable pose of the ligands bound to the H1 site that could be superimposed to the prequilibrated membrane models for further procedures. However, conformational changes of the protein surface prevented this procedure, as they would have introduced large steric clashes. Thus, we evaluated these simulations to obtain additional insight into the preference of small-molecule association on the enzyme surface. We used an in-house python routine computing the cumulative number of ligand heavy atoms within 5 Å distance to protein $\alpha$-carbons for each MD frame and normalized this count to the total number of ligand heavy atoms.

To validate our membrane models with experimental parameters, we determined the heme tilt angle as well as the burying depth of the membrane-anchored proteins during our equilibration simulations. The heme tilt angle was defined as the angle between the heme plane defined by the porphyrin nitrogens and the membrane normal (z-axis) [11, 21]. The burying depth was defined as distance between the mass center of the protein considering $\alpha$-carbons and the centroid of the POPC C1-carbons [21, 48]. Both of these validation parameters were determined for every frame of the trajectories in analogy to our previous work [8].

Tunnels connecting the buried active site of CYPs to the surrounding solvent environment were computed using CAVER (v3.0) [27] for the microsecond simulations of the full-length proteins embedded in a membrane. The starting point for the tunnel computation was determined in CAVER Analyst (v1.0) [49] by selecting the heme, as well as two additional residues in the binding site based on a structural alignment to CYP2D6, for which we defined suitable residues in our previous work (Table S10). Similarly, as reported in our previous analyses on CYP2D6 [8, 21], we selected a clustering threshold of 4.5 for the computation. As we observed four simulations presenting a dissociation of the allosteric ligand during the preequilibration of 5 ns without restraints, we discarded these trajectories from the tunnel computation (Table S11). The radius of gyration was determined from MD frames in according to the publication of Lobanov and colleagues [29].

While we visually determined the simulation time $\Delta$T when the ligand completely dissociated from the protein in the metadynamics simulations, we derived the maximal potential $P_{max}$ during the egress process using the metadynminer toolkit based on R scripting language [50]. The statistical significance of the average simulation time until ligand egress, the maximal potentials, as well as the bottleneck radii was evaluated using ttest_ind_from_stats routine that comes with the python-scipy module at the p=0.1 significance level.

## References

[1] Ulrich M. Zanger and Matthias Schwab. Cytochrome P450 enzymes in drug metabolism: Regulation of gene expression, enzyme activities, and impact of genetic variation. *Pharmacology and Therapeutics*, 138(1):103–141, 2013.

[2] Palrasu Manikandan and Siddavaram Nagini. Cytochrome P450 Structure, Function and Clinical Significance: A Review. *Current drug targets*, 19(1):38–54, 2018.

[3] S Casey Laizure, Vanessa Herring, Zheyi Hu, Kevin Witbrodt, and Robert B Parker. The role of human carboxylesterases in drug metabolism: have we overlooked their importance? *Pharmacotherapy*, 33(2):210–222, 2 2013.

[4] Jonathan D Tyzack and Johannes Kirchmair. Computational methods and tools to predict cytochrome P450 metabolism for drug discovery. *Chemical biology and drug design*, 93 (4):377–386, 4 2019.

[5] Han van de Waterbeemd, Eric Gifford, and Ann Arbor. ADMET in silico modelling: towards prediction paradise? *Nat Rev Drug Discov*, 2(3):192–204, 3 2003.

[6] Ismail Kola and John Landis. Can the pharmaceutical industry reduce attrition rates? *Nature Reviews Drug Discovery*, 3(8):711–715, 2004.

[7] Maximilian J L J Fürst, Filippo Fiorentini, and Marco W Fraaije. Beyond active site residues: overall structural dynamics control catalysis in flavin-containing and heme-containing monooxygenases. *Current Opinion in Structural Biology*, 59:29–37, 2019.

[8] André Fischer and Martin Smieško. Spontaneous Ligand Access Events to Membrane-Bound Cytochrome P450 2D6 Sampled at Atomic Resolution. *Scientific Reports*, 9(1): 16411, 2019.

[9] Philippe Urban, Thomas Lautier, Denis Pompon, and Gilles Truan. Ligand Access Channels in Cytochrome P450 Enzymes: A Review. *Int J Mol Sci.*, 19(6), 5 2018.

[10] Karel Berka, Tereza Hendrychová, Pavel Anzenbacher, and Michal Otyepka. Membrane position of ibuprofen agrees with suggested access path entrance to cytochrome P450 2C9 active site. *Journal of Physical Chemistry A*, 115(41):11248–11255, 2011.

[11] Karel Berka, Markéta Paloncýová, Pavel Anzenbacher, and Michal Otyepka. Behavior of human cytochromes P450 on lipid membranes. *Journal of Physical Chemistry B*, 117(39): 11556–11564, 2013.

[12] Artur Gora, Jan Brezovsky, and Jiri Damborsky. Gates of enzymes. *Chemical Reviews*, 113(8):5871–5923, 2013.

[13] André Fischer, Gabriela Frehner, Markus A Lill, and Martin Smieško. Conformational Changes of Thyroid Receptors in Response to Antagonists. *Journal of Chemical Information and Modeling*, 2021.

[14] Alec H Follmer, Mavish Mahomed, David B Goodin, and Thomas L Poulos. Substrate-Dependent Allosteric Regulation in Cytochrome P450cam (CYP101A1). *Journal of the American Chemical Society*, 140:16222–16228, 2018.

[15] Vlad Cojocaru, Peter J. Winn, and Rebecca C. Wade. The ins and outs of cytochrome P450s. *Biochim Biophys Acta.*, 1770(3):390–401, 2007.

[16] Gennady M. Verkhivker, Steve Agajanian, Guang Hu, and Peng Tao. Allosteric Regulation at the Crossroads of New Technologies: Multiscale Modeling, Networks, and Machine Learning. *Frontiers in Molecular Biosciences*, 7(July), 2020.

[17] George P. Lisi and J. Patrick Loria. Allostery in enzyme catalysis. *Current Opinion in Structural Biology*, 47:123–130, 2017.

[18] Phani Ghanakota and Heather A. Carlson. Driving Structure-Based Drug Discovery through Cosolvent Molecular Dynamics. *Journal of Medicinal Chemistry*, 59(23):10383–10399, 2016.

[19] Marco De Vivo, Matteo Masetti, Giovanni Bottegoni, and Andrea Cavalli. Role of Molecular Dynamics and Related Methods in Drug Discovery. *Journal of Medicinal Chemistry*, 59(9):4035–4061, 2016.

[20] Adam Hospital, Josep Ramón Goñi, Modesto Orozco, and Josep Gelpi. Molecular dynamics simulations: Advances and applications. *Advances and Applications in Bioinformatics and Chemistry*, 8:37–47, 2015.

[21] André Fischer, Charleen G. Don, and Martin Smieško. Molecular Dynamics Simulations Reveal Structural Differences among Allelic Variants of Membrane-Anchored Cytochrome P450 2D6. *Journal of Chemical Information and Modeling*, 58(9):1962–1975, 2018.

[22] Yoshihiro Ohta, Suguru Kawato, Hiroko Tagashira, Shigeki Takemori, and Shiro Kominami. Dynamic Structures of Adrenocortical Cytochrome P-450 in Proteoliposomes and Microsomes: Protein Rotation Study. *Biochemistry*, 31(50):12680–12687, 1992.

[23] T. H. Bayburt and S. G. Sligar. Single-molecule height measurements on microsomal cytochrome P450 in nanometer-scale phospholipid bilayer disks. *Proceedings of the National Academy of Sciences*, 99(10):6725–6730, 2002.

[24] Young-Tae Lee, Richard F Wilson, Igor Rupniewski, and David B Goodin. P450cam visits an open conformation in the absence of substrate. *Biochemistry*, 49(16):3412–3419, 4 2010.

[25] P. Haritha, P. Shanmugavadivu, and S. Dhamodharan. A Comprehensive Review on Protein Sequence Analysis Techniques. *International Journal of Computer Sciences and Engineering*, 6(7):1433–1442, 2018.

[26] Markéta Paloncýova, Veronika Navrátilova, Karel Berka, Alessandro Laio, and Michal Otyepka. Role of Enzyme Flexibility in Ligand Access and Egress to Active Site: Bias-Exchange Metadynamics Study of 1,3,7-Trimethyluric Acid in Cytochrome P450 3A4. *Journal of Chemical Theory and Computation*, 12(4):2101–2109, 2016.

[27] Eva Chovancova, Antonin Pavelka, Petr Benes, Ondrej Strnad, Jan Brezovsky, Barbora Kozlikova, Artur Gora, Vilem Sustr, Martin Klvana, Petr Medek, Lada Biedermannova, Jiri Sochor, and Jiri Damborsky. CAVER 3.0: A Tool for the Analysis of Transport Pathways in Dynamic Protein Structures. *PLoS Computational Biology*, 8(10):23–30, 2012.

[28] Axel Sündermann and Chris Oostenbrink. Molecular dynamics simulations give insight into the conformational change, complex formation, and electron transfer pathway for cytochrome P450 reductase. *Protein Science*, 22(9):1183–1195, 2013.

[29] M Yu. Lobanov, N S Bogatyreva, and O V Galzitskaya. Radius of gyration as an indicator of protein structure compactness. *Molecular Biology*, 42(4):623–628, 2008.

[30] Benjamin Trendelkamp-Schroer and Frank Noé. Efficient Estimation of Rare-Event Kinetics. *Phys. Rev. X*, 6(1):11009, 1 2016.

[31] Alessandro Laio and Francesco L. Gervasio. Metadynamics: A method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Reports on Progress in Physics*, 71(12), 2008.

[32] Donald Hamelberg, John Mongan, and J Andrew McCammon. Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules. *Journal of Chemical Physics*, 120(24):11919–11929, 2004.

[33] Kenneth W. Borrelli, Andreas Vitalis, Raul Alcantara, and Victor Guallar. PELE: Protein energy landscape exploration. A novel Monte Carlo based technique. *Journal of Chemical Theory and Computation*, 1(6):1304–1311, 2005.

[34] Andrea Bortolato, Francesca Deflorian, Dahlia R Weiss, and Jonathan S Mason. Decoding the Role of Water Dynamics in Ligand–Protein Unbinding: CRF1R as a Test Case. *Journal of Chemical Information and Modeling*, 55(9):1857–1866, 9 2015.

[35] Huiyong Sun, Youyong Li, Mingyun Shen, Dan Li, Yu Kang, and Tingjun Hou. Characterizing Drug–Target Residence Time with Metadynamics: How To Achieve Dissociation Rate Efficiently without Losing Accuracy against Time-Consuming Approaches. *Journal of Chemical Information and Modeling*, 57(8):1895–1906, 8 2017.

[36] The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1):D506–D515, 1 2019.

[37] Julie D. Thompson, Desmond G. Higgins, and Toby J. Gibson. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22(22): 4673–4680, 1994.

[38] Konstantin Okonechnikov, Olga Golosova, Mikhail Fursov, Alexey Varlamov, Yuri Vaskin, Ivan Efremov, O. G. German Grehov, Denis Kandrov, Kirill Rasputin, Maxim Syabro, and Timur Tleukenov. Unipro UGENE: A unified bioinformatics toolkit. *Bioinformatics*, 28 (8):1166–1167, 2012.

[39] André Fischer and Martin Smieško. Allosteric binding sites on nuclear receptors: Focus on drug efficacy and selectivity. *International Journal of Molecular Sciences*, 21(2):6–8, 2020.

[40] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, T N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 1 2000.

[41] Schrödinger LCC. Maestro Small-Molecule Drug Discovery Suite 2019-3. 2019.

[42] G. Madhavi Sastry, Matvey Adzhigirey, Tyler Day, Ramakrishna Annabhimoju, and Woody Sherman. Protein and ligand preparation: Parameters, protocols, and influence

on virtual screening enrichments. *Journal of Computer-Aided Molecular Design*, 27(3): 221–234, 2013.

[43] Slobodan Rendic. Summary of information on human CYP enzymes: Human P450 metabolism data. *Drug Metabolism Reviews*, 34(1-2):83–448, 2002.

[44] Mikhail A. Lomize, Irina D Pogozheva, Hyeon Joo, Henry I Mosberg, and Andrei L Lomize. OPM database and PPM web server: Resources for positioning of proteins in membranes. *Nucleic Acids Research*, 40(D1):370–376, 2012.

[45] Kevin Bowers, Edmond Chow, Huafeng Xu, Ron Dror, Michael Eastwood, Brent Gregersen, John Klepeis, Istvan Kolossvary, Mark Moraes, Federico Sacerdoti, John Salmon, Yibing Shan, and David Shaw. Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters. *ACM/IEEE SC 2006 Conference (SC'06)*, (November):43, 2006.

[46] David E. Shaw, J. P. Grossman, Joseph A. Bank, Brannon Batson, J. Adam Butts, Jack C. Chao, Martin M. Deneroff, Ron O. Dror, Amos Even, Christopher H. Fenton, Anthony Forte, Joseph Gagliardo, Gennette Gill, Brian Greskamp, C. Richard Ho, Douglas J. Ierardi, Lev Iserovich, Jeffrey S. Kuskin, Richard H. Larson, Timothy Layman, Li Siang Lee, Adam K. Lerer, Chester Li, Daniel Killebrew, Kenneth M. Mackenzie, Shark Yeuk Hai Mok, Mark A. Moraes, Rolf Mueller, Lawrence J. Nociolo, Jon L. Peticolas, Terry Quan, Daniel Ramot, John K. Salmon, Daniele P. Scarpazza, U. Ben Schafer, Naseer Siddique, Christopher W. Snyder, Jochen Spengler, Ping Tak Peter Tang, Michael Theobald, Horia Toma, Brian Towles, Benjamin Vitale, Stanley C. Wang, and Cliff Young. Anton 2: Raising the Bar for Performance and Programmability in a Special-Purpose Molecular Dynamics Supercomputer. *International Conference for High Performance Computing, Networking, Storage and Analysis, SC*, 2015-Janua(January):41–53, 2014.

[47] Thomas A. Halgren, Robert B. Murphy, Richard A. Friesner, Hege S. Beard, Leah L. Frye, W. Thomas Pollard, and Jay L. Banks. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. *Journal of Medicinal Chemistry*, 47(7):1750–1759, 2004.

[48] Lionel Ducassou, Laura Dhers, Gabriella Jonasson, Nicolas Pietrancosta, Jean Luc Boucher, Daniel Mansuy, and François André. Membrane-bound human orphan cytochrome P450 2U1: Sequence singularities, construction of a full 3D model, and substrate docking. *Biochimie*, 140:166–175, 2017.

[49] Barbora Kozlikova, Eva Sebestova, Vilem Sustr, Jan Brezovsky, Ondrej Strnad, Lukas Daniel, David Bednar, Antonin Pavelka, Martin Manak, Martin Bezdeka, Petr Benes, Matus Kotry, Artur Gora, Jiri Damborsky, and Jiri Sochor. CAVER Analyst 1.0: Graphic tool for interactive visualization and analysis of tunnels and channels in protein structures. *Bioinformatics*, 30(18):2684–2685, 2014.

[50] Dalibor Trapl and Vojtěch Spiwok. Analysis of the Results of Metadynamics Simulations by metadynminer and metadynminer3d. *arXiv.org*, XX:1–11, 2020.

## 3.1 Supporting Information

# Supporting Results and Discussion

## Simulation techniques and model validation

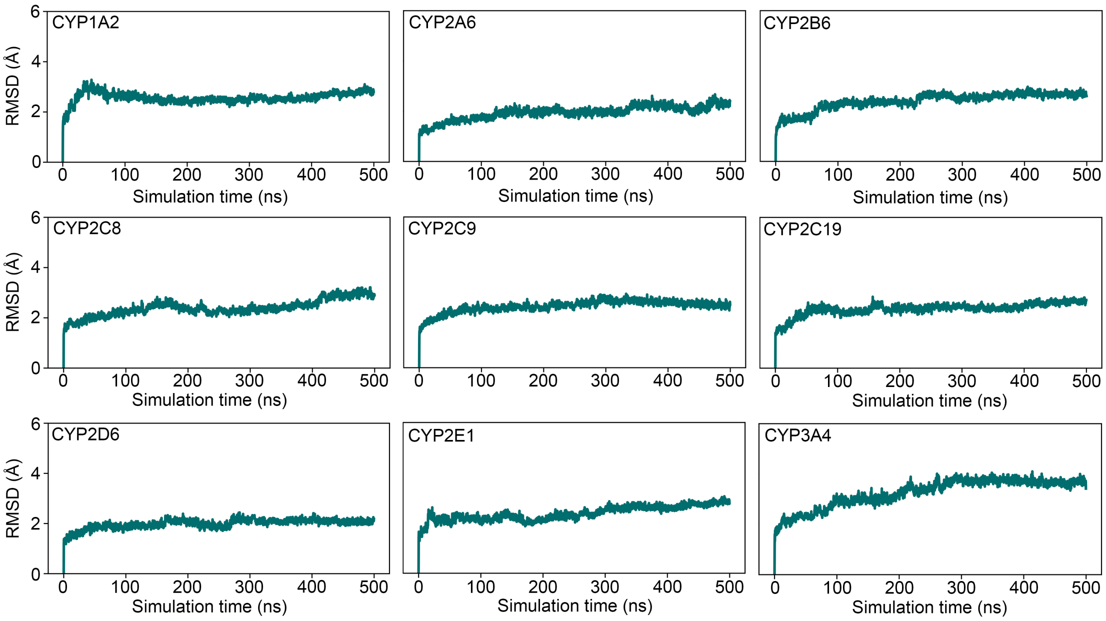


**Figure S 1** RMSD of association simulations.

**Table S 1** Membrane model validation.

| Enzyme | Heme tilt angle (°) | Burying depth (Å) |
|---|---|---|
| CYP1A2 | 66.7 ± 3.7 | 36.0 ± 1.1 |
| CYP2A6 | 39.1 ± 3.8 | 38.5 ± 1.7 |
| CYP2B6 | 66.3 ± 3.4 | 38.9 ± 1.6 |
| CYP2C8 | 63.8 ± 3.5 | 38.6 ± 1.4 |
| CYP2C9 | 68.6 ± 3.4 | 37.3 ± 1.3 |
| CYP2C19 | 62.5 ± 3.7 | 36.9 ± 1.9 |
| CYP2D6 | 55.3 ± 3.5 | 38.1 ± 1.3 |
| CYP2E1 | 53.1 ± 3.9 | 38.4 ± 1.6 |
| CYP3A4 | 75.3 ± 3.0 | 35.7 ± 1.0 |

Average values are given with standard deviation.

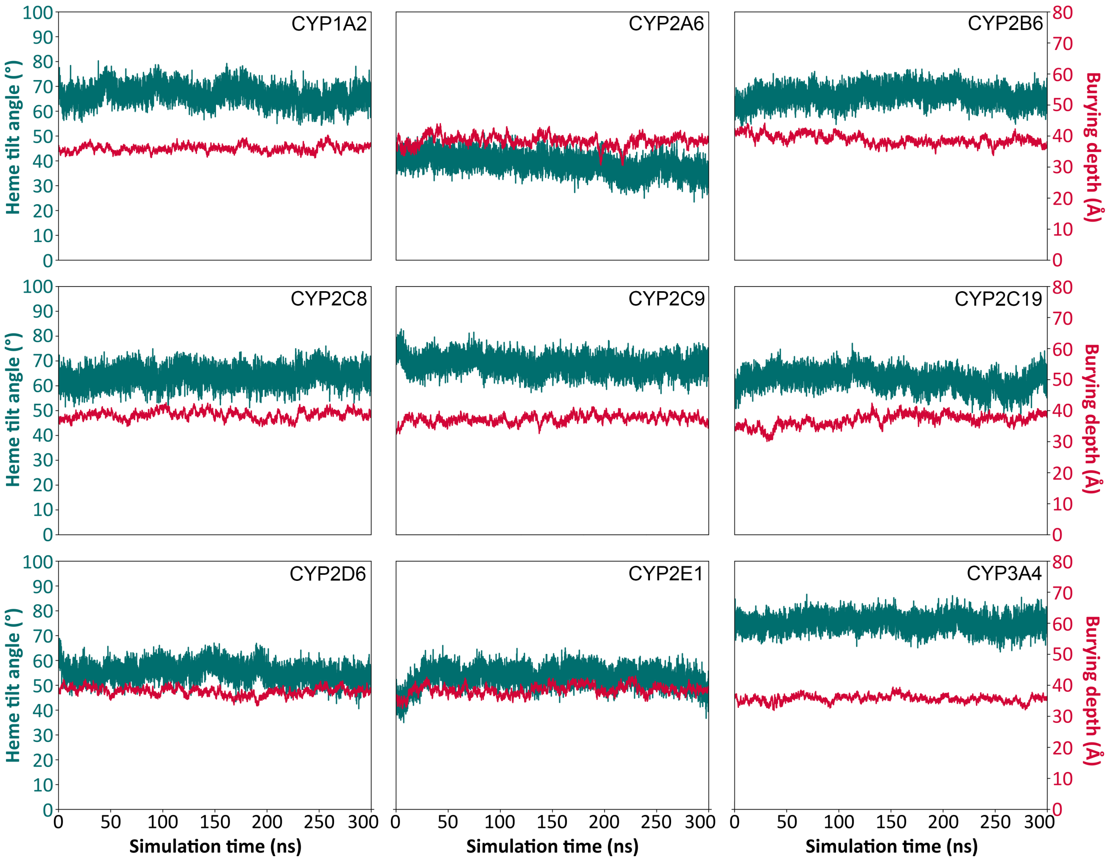


**Figure S 2** Time-evolved values of heme tilt angle (pine green) and buyring depth (red) during association simulations.

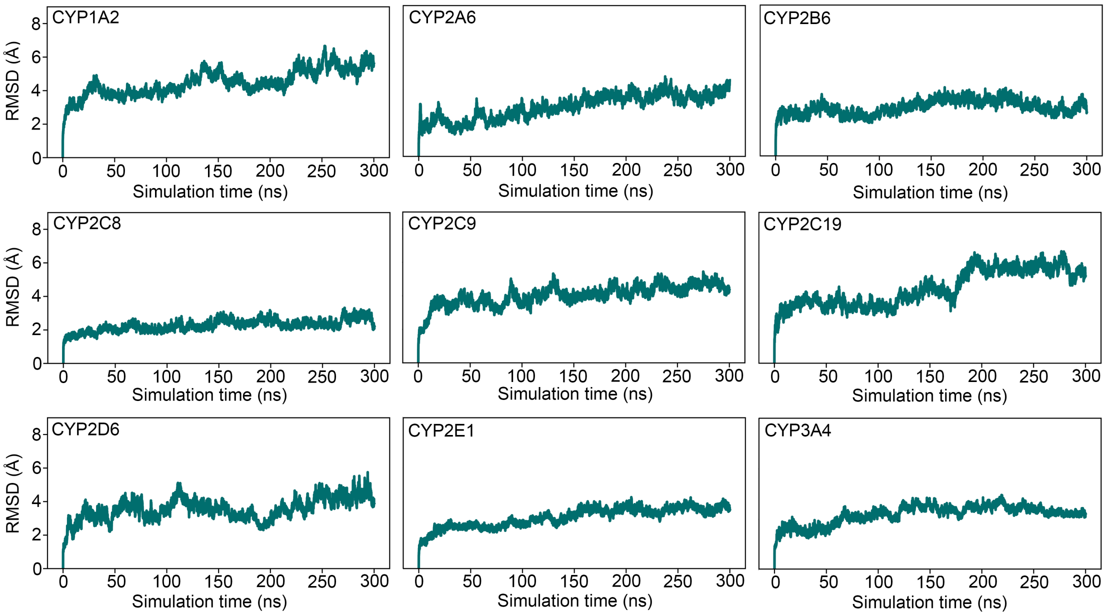


**Figure S 3** RMSD of membrane equilibration simulations.

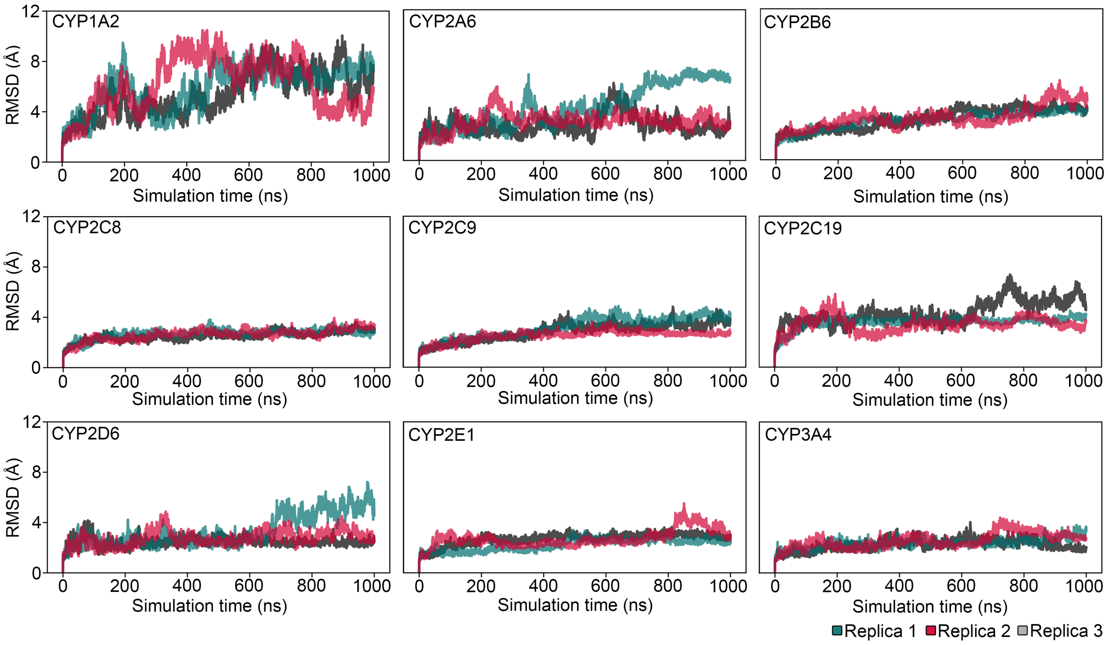


**Figure S 4** RMSD of sampling simulations.

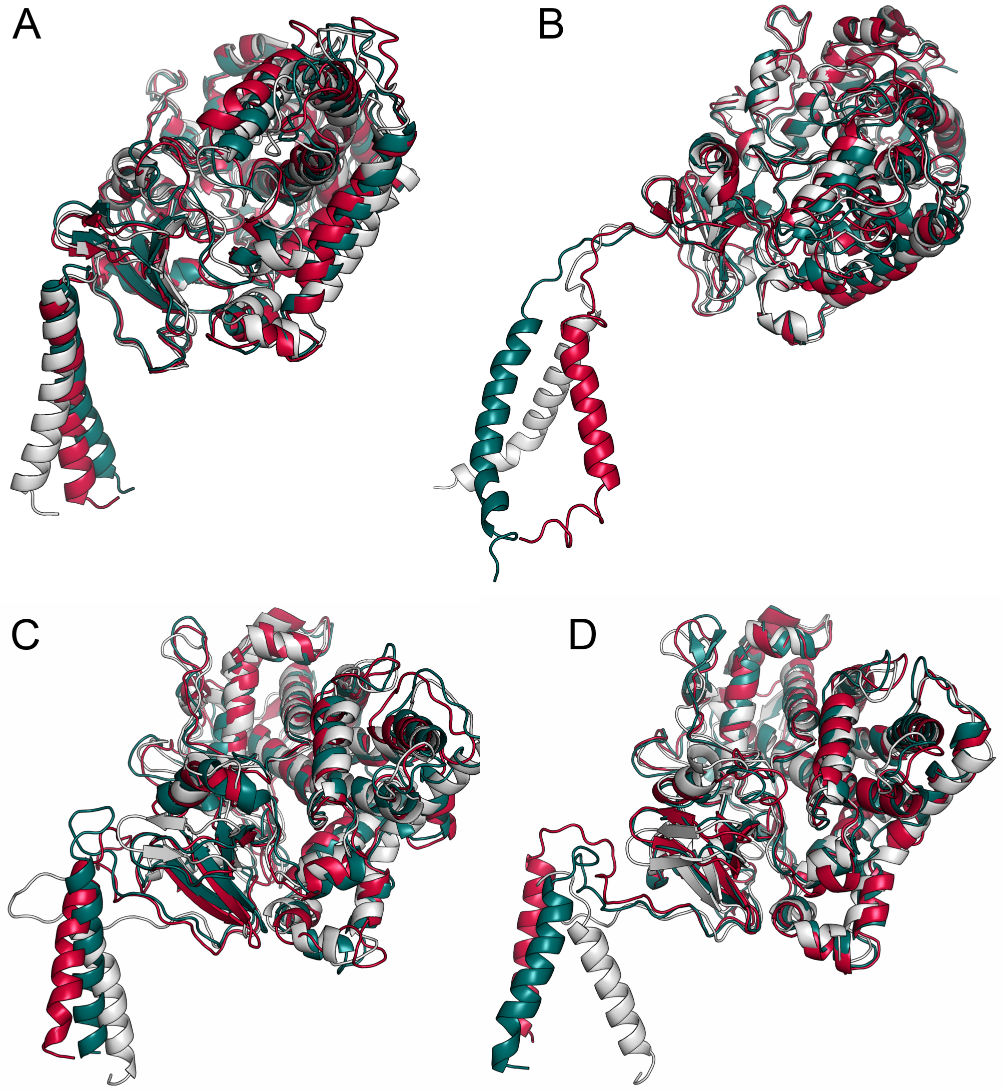


**Figure S 5** Depiction of conformational changes in sampling simulations for (A) CYP2C8, (B) CYP1A2, (C) CYP2C19, and (D) CYP2C19. The three replica simulations of each system are shown in different colors.

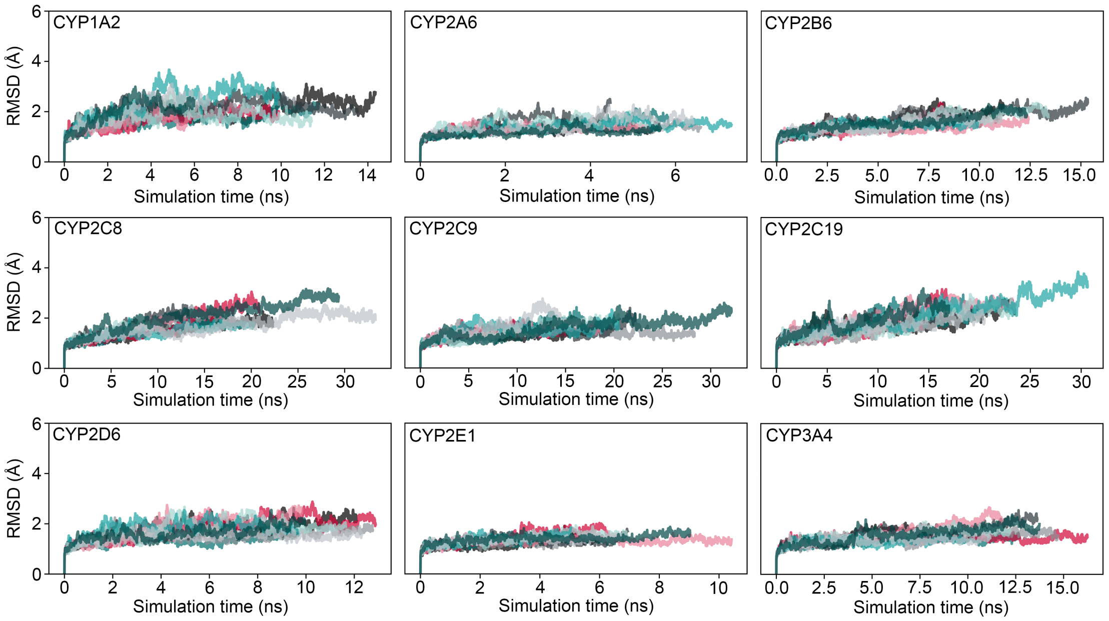


**Figure S 6** RMSD of metadynamics simulations (with allosteric ligand) in all enzymes studies here. Different replica simulations are indicated by different colors.

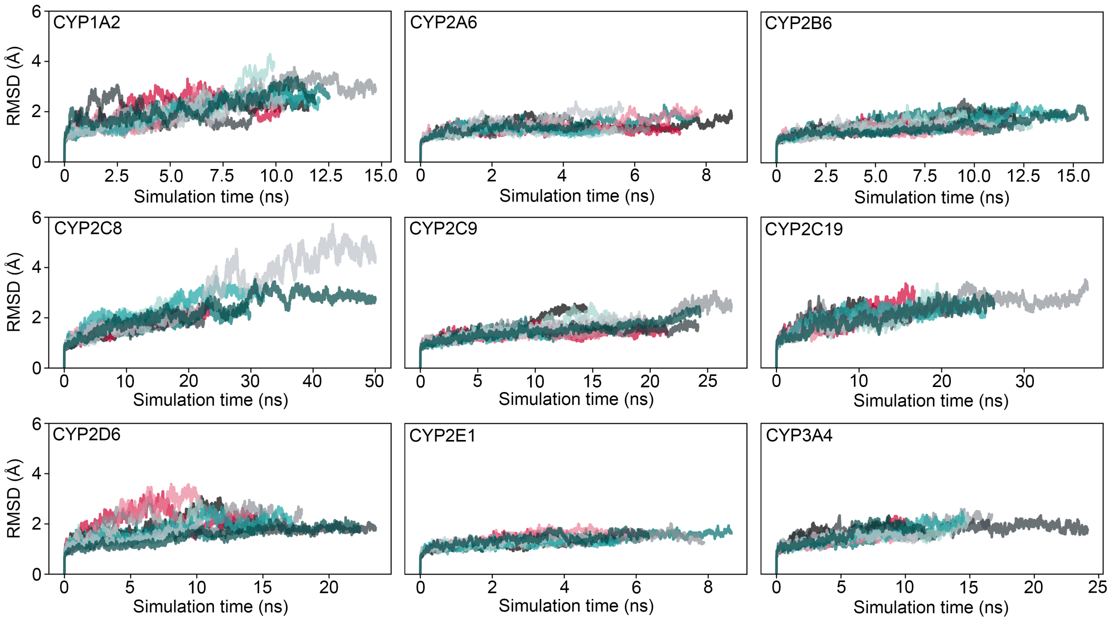


**Figure S 7** RMSD of metadynamics simulations (without allosteric ligand) in all enzymes studies here. Different replica simulations are indicated by different colors.

## The effect of allosteric ligands bound to the H1 site is isoform-dependent.

**Table S 2** Statistics of bottleneck radii.

| Enzyme | Lig$_{allo}$ | Tunnel | n | Mean | SD |
|---|---|---|---|---|---|
| CYP1A2 | no | 2b | 265 | 1.033 | 0.108 |
| CYP1A2 | yes | | 224 | 1.023 | 0.120 |
| CYP1A2 | no | 2c | 396 | 1.025 | 0.089 |
| CYP1A2 | yes | | 625 | 1.063 | 0.141 |
| CYP1A2 | no | 2f | 508 | 1.141 | 0.123 |
| CYP1A2 | yes | | 773 | 1.129 | 0.141 |
| CYP2A6 | no | 2b | 70 | 0.941 | 0.039 |
| CYP2A6 | yes | | 242 | 0.967 | 0.071 |
| CYP2A6 | no | 2c | 97 | 0.969 | 0.069 |
| CYP2A6 | yes | | 327 | 0.955 | 0.059 |
| CYP2A6 | no | 2f | 15 | 0.920 | 0.015 |
| CYP2A6 | yes | | 35 | 0.939 | 0.035 |
| CYP2B6 | no | 2b | 535 | 1.181 | 0.176 |
| CYP2B6 | yes | | 892 | 1.309 | 0.290 |
| CYP2B6 | no | 2f | 126 | 1.038 | 0.104 |
| CYP2B6 | yes | | 346 | 1.074 | 0.119 |
| CYP2C8 | no | 2b | 555 | 1.641 | 0.209 |
| CYP2C8 | yes | | 1455 | 1.468 | 0.274 |
| CYP2C8 | no | 2c | 555 | 1.605 | 0.262 |
| CYP2C8 | yes | | 1262 | 1.395 | 0.308 |
| CYP2C8 | no | 2f | 11 | 0.953 | 0.064 |
| CYP2C8 | yes | | 1154 | 1.326 | 0.253 |
| CYP2C9 | no | 2b | 5 | 0.982 | 0.060 |
| CYP2C9 | yes | | 526 | 1.118 | 0.145 |
| CYP2C9 | no | 2c | 556 | 1.512 | 0.215 |
| CYP2C9 | yes | | 370 | 1.055 | 0.131 |
| CYP2C9 | no | 2f | 62 | 0.959 | 0.044 |
| CYP2C9 | yes | | 222 | 0.973 | 0.068 |

**Table S 3** Statistics of bottleneck radii (continued).

| **Enzyme** | **Lig$_{\text{allo}}$** | **Tunnel** | **n** | **Mean** | **SD** |
|---|---|---|---|---|---|
| CYP2C19 | no | 2b | 465 | 1.342 | 0.240 |
| CYP2C19 | yes | | 133 | 1.109 | 0.251 |
| CYP2C19 | no | 2c | 76 | 1.178 | 0.338 |
| CYP2C19 | yes | | 505 | 1.510 | 0.306 |
| CYP2C19 | no | 2f | 238 | 1.093 | 0.227 |
| CYP2C19 | yes | | 839 | 1.609 | 0.477 |
| CYP2D6 | no | 2b | 354 | 1.289 | 0.244 |
| CYP2D6 | yes | | 1342 | 1.214 | 0.189 |
| CYP2D6 | no | 2c | 119 | 1.007 | 0.141 |
| CYP2D6 | yes | | 570 | 1.046 | 0.139 |
| CYP2D6 | no | 2f | 344 | 1.328 | 0.242 |
| CYP2D6 | yes | | 1119 | 1.192 | 0.178 |
| CYP2E1 | no | 2b | 302 | 1.054 | 0.153 |
| CYP2E1 | yes | | 333 | 1.087 | 0.156 |
| CYP2E1 | no | 2c | 556 | 1.640 | 0.203 |
| CYP2E1 | yes | | 498 | 1.454 | 0.295 |
| CYP2E1 | no | 2f | 61 | 0.956 | 0.059 |
| CYP2E1 | yes | | 35 | 0.962 | 0.062 |
| CYP3A4 | no | 2c | 38 | 0.983 | 0.085 |
| CYP3A4 | yes | | 34 | 1.054 | 0.331 |
| CYP3A4 | no | 2f | 555 | 2.161 | 0.261 |
| CYP3A4 | yes | | 1433 | 1.78 | 0.496 |

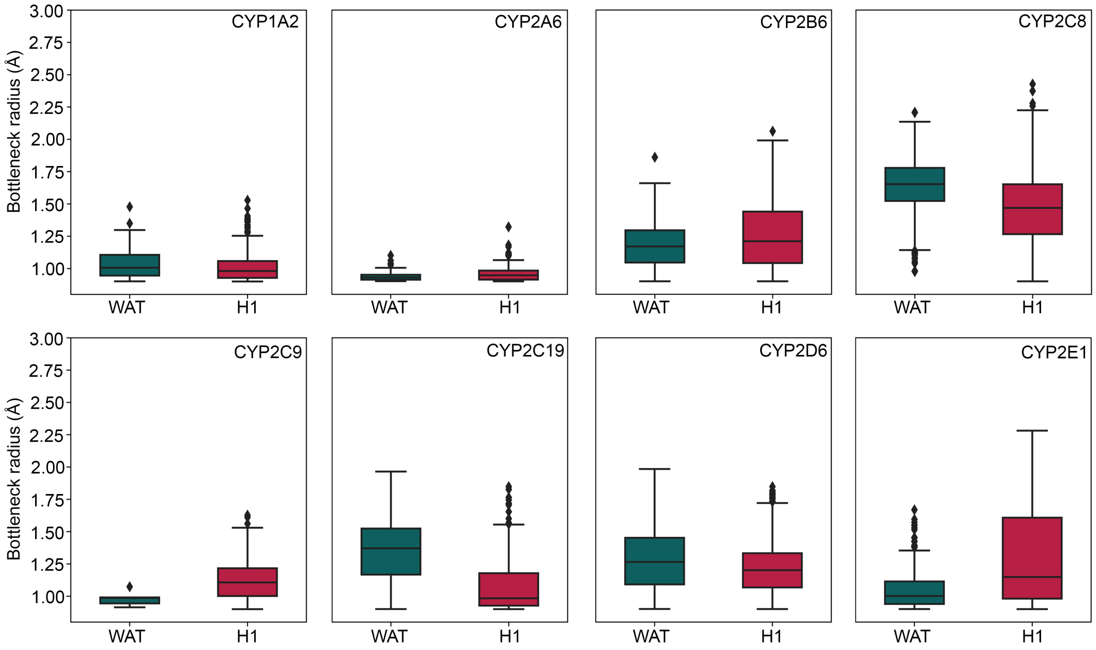


**Figure S 8** Boxplots of bottleneck radii for tunnel 2b. While "WAT" indicates no allosteric ligand present, "H1" indicates if a ligand was bound to H1.

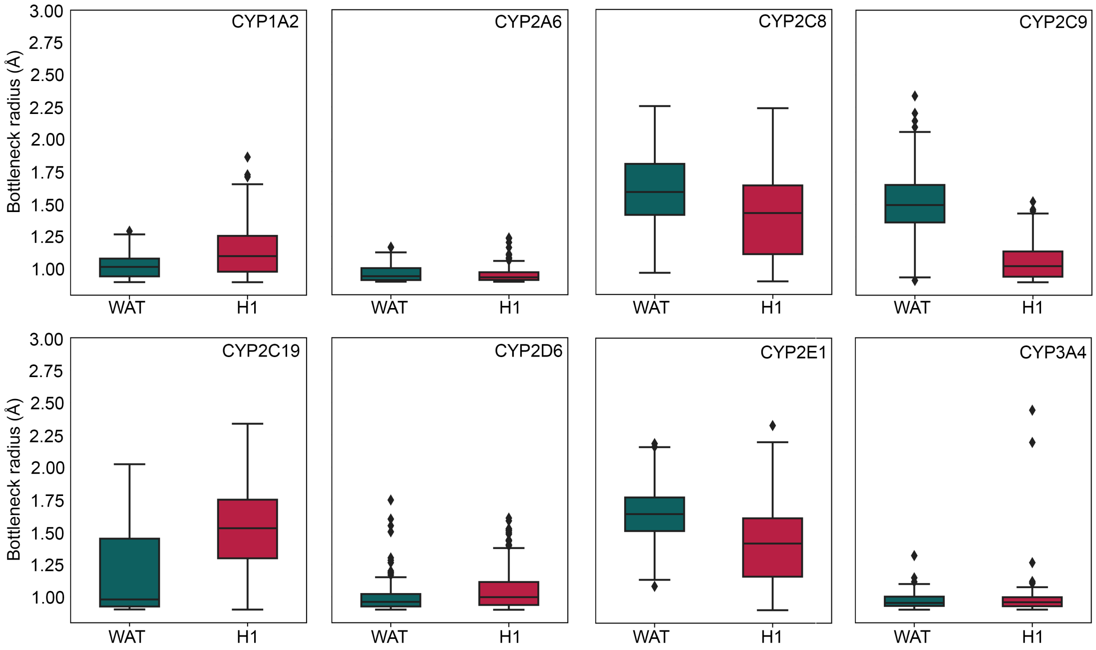


**Figure S 9** Boxplots of bottleneck radii for tunnel 2c. While "WAT" indicates no allosteric ligand present, "H1" indicates if a ligand was bound to H1.

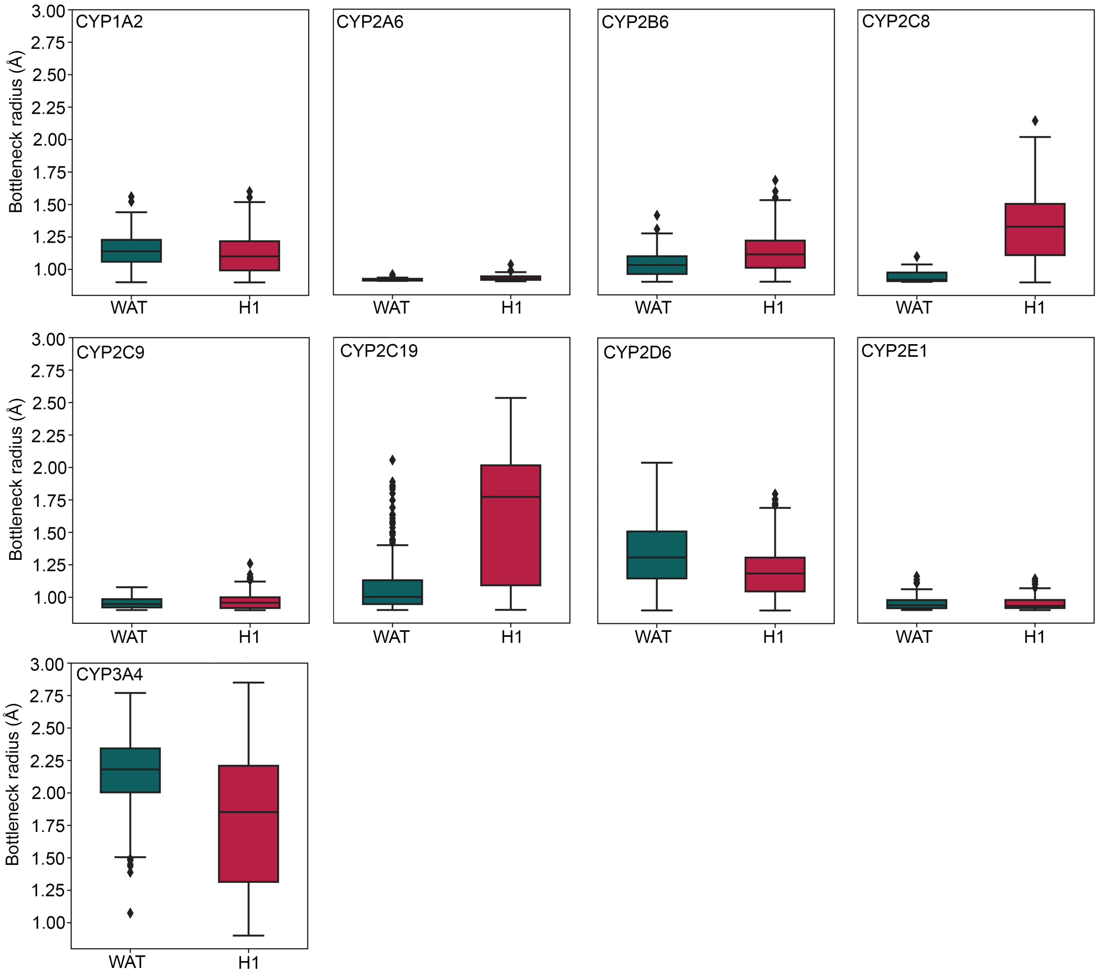


**Figure S 10** Boxplots of bottleneck radii for tunnel 2f. While ”WAT” indicates no allosteric ligand present, ”H1” indicates if a ligand was bound to H1.

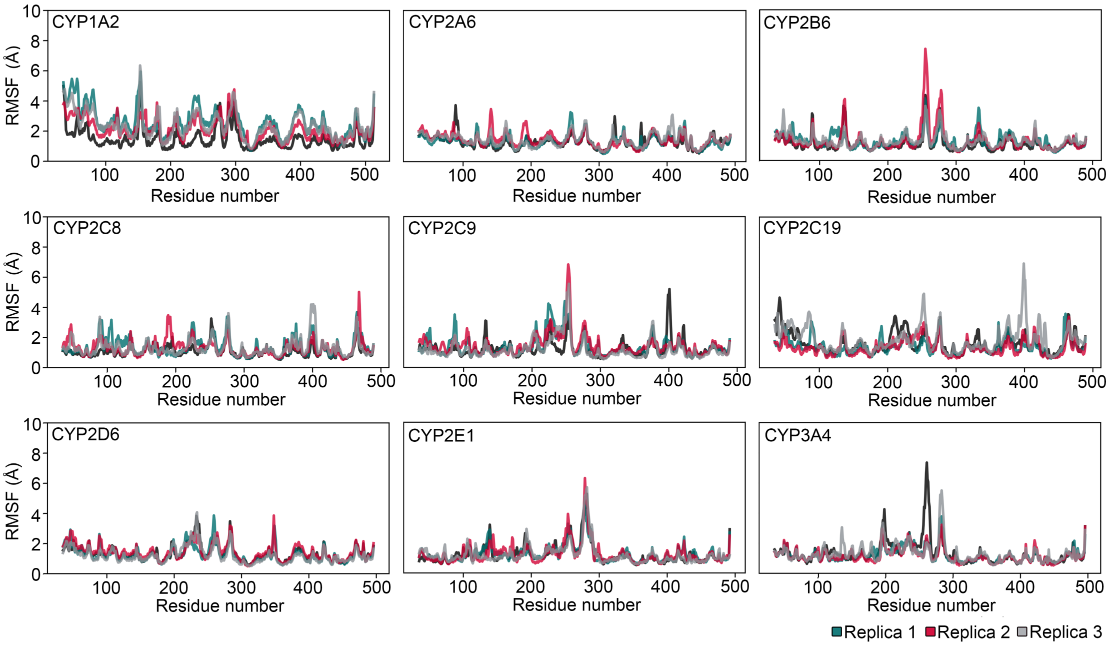


**Figure S 11** RMSF of sampling simulations for all studied enzymes.

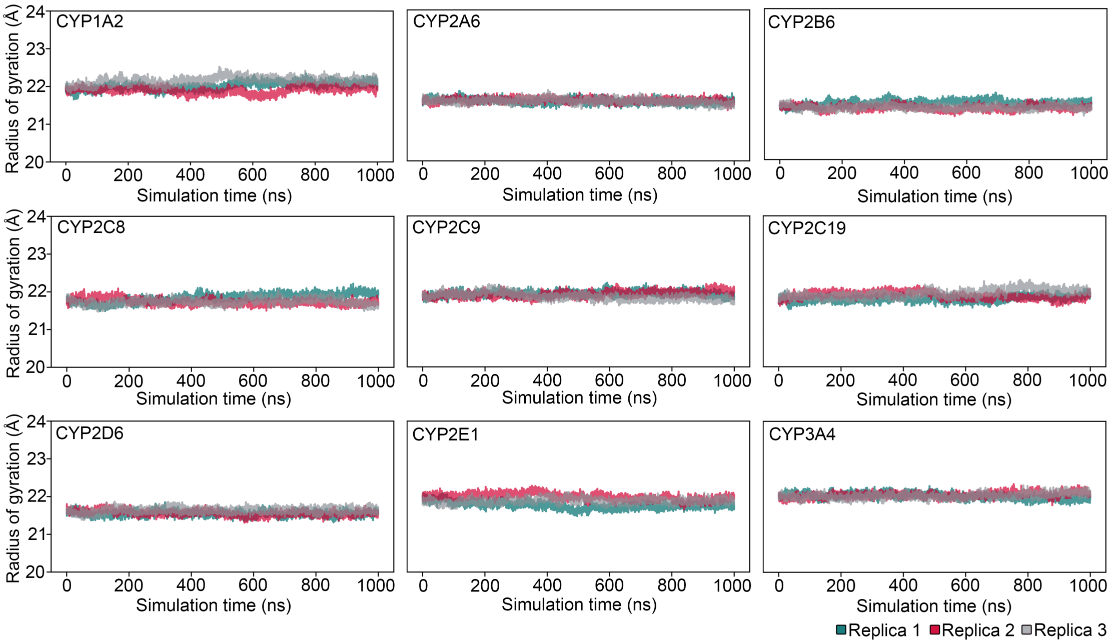


**Figure S 12** Radius of gyration during sampling simulations.

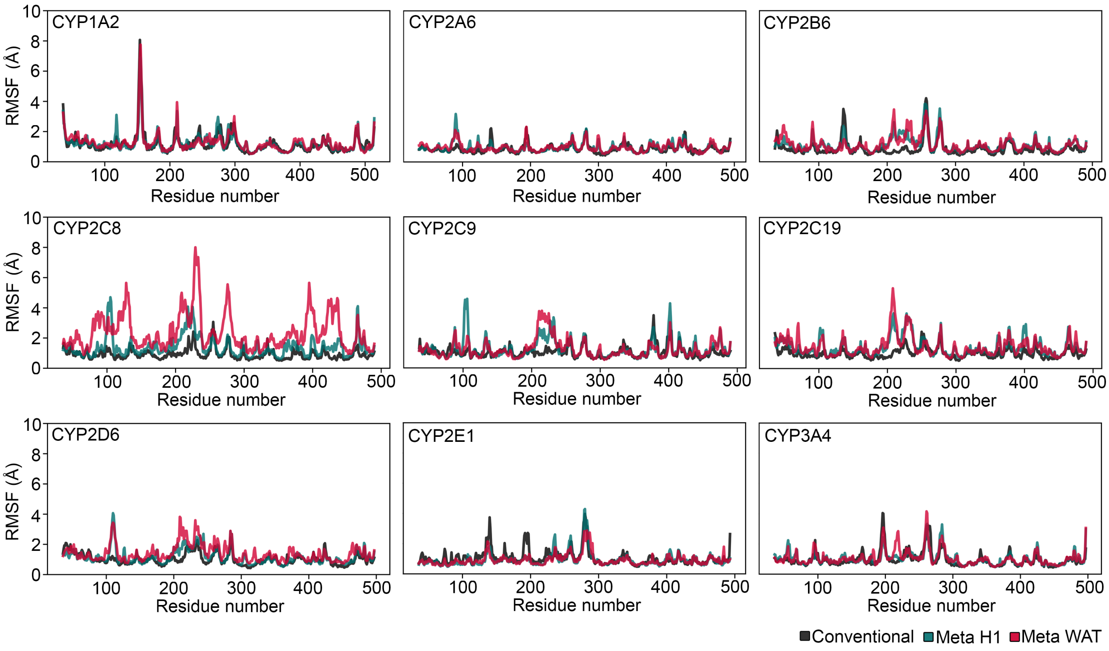


**Figure S 13** RMSF of metadynamics simulations compared to the ones obtained from conventional MD.

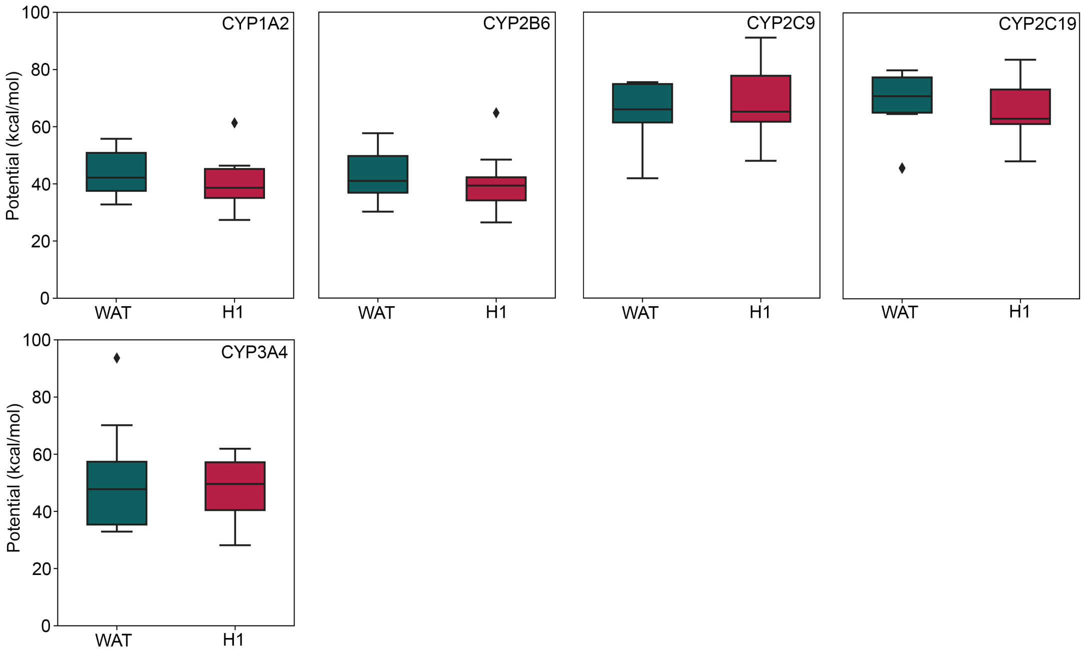


**Figure S 14** Boxplots of maximal potential ($P_{max}$) registered during metadynamics simulations.

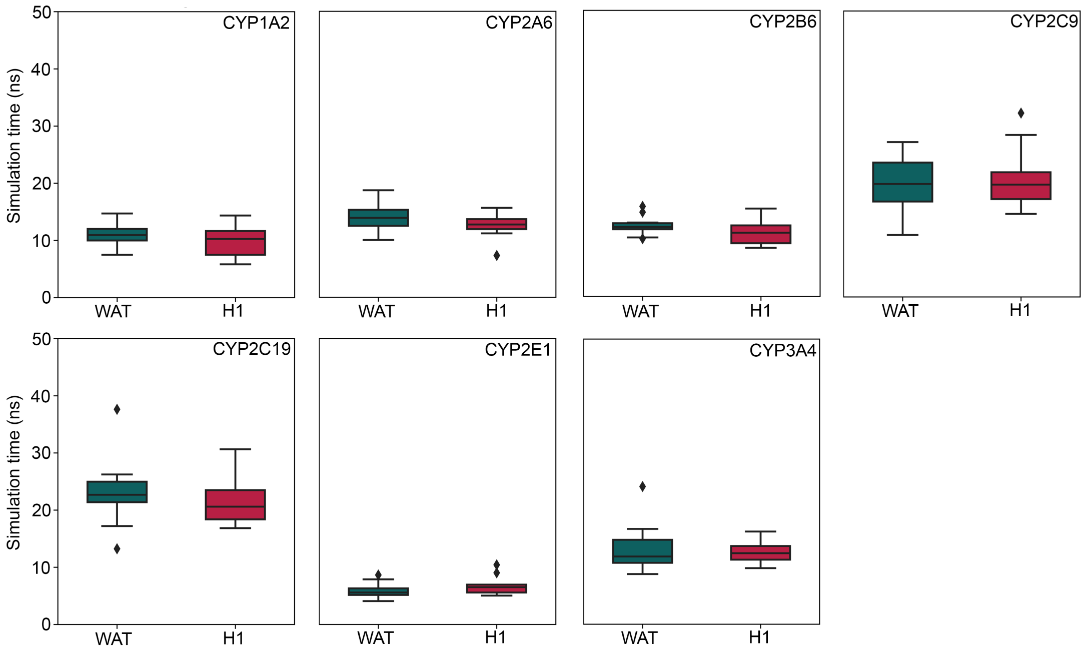


**Figure S 15** Boxplots of simulations times ($\Delta T$) registered during metadynamics simulations.

**Table S 4** Statistics of maximal biasing potential.

| | | | | $P_{max}$ (kcal/mol) | | ΔT (ns) | |
|---|---|---|---|---|---|---|---|
| **Enzyme** | **$Lig_{allo}$** | **Tunnel** | **n** | **Mean** | **SD** | **Mean** | **SD** |
| CYP1A2 | yes | 2c | 5 | 42.794 | 12.920 | 10.659 | 2.790 |
| CYP1A2 | no | | 2 | 44.104 | 10.188 | 10.883 | 0.909 |
| CYP1A2 | yes | 3 | 5 | 38.676 | 4.175 | 9.329 | 3.376 |
| CYP1A2 | no | | 8 | 44.105 | 8.363 | 11.036 | 2.228 |
| CYP1A2 | yes | all | 10 | 40.735 | 9.308 | 9.994 | 3.003 |
| CYP1A2 | no | | 10 | 44.105 | 8.119 | 11.005 | 1.989 |
| CYP2A6 | yes | 2c | 10 | 25.121 | 4.569 | 5.547 | 1.302 |
| CYP2A6 | no | | 9 | 27.076 | 3.984 | 6.051 | 1.327 |
| CYP2A6 | yes | all | 10 | 25.121 | 4.569 | 5.898 | 1.331 |
| CYP2A6 | no | | 10 | 28.111 | 4.982 | 6.318 | 1.509 |
| CYP2B6 | yes | 2c | 4 | 34.652 | 5.453 | 9.780 | 0.957 |
| CYP2B6 | no | | 2 | 36.933 | 4.627 | 10.190 | 0.148 |
| CYP2B6 | yes | 2f | 3 | 44.176 | 3.700 | 12.720 | 0.590 |
| CYP2B6 | no | | 6 | 46.432 | 10.666 | 13.232 | 1.616 |
| CYP2B6 | yes | 4 | 2 | 51.483 | 18.815 | 13.258 | 2.952 |
| CYP2B6 | no | | 2 | 38.937 | 3.959 | 12.083 | 0.541 |
| CYP2B6 | yes | all | 10 | 40.056 | 10.977 | 11.228 | 2.185 |
| CYP2B6 | no | | 10 | 43.033 | 9.329 | 12.394 | 1.748 |
| CYP2C8 | yes | 2c | 6 | 48.788 | 5.873 | 12.398 | 2.617 |
| CYP2C8 | no | | 3 | 52.085 | 11.539 | 19.040 | 11.293 |
| CYP2C8 | yes | 2f | 3 | 63.399 | 6.637 | 20.468 | 2.088 |
| CYP2C8 | no | | 2 | 71.964 | 8.305 | 22.793 | 0.541 |
| CYP2C8 | yes | all | 10 | 54.716 | 9.358 | 14.729 | 4.532 |
| CYP2C8 | no | | 10 | 75.437 | 25.199 | 27.575 | 13.362 |
| CYP2C9 | yes | 2c | 7 | 67.903 | 12.373 | 19.571 | 4.760 |
| CYP2C9 | no | | 3 | 61.515 | 13.990 | 16.127 | 5.315 |
| CYP2C9 | yes | all | 10 | 67.941 | 14.328 | 20.658 | 5.681 |
| CYP2C9 | no | | 10 | 64.413 | 12.068 | 19.642 | 4.950 |

**Table S 5** Statistics of maximal biasing potential (continued).

| | | | | $P_{max}$ (kcal/mol) | | $\Delta$T (ns) | |
|---|---|---|---|---|---|---|---|
| **Enzyme** | **$Lig_{allo}$** | **Tunnel** | **n** | **Mean** | **SD** | **Mean** | **SD** |
| CYP2C19 | yes | 2f | 2 | 88.222 | 34.742 | 25.560 | 7.184 |
| CYP2C19 | no | | 3 | 77.834 | 21.233 | 27.447 | 8.862 |
| CYP2C19 | yes | 4 | 7 | 62.884 | 12.778 | 19.848 | 2.837 |
| CYP2C19 | no | | 5 | 68.267 | 13.106 | 22.229 | 3.391 |
| CYP2C19 | yes | all | 10 | 68.657 | 18.818 | 21.355 | 4.160 |
| CYP2C19 | no | | 10 | 70.053 | 16.574 | 23.196 | 6.370 |
| CYP2D6 | yes | 2c | 8 | 39.114 | 9.468 | 11.098 | 1.452 |
| CYP2D6 | no | | 6 | 56.748 | 9.729 | 19.781 | 6.570 |
| CYP2D6 | yes | all | 10 | 40.332 | 8.952 | 11.033 | 1.433 |
| CYP2D6 | no | | 10 | 59.120 | 11.906 | 21.618 | 5.997 |
| CYP2E1 | yes | all | 10 | 25.989 | 5.955 | 6.815 | 1.723 |
| CYP2E1 | no | | 10 | 21.144 | 5.937 | 5.943 | 1.411 |
| CYP3A4 | yes | all | 10 | 48.443 | 11.299 | 12.625 | 2.010 |
| CYP3A4 | no | | 10 | 51.072 | 19.207 | 13.513 | 4.434 |

## Supporting Materials and Methods

### Model building

**Table S 6** Structure overview.

| Enzyme | PDB ID | UniProt ID | Mutations to obtain wild-type |
|---|---|---|---|
| CYP1A2 | 2HI4 | P05177 | none |
| CYP2A6 | 1Z10 | P11509 | none |
| CYP2B6 | 5UAP | P20813 | D28G, R29K, Y226H, K262R |
| CYP2C8 | 2NNI | P10632 | none |
| CYP2C9 | 1OG5 | P11712 | K206E, I215V, C216Y, S220P, P221A, I222L, I223L, G296K |
| CYP2C19 | 4GQS | P33261 | V490I |
| CYP2D6 | 3TDA | P10635 | A31G, R32K, Y33L |
| CYP2E1 | 3GPH | P05181 | N31K |
| CYP3A4 | 5TE8 | P08684 | L22A |

Accession codes for Protein DataBank and UniProt database given for all enzymes along with the amino acid mutations to obtain the wild-type sequence.

**Table S 7** Two-dimensional structures of ligands studied in this work.

**Table S 8** Substrates distributed around the enzymes.

| Enzyme | Residues |
|---|---|
| CYP1A2 | 16x acetaminophen, 4x caffeine |
| CYP2A6 | 15x acetaminophen, 5x nicotine |
| CYP2B6 | 3x quinoline, 10x propofol, 7x nicotine |
| CYP2C8 | 10x ibuprofen, 5x propofol, 5x nicotine |
| CYP2C9 | 15x acetaminophen, 5x ibuprofen |
| CYP2C19 | 15x phenacetin, 5x nicotine |
| CYP2D6 | 20x acetaminophen |
| CYP2E1 | 15x acetaminophen, 2x phenacetin, 3x chlorzoxazone |
| CYP3A4 | 16x acetaminophen, 4x chlorzoxazone |

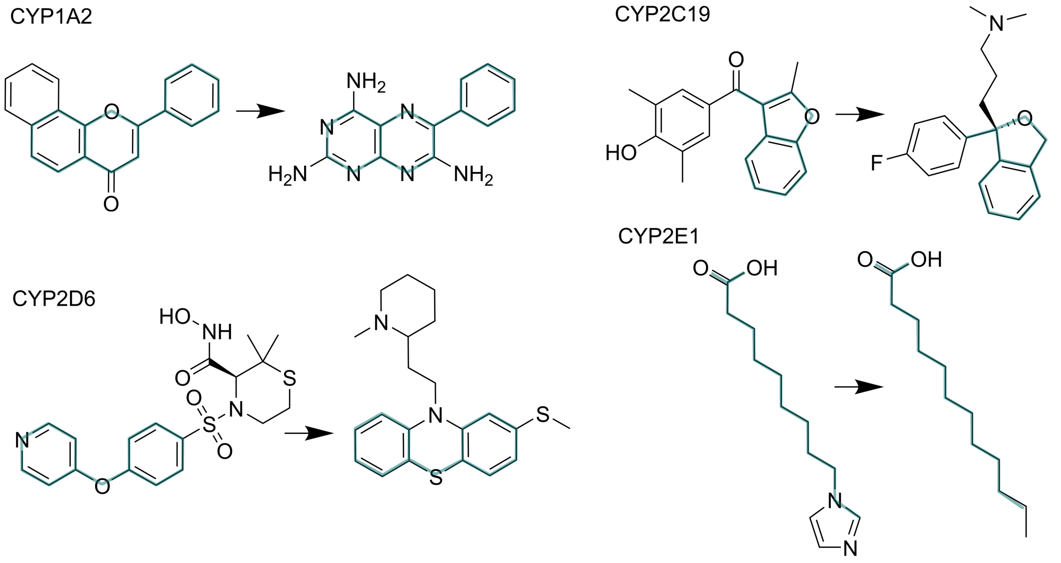


**Figure S 16** Ligands that were manually replaced by superposition during model building procedures based on the highlighted common scaffolds, which are depicted in pine green.

**Table S 9** Ligands studied in this work to H1 site.

| Enzyme | Allosteric | Orthosteric |
|---|---|---|
| CYP1A2 | acetaminophen | triamterene |
| CYP2A6 | acetaminophen | coumarin |
| CYP2B6 | acetaminophen | ZINC49942680 |
| CYP2C8 | ibuprofen | montelukast |
| CYP2C9 | acetaminophen | (S)-warfarin |
| CYP2C19 | phenacetin | 60122187[a] |
| CYP2D6 | acetaminophen | thioridazine |
| CYP2E1 | acetaminophen | undecanoic acid |
| CYP3A4 | acetaminophen | midazolam |

[a] PubChem identifier (compound ID) given.

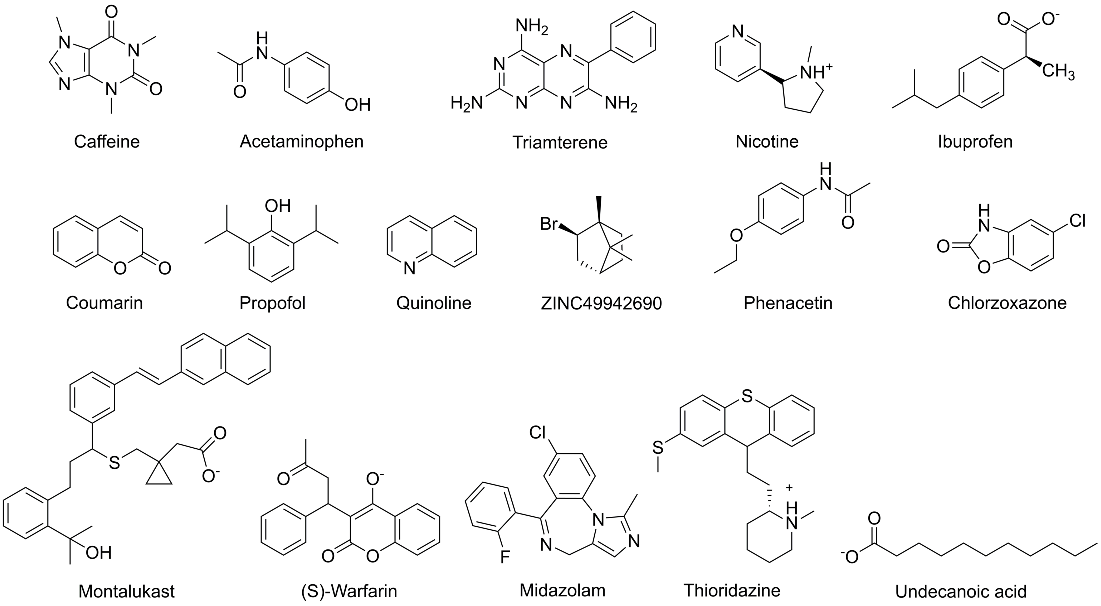


**Figure S 17** Two-dimensional structures of ligands studied in this work.

## MD simulations

**Table S 10** Residues for harmonic distance restraints.

| Enzyme | Helix C | Helix E | Helix I |
|---|---|---|---|
| CYP1A2 | A140 | V199 | V311 |
| CYP2A6 | S131 | V181 | T295 |
| CYP2B6 | S128 | I178 | T292 |
| CYP2C8 | S127 | V177 | V291 |
| CYP2C9 | S127 | V177 | A291 |
| CYP2C19 | S127 | V177 | A291 |
| CYP2D6 | S135 | V185 | V299 |
| CYP2E1 | S129 | V179 | V293 |
| CYP3A4 | L133 | V183 | I300 |

## Evaluation of the MD trajectories

**Table S 11** Residues for tunnel computation starting points.

| Enzyme | Residues |
|---|---|
| CYP1A2 | A230, D313 |
| CYP2A6 | T212, N297 |
| CYP2B6 | I209, S294 |
| CYP2C8 | L208, L294 |
| CYP2C9 | L233, D293 |
| CYP2C19 | V208, D293 |
| CYP2D6 | E216, D301 |
| CYP2E1 | L210, D295 |
| CYP3A4 | L216, I301 |

**Table S 12** Occupancy of H1 site during free MD.

| Enzyme | Replica 1 | Replica 2 | Replica 3 |
|---|---|---|---|
| CYP1A2 | yes | no | yes |
| CYP2A6 | yes | yes | yes |
| CYP2B6 | yes | yes | no |
| CYP2C8 | yes | yes | yes |
| CYP2C9 | yes | yes | yes |
| CYP2C19 | yes | yes | yes |
| CYP2D6 | yes | yes | yes |
| CYP2E1 | no | yes | no |
| CYP3A4 | yes | yes | yes |

## CHAPTER 4

# Computational Prediction of Ester Hydrolysis by Human Carboxylesterases 1 and 2

Esters are abundant in drug-like molecules. During my involvement in several virtual screening projects I realized a lack of predictive methods for ester hydrolysis. Due to my experience with metabolic enzymes as well as previous research conducted by Prof. Beat Ernst and colleagues, it was natural to follow up on this issue. The presented work focused on the development of a model predicting the specificity of ester compounds for the carboxylesterases hCE-1 and hCE-2. Several aspects of molecular recognition are involved in governing the specificity of ligands for these enzymes. This topic is of high relevance for the development of ester prodrugs due to the significantly different expression pattern of the enzymes. Even though the work is primarily centered around cheminformatics methodology, this project allowed me to leverage a broad range of modeling skills I previously acquired.

---

**Author contributions:** Conceptualization, A.F.; methodology, A.F., P.R.; formal analysis, A.F., P.R.; writing and original draft preparation, A.F.; writing, review and editing, A.F., P.R., M.A., M.S.; visualization, A.F.; supervision, M.A., M.S.

---

*Based on a manuscript submitted to J. Cheminformatics:*

Fischer, A.; Peter, R; Lill, M.A.; Smieško M., Computational Prediction of Ester Hydrolysis by Human Carboxylesterases 1 and 2.

# Abstract

Human carboxylesterases (hCE) are responsible for the majority of hydrolytic reactions on drugs *in vivo*. While predictive models exist for other phase-I metabolic enzymes, predictions on substrate specificity of the two main isoforms hCE-1 and hCE-2 were previously disregarded. Knowledge on compound selectivity for either isoform is crucial for a successful prodrug approach due to the predominance of hCE-2 over hCE-1 in the small intestine. To achieve controlled systemic release of the active principle from a prodrug, selective hydrolysis by hCE-1 rather than hCE-2 is desired. Here, we applied a combination of ligand-based and structure-based computational methods for training a machine learning classifier to predict likelihood of compound hydrolysis by either of these two enzymes. Our model achieved an accuracy of 92% during internal validation and 86% when challenged with an external test set. Among the most relevant features for the predictions were metrics describing the acyl and alcohol moieties of a compound, quantum mechanical descriptors, steric indices, and structure-based metrics. In contrast to the current rationale for substrate specificity, we found the topological polar surface area of the hydrolytic products to outperform metrics describing their size. The present study may advance the rational design of prodrugs and contribute to an improved prediction of drug-drug interactions at an early stage of drug discovery.

# Introduction

Esterification of compounds featuring a carboxylic acid or a free hydroxyl group to the respective ester prodrug is a frequently exploited technique of medicinal chemistry to overcome limitations such as low bioavailability by improving the passive transport of the typically more lipophilic ester. Optimally, after uptake from the gastrointestinal tract (GIT) the active principle of the drug should be released in a controlled manner by hydrolysis[1, 2, 3, 4]. Even though other esterases such as butylcholinesterase or acetylcholinesterase are present in humans, ester prodrugs are predominately hydrolyzed by human carboxylesterases (hCEs) with well-established examples such as methylphenidate, oseltamivir, clopidogrel, irinotecan, and angiotensin-converting enzyme inhibitors [5, 3, 6, 1, 7]. Moreover, the two main isoforms hCE-1 and hCE-2 can be responsible for the metabolic inactivation of drugs to facilitate their excretion and have also been implicated in endogenous processes such as cholesterol homeostasis. The latter renders hCEs as drug target for hypertriglyceridemia or diabetes in addition to the potential modulation of drug metabolism by their selective inhibition [1, 8, 9, 10]. Due to their different tissue distribution, with hCE-1 being mainly expressed in the liver and hCE-2 mostly limited to the GIT, knowledge on substrate specificity of the enzymes becomes pivotal for the success of a prodrug approach. In this regard, it was speculated that the substrate specificity is influenced by the size of the acyl and alcohol moieties forming the ester compound. While hCE-1 seems to prefer compounds with a small alcohol group, hCE-2 prefers cleaving off small acyl groups, as a result of the shape of their active sites and, correspondingly, steric limitations [1, 3, 5, 8]. In addition to therapeutic drugs, hCEs are also responsible for the metabolism of narcotics such as heroin or cocaine [8, 11]. The metabolic reactions of the latter exemplarily illustrate the substrate specificity of hCE-1 and hCE-2 regarding their preference for differently sized alcohol and acyl moieties (Figure 1A). Further, as pharmacokinetic drug-drug interactions can occur if multiple compounds bind to the same metabolic enzyme, information on metabolic pathways is crucial, especially as a large share of clinical drug candidates fail due to a poor pharmacokinetic profile [3, 12, 13, 14]. Altogether, the prediction of hCE-mediated metabolism early in the discovery stage is highly beneficial for designing of compounds with optimal pharmacokinetics and prodrugs with controlled systemic release.

Due to their comparatively low cost and high throughput, computational methods have been developed to predict metabolic reactions and to design compounds with an appropriate pharmacokinetic profile [12, 15, 16]. While there are several tools available to predict the regioselectivity and reactivity of substrates for cytochrome P450 enzymes or glutathion-S-transferase, there is a lack of such predictive tools for esterases [16, 17, 18]. Computational approaches for predictive metabolism include quantitative structure-activity relationships (QSAR), pharmacophore-based, structure-based, and rule–based methods [16, 18]. In the field of esterases, some effort was directed to establishing predictive models to determine selective inhibitors with regard to drug-drug interactions as well as potential therapeutics. While one study focused on a combination of structure-based considerations and machine learning [10], another one introduced a

2D-QSAR model for protease inhibitors [19]. The only efforts towards the prediction of hydrolysis by hCE-1 and hCE-2 were limited to a small set of 40 compounds with a focus on quantitative data [12, 13]. Further, the predefined test set was composed of compounds structurally highly similar to those in the training set, including stereoisomers. Other computationally assisted studies only focused on small congeneric series or specific substrates [20, 21]. Thus, there remains a need to develop tools predicting if a compound is metabolized by either hCE-1 or hCE-2, especially to determine if the hydrolysis takes place systemically or already in the GIT after oral administration.

In this study, we compiled a library of 166 compounds based on biochemical databases and the literature to train a machine learning model with structure-based and ligand-based metrics predicting if a compound is hydrolyzed by hCE-1 or hCE-2. Structure-based metrics contained information about the formation of the enzymatic transition state, as well as the distance from the catalytic serine to the reactive center (carbon atom of the carbonyl in the ester group). Ligand based metrics included semi-empirical quantum mechanics (QM) parameters and several topological descriptors characterizing structure as well as steric accessibility. The obtained model based on decision trees predicted the selectivity of a compound for hCE-1 or hCE-2 with an accuracy of 92% during internal validation and 86% when challenged with an external test set. Moreover, the applied methodology allowed us to give an estimate on how reliable the result was and which features were important for the prediction. By supplying all computed metrics and structures, we urge for the future use of the obtained data. This work will improve the rational development of prodrugs with a controlled systemic release of the active principle and the prediction of potential drug-drug interactions.

## Results and Discussion

**Library.** As previous efforts to predict hydrolysis by hCE-1 and hCE-2 were limited to a small number of structurally similar compounds [12, 13], we aimed at substantially extending our compound library in comparison. Thus, we enhanced the applicability of our model to a broader chemical space, to address this well-recognized limitation of data-driven procedures [16]. We screened several databases including PubChem BioAssay [22], DrugBank [23], BRENDA [24], the latter being specifically initiated to cover enzymatic reactions, as well as the referenced literature to collect substrates of hCE-1 and hCE-2. Based on the available data, we limited our search to esters and carbamates, even though these enzymes have also been reported to hydrolyze thioesters and amides [8, 13]. As we were focused on categorical rather than absolute quantitative data, our dataset was less prone to bias from varying experimental assay methods and related inaccuracies [25]. The resulting database consisted of 166 unique ligands, with 16 additional entries either being isomers with different stereospecificity or compounds with multiple ester groups, which we considered separately (Figure S1). Due to the promiscuity of hCE-1 and hCE-2 [1, 10], a large share of compounds can be metabolized by both hCE-1 and hCE-2 at different rates. For our training set, we retained compounds with at least least 2-fold selectivity toward either enzyme, while we grouped the ones with a selectivity factor below 2-fold in a separate set (External A). Clearly, this set is relatively challenging to predict. Further, there were several compounds, for which either hCE-1 or hCE-2 were proposed as major isoform for hydrolysis, but no selectivity data was reported in the literature, leading us to create a third set comprising of these compounds (External B). The majority of compounds included in our library were drugs and, thus, complied with the boundaries proposed by Lipinski [26]. As several compounds presented properties beyond the drug-like chemical space, which is an increasing trend for marketed drugs observed in recent years [27], the predictions of our model were centered, but not limited to drugs fulfilling the Lipinski criteria (Figure 1B). A drawback of QSAR models is the frequently limited chemical space of the considered compounds [28]. To verify the chemical diversity of the ligands in our database, we computed pair-wise similarity based on extended-connectivity fingerprints (ECFP2) followed by comparison of Tanimoto coefficients. The obtained average values ranged from 0.16 to 0.37 indicating high chemical diversity within the training set, which is important for applicability of our model to new compounds [29]. The separated, rightmost peak on the histogram represented a congeneric series of statin derivatives (Figure 1B). Next, we computed the average and maximal similarity of the external sets A and B to the training set, assessing if the model was trained with similar compounds (Figure S2). While the majority of compounds in both external A and B sets presented a comparable similarity within them similar to the training set, the external set A included several (approximately 50%) similar compounds if the training set was taken as a reference. Some degree of similarity between the compounds can be attributed to congeneric series assessed in the underlying studies. In contrast, the external set B presented low similarity to the training set with the majority of compounds having a maximal Tanimoto coefficient of 0.5 or lower.

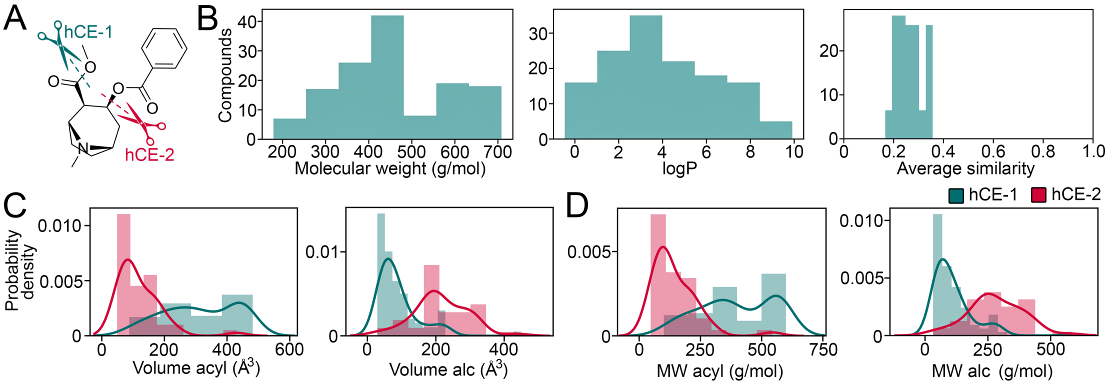


**Figure 1** Library characteristics and ligand-based descriptors. (A) Specificity of hCE-1 and hCE-2 for the ester groups of cocaine. (B) Molecular weight, logP, and pairwise similarity among compounds in the training set. (C) Molecular volume of acyl and alcohol moieties in the training set. (D) MW of alcohol and acyl moieties in the training set.

**Ligand-based metrics.** The majority of tools for predicting metabolic reactions are centered around ligand-based descriptors [16, 17, 30]. As our focus was the distinction between hCE-1 and hCE-2 substrates, we initially relied on the proposed rationale for their substrate specificity based on the size of the acyl and alcohol moieties [1, 3, 5, 8]. To quantify the preference, we determined the products of the hydrolysis and computed their molecular weight (MW), volume, number of atoms, logP, and topological polar surface area (tPSA). Further, we included the ratios of these properties between acyl and alcohol products. As depicted in Figures 1C and D, the volume or MW of the acyl and alcohol moieties did not allow to completely distinguish hCE-1 or hCE-2 substrates, albeit there was a considerable degree of separation between the sets. An analysis of frequently occurring acyl and alcohol moieties additionally confirmed the specificity rationale in the literature (Figure S3). When we trained a random forest (RF) classifier to predict if a compound is hydrolyzed by hCE-1 or hCE-2 based on the above-mentioned features, we obtained a Matthews correlation coefficient (MCC) of 0.79±0.12 during internal validation and 0.51 and 0.35 for the external sets A and B, respectively (Table S1). Even though the results were promising, we aimed to improve the prediction by including additional features. In general, topological descriptors can be used to translate the chemical constitution of a compound into numerical values [31]. However, many descriptors lack physical interpretability for the prediction of hydrolysis reactions, which is a drawback recognized in the QSAR field [28, 29]. As we will elaborate in the following section, the active sites of hCE-1 and hCE-2 vary in size, which is thought to contribute to their substrate specificity [21, 32]. Thus, features describing the size of compounds such as molecular eccentricity for both the whole molecule (Figure 2A), as well as the acyl and alcohol products, might affect the prediction. Using these three eccentricity features, we trained a RF classifier for our prediction task resulting in MCC values of 0.68±0.15, 0.57, and 0.35 for the training set and the external test sets A and B, respectively (Table S2). Moreover, we selected the number of rotatable bonds of the acyl and alcohol moieties as an additional feature, as they account for the flexibility of a substructure which might compensate steric limitations described by the above-mentioned descriptors. Indeed, some compounds could be separated regarding their selectivity for hCE-1 or hCE-2 based on the number of rotatable bonds (Figure 2B).

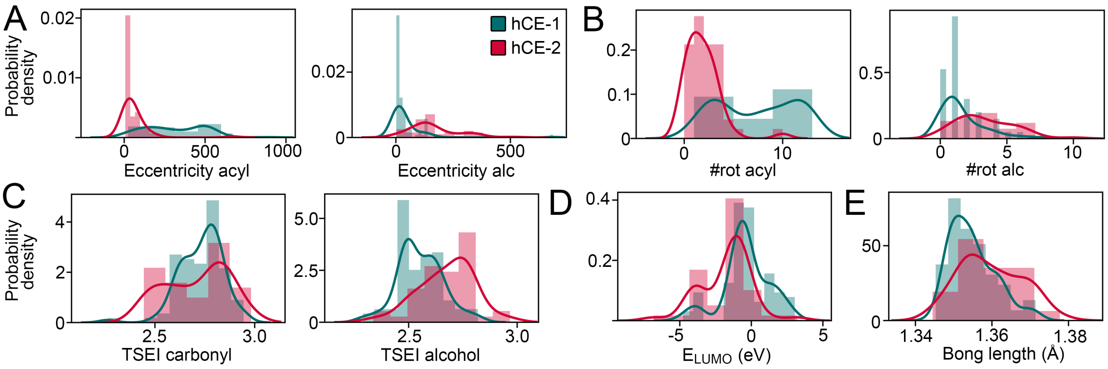


**Figure 2** Ligand-based descriptors. (A) Number of rotatable bonds (#rot) of acyl and alcohol moieties for the training set. (B) Eccentricity of acyl and alcohol moieties for the training set. (C) TSEI of the carbonyl carbon atom, as well as the oxygen atom of the alcohol moiety of the compounds. (D) LUMO energy for the training set. (E) Bond length between the carbonyl carbon atom and the oxygen of the alcohol moiety of compounds in the training set.

Due to the underlying enzymatic mechanism, which is based on a nucleophilic attack of the substrate, the steric accessibility of its carbonyl atom is a crucial factor for a successful hydrolysis. It was described that steric hindrance of the ester or carbamate group reduces the reaction rate for hydrolysis [5]. Thus, we computed the topological steric effect index (TSEI) according to Cao *et al.* describing the relative specific volume of a reaction center screened by its surrounding atoms [33]. While TSEI is typically evaluated for reactive centers of a ligand, we additionally computed this metric for the surrounding atoms of the ester or carbamate groups. The visualization of the distribution of TSEI values revealed a slight separation of hCE-2 ligands (Figure 2C) leading us to retain these metrics for further procedures.

When estimating the reactivity of compounds, QM methods have been applied to predict various enzymatic reactions [17, 21, 32, 34]. During the nucleophilic attack of the catalytic serine, electrons are transferred from its highest occupied molecular orbital (HOMO) to the lowest unoccupied molecular orbital (LUMO) of the substrate in the so-called frontier molecular orbital approach [21, 32]. Correspondingly, we computed the LUMO energy of the substrates using semi-empirical QM methodology with PM7 parameterization. Further, we evaluated the bond length of the carbonyl double bond, as well as the single bond between the carbonyl carbon atom and the oxygen atom of the alcohol moiety. The latter bond is cleaved during the catalytic reaction, while the former undergoes a change of bond order during hydrolysis (Figure S4). The semi-empirical approach was selected due to the comparatively short computation time. Although the contribution of reactivity to the substrate specificity between hCE-1 and hCE-2 is not obvious, as both enzymes share the same catalytic mechanism [1, 6], we could detect a slight separation of the LUMO energy and C-O bond length between compounds selective for either of the esterases (Figures 2D and 2E). In addition, we computed the atomic charges of the carbonyl function, the nucleophilic delocalizability ($D_N$) of the carbonyl carbon atom, as well as the hardness of the substrates [17, 35, 36]. Using both the steric considerations and the QM parameters, we trained a RF model resulting in a MCC values of 0.57±0.13 (training set), 0.48 (external set A), and 0.25 (external set B). Even though this model performed worse than the ones based on topological or acyl/alcohol descriptors, it could distinguish hCE-1 from hCE-2 substrates at an accuracy of approximately 82% during internal cross-validation (Table S3).

**Structure-based metrics.** Due to the different topology of the active sites between hCE-1 and hCE-2 [12, 21], structure-based techniques might allow to separate between substrates for either enzyme. The volume of the binding site has been reported to be larger in hCE-2 compared to hCE-1 [21, 32]. Indeed, our calculation of the binding site volumes using SiteMap confirmed this observation (Figures 3A and 3B). However, when the eccentricity was computed for the substrates in our training set, the compounds primarily hydrolyzed by hCE-1 presented higher values. This further underlined the flexibility of the hCE-1 active site, as it needs to structurally adapt to a broad range of differently sized substrates. The sequence identity between hCE-1 and hCE-2 amounts to 48% (Figure S5) [1]. As there was no crystal structure available for hCE-2, we constructed several homology models based on different hCE-1 template structures and evaluated them in a decoy docking approach. The obtained quality estimates of the homology models indicated a high nativeness of the structures (Table S4). Structural differences between the homology models derived from the two algorithms were apparent around residues E105, M309, S429, and K475 (Figure S6). In addition, the models were subjected to short molecular dynamics simulations in order to obtain representative structures. The representative time-evolved structure of the model (Template PDB ID: 1MX1) we generated using SWISS-MODEL, presented the best area under the curve (AUC) of the receiver operating characteristic (ROC) in the decoy docking validation (Table S5). Importantly, we ensured to model the resting state of the enzyme with a hydrogen bond between the catalytic serine and histidine residues, as well as an additional one between the glutamic acid and histidine during the preparation of all structures (Figure S4B). Similar to previous studies [13, 20], we detected a difference of a loop in vicinity to the active site (Figure 3C) at residues Asp307 to Thr321 in hCE-1. In hCE-2 this loop, which is located at the putative entrance of the catalytic cavity, was absent, potentially contributing to the substrate specificity of hCE-1 and hCE-2. The remaining parts of the structures were highly similar (Figure S7). As there were several crystal structures available for hCE-1, we selected the most promising candidates based on their performance in reproducing binding modes of cocrystallized ligands using two different docking protocols (Figure S8). The smina docking protocol produced superior results regarding pose prediction, with one structure reproducing six of seven cocrystallized ligands to a satisfactory degree (RMSD below 3.5 Å). Again, the most promising structures were subjected to a decoy docking procedure. Unfortunately, the highest

AUC value we could reach was 0.524, indicating only a very slight discrimination of the 235 actives from the 14330 decoys (Tables S5 and S6). Despite this drawback, the binding modes of chemically diverse ligands could be reproduced, leading us to retain the respective structure for further procedures.

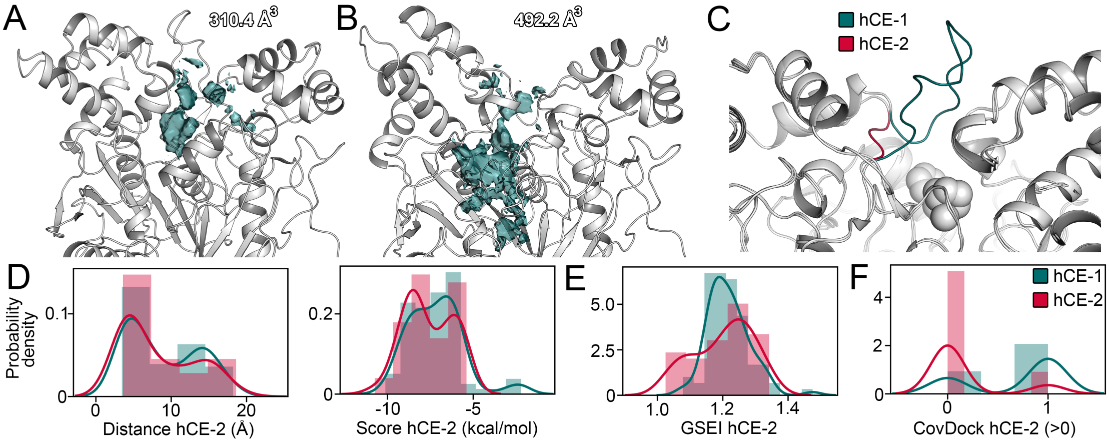


**Figure 3** Structure-based considerations. (A) Active site volume of hCE-1. (B) Active site volume of hCE-2. (C) Comparison of the hCE-1 crystal structure (PDB ID: 1MX1) and the selected hCE-2 homology model. (D) Distance between the reactive centers in docked poses of hCE-2 (left) and their docking scores. (E) GSEI of docked poses to hCE-2. (F) Boolean parameter of covalent docking was successful for hCE-2.

After the validation of the protocols, we docked our substrate library to the selected structures. As in several previous studies [21, 13, 20, 12], we computed the distance between the side chain oxygen of the catalytic serine residue to the carbonyl carbon atom of each substrate docked to either of the esterases. The inspection of the binding modes of two highly selective substrates for hCE-1 revealed the ester groups distant from the catalytic center, if they were docked to hCE-2 (Figure S9). Nevertheless, both the docking scores, as well as the computed distances did not properly separate the substrates according to their selectivity besides a limited number of hCE-1 compounds presenting bad scores in hCE-2 (Figures 3D and S10). Thus, we tried to ameliorate the results by computing the geometric steric effect index (GSEI), similar to the ligand-based procedures where we computed the TSEI parameter. Even though the GSEI values of the carbonyl atoms based on the hCE-2 docking poses presented improved separation of the substrates as opposed to the distances and docking scores (Figure 3E), a considerable overlap between them remained. Next, we applied a covalent docking approach based on the Glide engine [37] to model the tetrahedral transition state formed after the attack of the catalytic serine. Due to the high computational expense of the protocol, we used it in enrichment mode and determined if the algorithm could produce a valid pose with the selected structures (Figure 3F). As we also validated the use of the Glide protocol, we used the best-performing ensemble of structures to reproduce a maximal number of cocrystallized binding modes for hCE-1. Further, we selected three structures of hCE-2, as decoy docking using the Glide protocol revealed no structure with superior AUC values (Table S7). Using the described structure-based metrics, we trained a RF classifier which presented MCC values of 0.47±0.12 for the training set during internal validation and 0.46 and 0.32 for the external sets A and B, respectively. While the accuracy of the predictions between 66 and 77% were the lowest among the tested features (Table S8), the structure-based metrics still allowed to distinguish a reasonable number of substrates.

**Predictive Model.** In recent years, machine learning has received great attention in the prediction of metabolic reactions [16, 17, 18, 30]. Using all descriptors we introduced above, we evaluated several algorithms suitable for classification tasks including RF, extreme gradient boosting (XGBoost) decision trees, k-nearest neighbors (k-NN), support vector machine (SVM), linear discriminant analysis (LDA), and logistic regression [38, 39, 40, 41]. For RF, XGBoost, k-NN, and SVM, we conducted a grid search to determine the optimal hyperparameters maximizing the MCC. For LDA and logistic regression, we retained default parameters. To determine which approach was superior, we conducted a 5-fold internal cross-validation with the randomly shuffled training dataset. As RF, k-NN, and XGBoost performed best during internal validation, we subjected the obtained models to the external test set A and found RF to be optimal (Tables S9 and S10). In the final step, we aimed to reduce the number of features by recursive elimination and retained 28 features producing a maximal MCC (Tables S11 and S12).

**Table 1** Performance metrics of the final RF model.

| **Set** | **hCE-1**[a] | **hCE-2**[a] | **Misclassified** | **Accuracy** | **AUC** | **MCC** |
|---|---|---|---|---|---|---|
| Training | 90 | 46 | 11 | 0.92±0.06 | 0.97±0.03 | 0.84±0.12 |
| External A | 16 | 12 | 4 | 0.86 | 0.85 | 0.73 |
| External B | 13 | 4 | 4 | 0.76 | 0.79 | 0.35 |

[a] Number of entries in set.

In QSAR modeling, internal and external validation has been suggested as good practice [29, 39, 42]. During internal validation, we found an accuracy of 92±0.06% (Table 1) as well as high AUC and MCC, indicating good discrimination performance. Next, we conducted an external validation using the two test sets that were not used for training. For the External A set, which included compounds with below 2-fold selectivity and, thus, was inherently more challenging to predict, we found a slightly decreased performance with an accuracy of 86%, as well as lower AUC and MCC (Figure 4A). However, the metrics were still acceptable and close to the standard reported by comparable computational approaches focused on different enzyme systems [16, 17, 30]. In the external set B, we included compounds for which no selectivity testing was conducted, but a major isoform involved in ester hydrolysis was suggested. For this set, the performance metrics dropped again, which could be explained by the reduced confidence into the proposed main driver of hydrolysis. Still, the accuracy reached 76%, which we deemed acceptable for this specific test set. As discussed above, the majority of substrates in the external libraries were dissimilar to those in the training set, underlining the applicability of our predictions to a broad chemical space. The complete performance metrics are given in Table S13. To obtain more insight into the importance of individual features, we analyzed the decision trees in our final RF model. As we already identified in feature engineering, the description of the acyl and alcohol hydrolysis products, as well as the respective ratios were among the top six features (Figure 4B). This stands in accordance with the frequently described rationale that hCE-1 prefers small alcohol moieties, while hCE-2 prefers small acyl moieties [1, 3, 5, 43]. Interestingly, the ratio between the tPSA of the acyl and alcohol moieties was the most important feature, suggesting fragment contributions of polar atoms [44] to be more relevant than their size or volume alone. Reactivity parameters obtained from semi-empirical QM calculations such as the LUMO energy and the charge of the carbonyl carbon were also among the features having a high information value. Additionally, features stemming from structure-based considerations including covalent and conventional docking contributed positively to the accuracy of our model.

Vistoli and colleagues reported a regression equation distinguishing hCE-1 from hCE-2 substrates based on logP as well as the volume of alcohol and acyl moieties [13]. When we tested the respective equation with our training dataset, we obtained an accuracy of only 66.2%, demonstrating the superiority of our approach.

Next, we investigated compounds that were misclassified by our model. Our training set contained eight compounds with multiple stereoisomers for which a different metabolic fate was reported. Among them were propranolol derivatives [45] and pyrethroid insecticides [46], which are difficult to predict, especially if ligand-based descriptors are considered. Of those eight stereoisomers, we encountered two of them among the misclassified ligands during training. In analogy, two of the four misclassified compounds in the external set A also present stereospecific selectivity for hCE-1 and hCE-2, indicating a large share of the outliers to be based on small differences in their binding modes (Tables S14 and S15). The remaining outliers in the external set A were eslicarbamazepine acetate and fosinopril, for which we selected the Shapley additive explanations (SHAP) method [47] to extract additional insight into the classification process (Figure S11). In both cases, the ratios of tPSA, volume, and MW between acyl and alcohol moieties contributed the most toward the misclassification, while none of the structure-based features appeared in the SHAP analysis. The outliers in the external set A displayed a comparatively low similarity to the compounds in the training set (Table S15). The underlying RF model allowed us to estimate a probability for the respective outcomes based on the voting of individual trees in the forest. Regarding the prediction probabilities, the misclassified compounds in the training set generally presented lower values (statistically significant) indicating this readout to be useful to identify potential outliers (Figure S12). If the probability threshold was selected above 0.9, the number of outliers could be minimized to only two in the training set, and only one in the external sets. However, by applying this threshold, 58.8%, 46.4%, and 47.1% of compounds could not be predicted in the training set, the external sets A and B, respectively. In the comparatively smaller external set A containing the more challenging compounds, however, such a trend could not be deduced.

To ensure the reliability of our predictions, we considered additional validation procedures as recommended in the literature [29, 39, 42]. To decrease the probability of chance correlations, we used *y*-scrambling [29, 48], for which the selectivity labels of the training set were randomly reorganized followed by a 5-fold internal cross-

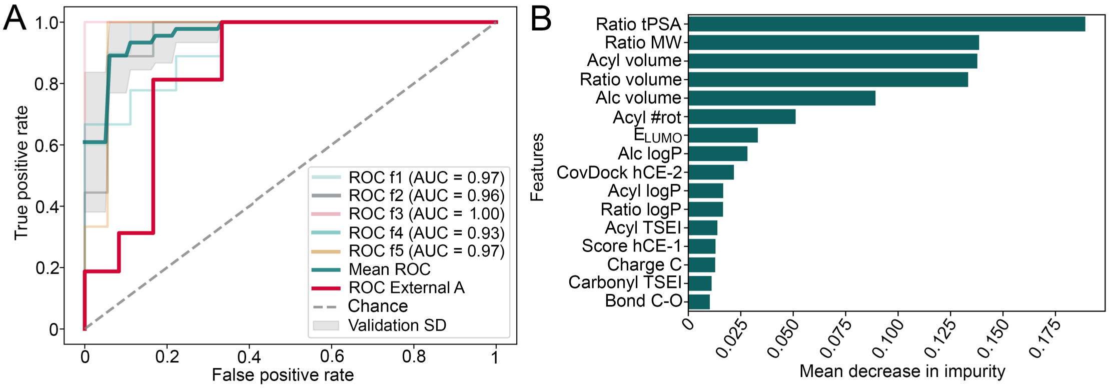


**Figure 4** Prediction metrics and features. (A) ROC curves for the training set as well as the external set A. The AUC values of the individual test sets during internal validation are indicated, together with their mean ROC curve. (B) Feature importance deduced from recursive feature elimination.

validation. All parameters presented near to random prediction performance with an AUC value of 0.43±0.11 and a MCC of -0.21±0.21, confirming the robustness and sensitivity of our final model. In conclusion, our predictive model presented a good accuracy in predicting selectivity of compounds for hCE-1 and hCE-2 based on internal and external validation, and the exclusion of chance predictions. One obvious drawback was the number of compounds, for which selectivity data was available in the literature and databases. However, as only approximately 10% of marketed drugs are prodrugs, of which 50% are activated by hydrolysis [2], the total number of compounds we could consider for our predictions was limited. As recommended [29], we provide all structures and computed metrics to support our findings.

## Materials and Methods

**Library generation.** We screened the PubChem BioAssay, DrugBank, BRENDA databases to collect substrates of hCE-1 and hCE-2 [22, 23, 24]. Using the accompanying literature, as well as additional studies, we determined if a compound is hydrolyzed and which enzyme is predominantly responsible for the reaction. Further, we documented the method as well as the source of the protein used for the experiments. If a compound presented a selectivity of at least 2-fold for either enzyme, with the exception of data obtained from experiments with unpurified microsomal enzymes for which we used 10-fold as threshold, we added it to our main dataset. The remaining compounds were separated in two test sets: (i.) a set containing compounds with a selectivity below

2-fold or with variable selectivity values among multiple publications and (ii.) a set of compounds, for which no selectivity testing was conducted, but the preference of either hCE-1 or hCE-2 was suggested in the literature.

In a next step, we either manually drew the compounds in the Maestro Small-Molecule Drug Discovery Suite (v2019-4) [49], or, if possible, obtained SMILES codes from the PubChem database. Using the LigPrep routine, we generated three-dimensional conformers with Epik predicting the protonation states at a pH of 7.4 and the OPLS3e force field for geometric optimization. In order to confirm the protonation states, we selected the Marvin suite of tools (v17.27.0) by ChemAxon [50] and retained all protomers with a predicted occurrence of at least 20%. Stereoisomers were manually selected to represent the compounds in the literature. To determine the heterogeneity of the compounds in our library, we computed pair-wise similarity based on FP2 extended-connectivity fingerprints using OpenBabel (v2.3.2) [51], compared them using the Tanimoto coefficient, and averaged the coefficients for each compound. Further, maximal values were computed as well. The similarity of the two test sets to the training set was computed in the same way with an additional comparison of maximal similarity. If multiple ester or carbamate groups were present in a single substrate, we separated them to represent individual entries. The respective assignment of these groups is given in Figure S13.

**Structure-based procedures.** Due to the absence of a crystal structure for hCE-2, we generated homology models using the SWISS-MODEL web server [52] based on a several different templates (PDB IDs: 1MX1, 2HRQ, and 5AG7) of hCE-1. Furthermore, we used the MODELLER tool [53], to obtain an additional structure (Template PDB ID: 1MX1). Next, we conducted molecular dynamics simulations of each model to derive representative structures for further procedures as described in the Supporting Materials and Methods. In the case of hCE-1, we selected 14 different crystal structures cocrystallized to 7 individual ligands (Figure S8). All protein structures were preprocessed using the Protein Preparation Wizard in Maestro [54] by adding hydrogen atoms, predicting protonation states at pH 7.4 using Epik, and assigning bond orders. Next, the hydrogen bonding network was reoriented with residue protonation prediction by PROPKA configured at pH 7.4. Finally, the structures were subjected to a restrained minimization with an RMSD convergence threshold of 0.3 Å for the protein heavy atoms. We chose to work with multiple structures for each enzyme due to the reports of high flexibility that was previously treated in an ensemble docking approach [10, 12]. Thus, we conducted a cross docking experiment using the Glide standard precision (SP) [55] and smina [56] docking protocols. For both protocols, the centroid of the search space was defined based on the mass center of a cocrystallized ligand, while we selected an exhaustiveness of 16, a random seed of 42, and a cubic box with a width of 21 Å for smina. The RMSD between the docked poses and the cocrystallized ligands was evaluated using the `rmsd.py` routine that comes with Maestro. Notably, even though one ligand was annotated as palmitic acid, the cocrystallized ligand in the crystal structure (PDB ID: 2DQZ) was nonanoic acid instead. The smina protocol presented superior pose prediction compared to Glide SP and, thus, we continued with this protocol. As an additional validation step, we evaluated the capability of the best-performing structures to distinguish known binders from random decoy molecules. Based on SMILES strings of known binders for either hCE-1 and hCE-2, we generated decoys using the DUD-E [57] web server (Table S16) and computed conformers using the LigPrep routine as described above. The ROC AUC metrics were obtained from the Screening Explorer [58] web server. Based on the validation data, we selected one structure per enzyme to study the interaction of our compound library with hCE-1 and hCE-2. Using the obtained poses, we computed the distance between the carbonyl carbon of the ester or carbamate functions and the side chain oxygen of the catalytic serine. Additionally, we selected the covalent docking protocol in Maestro based on Glide and Prime workflows [37]. We configured the protocol in enrichment mode to perform a nucleophilic addition to the carbonyl double bond, modeling the tetrahedral transition state of the first catalytic step. The remaining parameters in this protocol were left on default. From the resulting poses, we retained the predicted binding free energy and registered if no pose was found. Based on the latter information, we derived a boolean variable based on a threshold of unsuccessful docking attempts to an ensemble of structures. As we used different numbers of input structures for hCE-1 and hCE-2 (Table S6), we defined specific thresholds of four and one unsuccessful attempts to determine a valid pose for hCE-1 and hCE-2, respectively. The thresholds were selected based on inspecting the separation between the substrates into the correct categories.

**Ligand-based metrics.** We computed descriptors including molecular weight, logP, tPSA, number of rotatable bonds, as well as the molecular volume using RDKit for the whole compounds as well as the acyl and alcohol products [59]. The compounds were fragmented based on SMARTS substructure matching (Table S17). The remaining topological descriptors were computed in the Molecular Descriptors panel within Maestro. The steric accessibility metrics were computed using the cxcalc module of Marvin Sketch (v17.27.0) by ChemAxon coupled to RDKit for substructure matching. Semi-empirical QM data was obtained using MOPAC2016 [60] using PM7 parameterization coupled to a 1SCF single point calculation and the keyword for superpolarizabilites. Using RDKit, we determined the absolute charge of the ligands, while we converted the input ligands to a MOPAC compatible format using OpenBabel.

**Machine learning.** All machine learning procedures were conducted using the scikit-learn (v0.24.2) [61] module in python except for the XGBClassifier for gradient boosted decision trees [41]. The selected classification algorithms are listed in Table S18. Initially, the *y*-labels of the dataset with 46 features were encoded into binary flags and the optimal hyperparameters for RF, XGBoost, k-NN, and SVM were determined using the GridSearchCV module (Table S19). For both LDA and logistic regression, we retained default specifications. For k-NN, logistic regression, SVM, and LDA, numeric data was normalized before the prediction. Next, we examined the performance of each model for our training set using a stratified 5-fold cross-validation. As the number of hCE-2 substrates in our dataset was lower, we selected this approach to ensure that the ratio of the two classes remained constant during the train/test split procedure. Based on the obtained metrics including MCC, accuracy, and AUC, we selected RF as final model and used feature ranking with recursive elimination to reduce the number of features to 22, optimizing the MCC. In a final step, we evaluated the performance of our model using internal and external validation.

## Conclusion

In the last years, computational tools received great attention with respect to the prediction of the outcome and regioselectivity of metabolic reactions. However, the prediction of ester hydrolysis reactions, relevant for both the activation of prodrugs and the inactivation of drugs, was not previously considered. While hCE-2 is the main esterase expressed in the GIT, its expression in the liver is minor compared to hCE-1. Due to their different tissue distribution, knowledge on the selectivity of a prodrug for hCE-1 and hCE-2 becomes pivotal to control the formation of the active principle. Here, we computed a diverse set of physico-chemical meaningful features relating to the structure and topology, reactivity, steric accessibility, as well as structure-based considerations of 166 substrates. After evaluating several machine learning algorithms suited for classification tasks, we determined RF to perform optimally to predict the selectivity of a compound toward hCE-1 or hCE-2. During internal validation, we obtained a high accuracy of 92%, an AUC of 0.97, and an MCC of 0.84 for a diverse set of compounds. Moreover, when challenged with an external test set containing compounds with a selectivity factor below two, we obtained an accuracy of 86%, an AUC of 0.86, and a MCC of 0.73. Throughout our work we adhered to best practices in the QSAR, ligand-based and structure-based modeling. In contrast to previous observations, to predict selectivity we found the tPSA of the hydrolysis products to outperform metrics directly relating to their size. As for other phase-I metabolic reactions, prediction of ester hydrolysis could become a routine application in many drug discovery projects and advance the design of prodrugs with controlled systemic release, as well as the prediction of potential drug-drug interactions.

## References

[1] Teruko Imai. Human carboxylesterase isozymes: catalytic properties and rational drug design. *Drug metabolism and pharmacokinetics*, 21(3):173–185, 6 2006.

[2] Peter Ettmayer, Gordon L. Amidon, Bernd Clement, and Bernard Testa. Lessons Learned from Marketed and Investigational Prodrugs. *Journal of Medicinal Chemistry*, 47(10): 2393–2404, 2004.

[3] S Casey Laizure, Vanessa Herring, Zheyi Hu, Kevin Witbrodt, and Robert B Parker. The role of human carboxylesterases in drug metabolism: have we overlooked their importance? *Pharmacotherapy*, 33(2):210–222, 2 2013.

[4] Wojciech Schönemann, Simon Kleeb, Philipp Dätwyler, Oliver Schwardt, and Beat Ernst. Prodruggability of carbohydrates — oral FimH antagonists. *Canadian Journal of Chemistry*, 94(11):909–919, 3 2016.

[5] Masato Takahashi, Ibuki Hirota, Tomoyuki Nakano, Tomoyuki Kotani, Daisuke Takani, Kana Shiratori, Yura Choi, Masami Haba, and Masakiyo Hosokawa. Effects of steric

hindrance and electron density of ester prodrugs on controlling the metabolic activation by human carboxylesterase. *Drug metabolism and pharmacokinetics*, 38:100391, 3 2021.

[6] Masakiyo Hosokawa. Structure and catalytic properties of carboxylesterase isozymes involved in metabolic activation of prodrugs. *Molecules (Basel, Switzerland)*, 13(2):412–431, 2 2008.

[7] Shana V Stoddard, Xiaozhen Yu, Philip M Potter, and Randy M Wadkins. In Silico Design and Evaluation of Carboxylesterase Inhibitors. *Journal of pest science*, 35(3):240–249, 2010.

[8] Li-Wei Zou, Qiang Jin, Dan-Dan Wang, Qing-Kai Qian, Da-Cheng Hao, Guang-Bo Ge, and Ling Yang. Carboxylesterase Inhibitors: An Update. *Current medicinal chemistry*, 25 (14):1627–1649, 2018.

[9] Latorya D Hicks, Janice L Hyatt, Shana Stoddard, Lyudmila Tsurkan, Carol C Edwards, Randy M Wadkins, and Philip M Potter. Improved, selective, human intestinal carboxylesterase inhibitors designed to modulate 7-ethyl-10-[4-(1-piperidino)-1-piperidino]carbonyloxycamptothecin (Irinotecan; CPT-11) toxicity. *Journal of medicinal chemistry*, 52(12):3742–3752, 6 2009.

[10] Eliane Briand, Ragnar Thomsen, Kristian Linnet, Henrik B Rasmussen, Søren Brunak, and Olivier Taboureau. Combined Ensemble Docking and Machine Learning in Identification of Therapeutic Agents with Potential Inhibitory Effect on Human CES1, 2019.

[11] Jianzhuang Yao, Xiabin Chen, Fang Zheng, and Chang Guo Zhan. Catalytic Reaction Mechanism for Drug Metabolism in Human Carboxylesterase-1: Cocaine Hydrolysis Pathway. *Molecular Pharmaceutics*, 15(9):3871–3880, 2018.

[12] Giulio Vistoli, Alessandro Pedretti, Angelica Mazzolari, and Bernard Testa. In silico prediction of human carboxylesterase-1 (hCES1) metabolism combining docking analyses and MD simulations. *Bioorganic and Medicinal Chemistry*, 18(1):320–329, 2010.

[13] Giulio Vistoli, Alessandro Pedretti, Angelica Mazzolari, and Bernard Testa. Homology modeling and metabolism prediction of human carboxylesterase-2 using docking analyses by GriDock: a parallelized tool based on AutoDock 4.0. *Journal of computer-aided molecular design*, 24(9):771–787, 9 2010.

[14] Han van de Waterbeemd, Eric Gifford, and Ann Arbor. ADMET in silico modelling: towards prediction paradise? *Nat Rev Drug Discov*, 2(3):192–204, 3 2003.

[15] Myeong-Sang Yu, Hyang-Mi Lee, Aaron Park, Chungoo Park, Hyithaek Ceong, Ki-Hyeong Rhee, and Dokyun Na. In silico prediction of potential chemical reactions mediated by human enzymes. *BMC Bioinformatics*, 19(8):207, 2018.

[16] Johannes Kirchmair, Mark J Williamson, Avid M Afzal, Jonathan D Tyzack, Alison P K Choy, Andrew Howlett, Patrik Rydberg, and Robert C Glen. FAst MEtabolizer (FAME): A Rapid and Accurate Predictor of Sites of Metabolism in Multiple Species by Endogenous Enzymes. *Journal of Chemical Information and Modeling*, 53(11):2896–2907, 11 2013.

[17] Tyler B Hughes, Grover P Miller, and S Joshua Swamidass. Site of Reactivity Models Predict Molecular Reactivity of Diverse Chemicals with Glutathione. *Chemical Research in Toxicology*, 28(4):797–809, 4 2015.

[18] Gabriele Cruciani, Emanuele Carosati, Benoit De Boeck, Kantharaj Ethirajulu, Claire Mackie, Trevor Howe, and Riccardo Vianello. MetaSite: Understanding Metabolism in Human Cytochromes from the Perspective of the Chemist. *Journal of Medicinal Chemistry*, 48(22):6970–6979, 11 2005.

[19] Jenna A Rhoades, Yuri K Peterson, Hao-Jie Zhu, David I Appel, Charles A Peloquin, and John S Markowitz. Prediction and in vitro evaluation of selected protease inhibitor antiviral drugs as inhibitors of carboxylesterase 1: a potential source of drug-drug interactions. *Pharmaceutical research*, 29(4):972–982, 4 2012.

[20] M J Hatfield, L Tsurkan, J L Hyatt, X Yu, C C Edwards, L D Hicks, R M Wadkins, and P M Potter. Biochemical and molecular analysis of carboxylesterase-mediated hydrolysis of cocaine and heroin. *British journal of pharmacology*, 160(8):1916–1928, 8 2010.

[21] Bhawna Vyas, Shalki Choudhary, Pankaj Kumar Singh, Akashdeep Singh, Manjinder Singh, Himanshu Verma, Harpreet Singh, Renu Bahadur, Baldev Singh, and Om Silakari. Molecular dynamics/quantum mechanics guided designing of natural products based prodrugs of Epalrestat. *Journal of Molecular Structure*, 1171:556–563, 2018.

[22] Sunghwan Kim, Paul A Thiessen, Evan E Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A Shoemaker, Jiyao Wang, Bo Yu, Jian Zhang, and Stephen H Bryant. PubChem Substance and Compound databases. *Nucleic acids research*, 44(D1):1202–13, 1 2016.

[23] David S. Wishart, Yannick D. Feunang, An C. Guo, Elvis J. Lo, Ana Marcu, Jason R. Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, Nazanin Assempour, Ithayavani Iynkkaran, Yifeng Liu, Adam MacIejewski, Nicola Gale, Alex Wilson, Lucy Chin, Ryan Cummings, DIana Le, Allison Pon, Craig Knox, and Michael Wilson. DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Research*, 46(D1):D1074–D1082, 2018.

[24] Ida Schomburg, Antje Chang, and Dietmar Schomburg. BRENDA, enzyme data and metabolic information. *Nucleic acids research*, 30(1):47–49, 1 2002.

[25] Edward C. Hulme and Mike A. Trevethick. Ligand binding assays at equilibrium: Validation and interpretation. *British Journal of Pharmacology*, 161(6):1219–1237, 2010.

[26] W. Patrick Walters. Going further than Lipinski's rule in drug design. *Expert Opinion on Drug Discovery*, 7(2):99–107, 2012.

[27] Dean G Brown and Heike J Wobst. A Decade of FDA-Approved Drugs (2010–2019): Trends and Future Directions. *Journal of Medicinal Chemistry*, 64(5):2312–2338, 3 2021.

[28] Stephen R. Johnson. The trouble with QSAR (or how i learned to stop worrying and embrace fallacy). *Journal of Chemical Information and Modeling*, 48(1):25–26, 2008.

[29] T. Scior, J. Medina-Franco, Q.-T. Do, K. Martinez-Mayorga, J. Yunes Rojas, and P. Bernard. How to Recognize and Workaround Pitfalls in QSAR Studies: A Critical Review. *Current Medicinal Chemistry*, 16(32):4297–4313, 2009.

[30] Jed Zaretzki, Matthew Matlock, and S Joshua Swamidass. XenoSite: Accurately Predicting CYP-Mediated Sites of Metabolism with Neural Networks. *Journal of Chemical Information and Modeling*, 53(12):3373–3383, 12 2013.

[31] Sakander Hayat, Shaohui Wang, and Jia Bao Liu. Valency-based topological descriptors of chemical networks and their applications. *Applied Mathematical Modelling*, 60:164–178, 2018.

[32] Milica Markovic, Shimon Ben-Shabat, and Arik Dahan. Computational Simulations to Guide Enzyme-Mediated Prodrug Activation, 2020.

[33] Chenzhong Cao and Li Liu. Topological Steric Effect Index and Its Application. *Journal of Chemical Information and Computer Sciences*, 44(2):678–687, 3 2004.

[34] Ferruccio Palazzesi, Marc A. Grundl, Alexander Pautsch, Alexander Weber, and Christofer S. Tautermann. A Fast Ab Initio Predictor Tool for Covalent Reactivity Estimation of Acrylamides. *Journal of Chemical Information and Modeling*, 59(8):3565–3571, 2019.

[35] Gerrit Schüürmann. QSAR analysis of the acute fish toxicity of organic phosphorothionates using theoretically derived molecular descriptors. *Environmental Toxicology and Chemistry*, 9(4):417–428, 1990.

[36] R K Roy, S Krishnamurti, P Geerlings, and S Pal. Local Softness and Hardness Based Reactivity Descriptors for Predicting Intra- and Intermolecular Reactivity Sequences: Carbonyl Compounds. *The Journal of Physical Chemistry A*, 102(21):3746–3755, 5 1998.

[37] Kai Zhu, Kenneth W Borrelli, Jeremy R Greenwood, Tyler Day, Robert Abel, Ramy S Farid, and Edward Harder. Docking Covalent Inhibitors: A Parameter Free Approach To Pose Prediction and Scoring. *Journal of Chemical Information and Modeling*, 54(7): 1932–1940, 7 2014.

[38] Jessica Vamathevan, Dominic Clark, Paul Czodrowski, Ian Dunham, Edgardo Ferran, George Lee, Bin Li, Anant Madabhushi, Parantu Shah, Michaela Spitzer, and Shanrong Zhao. Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery*, 18(6):463–477, 2019.

[39] Artem Cherkasov, Eugene N Muratov, Denis Fourches, Alexandre Varnek, Igor I Baskin, Mark Cronin, John Dearden, Paola Gramatica, Yvonne C Martin, Roberto Todeschini, Viviana Consonni, Victor E Kuz'min, Richard Cramer, Romualdo Benigni, Chihae Yang, James Rathman, Lothar Terfloth, Johann Gasteiger, Ann Richard, and Alexander Tropsha. QSAR Modeling: Where Have You Been? Where Are You Going To? *Journal of Medicinal Chemistry*, 57(12):4977–5010, 6 2014.

[40] Subhash Ajmani, Kamalakar Jadhav, and Sudhir A Kulkarni. Three-Dimensional QSAR Using the k-Nearest Neighbor Method and Its Interpretation. *Journal of Chemical Information and Modeling*, 46(1):24–31, 1 2006.

[41] Tianqi Chen and Carlos Guestrin. *XGBoost: A Scalable Tree Boosting System*. 8 2016.

[42] Alexander Tropsha. Best Practices for QSAR Model Development, Validation, and Exploitation. *Molecular Informatics*, 29(6-7):476–488, 7 2010.

[43] Ting Zhou and Amedeo Caflisch. High-Throughput Virtual Screening Using Quantum Mechanical Probes: Discovery of Selective Kinase Inhibitors. *ChemMedChem*, 5(7): 1007–1014, 7 2010.

[44] Peter Ertl, Bernhard Rohde, and Paul Selzer. Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties. *Journal of Medicinal Chemistry*, 43(20):3714–3717, 10 2000.

[45] Teruko Imai, Megumi Taketani, Mayumi Shii, Masakiyo Hosokawa, and Kan Chiba. Substrate Specificity of Carboxylesterase Isozymes and Their Contribution to Hydrolase Activity in Human Liver and Small Intestine. *Drug Metabolism and Disposition*, 34(10): 1734 LP – 1741, 10 2006.

[46] Matthew K Ross, Abdolsamad Borazjani, Carol C Edwards, and Philip M Potter. Hydrolytic metabolism of pyrethroids by human and other mammalian carboxylesterases. *Biochemical Pharmacology*, 71(5):657–669, 2006.

[47] Raquel Rodríguez-Pérez and Jürgen Bajorath. Interpretation of Compound Activity Predictions from Complex Machine Learning Models Using Local Approximations and Shapley Values. *Journal of Medicinal Chemistry*, 63(16):8761–8777, 8 2020.

[48] Piotr F J Lipiński and Przemysław Szurmak. SCRAMBLE'N'GAMBLE: a tool for fast and facile generation of random data for statistical evaluation of QSAR models. *Chemical Papers*, 71(11):2217–2232, 2017.

[49] Schrodinger LCC. Maestro Small-Molecular Drug Discovery Suite 2019-4. 2019.

[50] ChemAxon. Marvin (v.20.4.0), 2020.

[51] Noel M O'Boyle, Michael Banck, Craig A James, Chris Morley, Tim Vandermeersch, and Geoffrey R Hutchison. Open Babel. *Journal of Cheminformatics*, 3(33):1–14, 2011.

[52] Marco Biasini, Stefan Bienert, Andrew Waterhouse, Konstantin Arnold, Gabriel Studer, Tobias Schmidt, Florian Kiefer, Tiziano Gallo Cassarino, Martino Bertoni, Lorenza Bordoli, Torsten Schwede, Tiziano Gallo Cassarino, Martino Bertoni, Lorenza Bordoli, and Torsten Schwede. SWISS-MODEL: Modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Research*, 42(Web Server issue):252–8, 7 2014.

[53] Benjamin Webb and Andrej Sali. Comparative Protein Structure Modeling Using MODELLER. *Current Protocols in Bioinformatics*, 54(1):1–5, 6 2016.

[54] G. Madhavi Sastry, Matvey Adzhigirey, Tyler Day, Ramakrishna Annabhimoju, and Woody Sherman. Protein and ligand preparation: Parameters, protocols, and influence on virtual screening enrichments. *Journal of Computer-Aided Molecular Design*, 27(3): 221–234, 2013.

[55] Thomas A. Halgren, Robert B. Murphy, Richard A. Friesner, Hege S. Beard, Leah L. Frye, W. Thomas Pollard, and Jay L. Banks. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. *Journal of Medicinal Chemistry*, 47(7):1750–1759, 2004.

[56] David Ryan Koes, Matthew P. Baumgartner, and Carlos J. Camacho. Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. *Journal of Chemical Information and Modeling*, 53(8):1893–1904, 2013.

[57] Michael M Mysinger, Michael Carchia, John. J Irwin, and Brian K Shoichet. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *Journal of Medicinal Chemistry*, 55(14):6582–6594, 7 2012.

[58] Charly Empereur-Mot, Jean-François Zagury, and Matthieu Montes. Screening Explorer–An Interactive Tool for the Analysis of Screening Results. *Journal of Chemical Information and Modeling*, 56(12):2281–2286, 12 2016.

[59] Gregory Landrum. RDKit: Open-Source Cheminformatics Software, 2021.

[60] J J P Stewart. MOPAC2016, 2016.

[61] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

## Supporting Results and Discussion

### Library

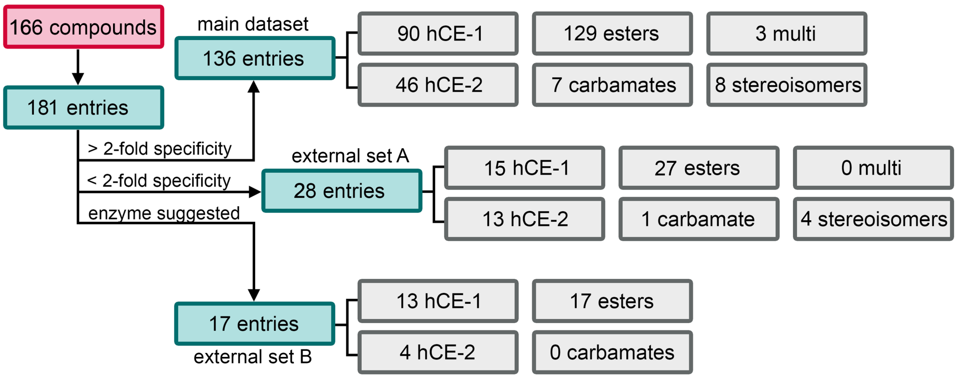


**Figure S 1** Composition of the compound library.

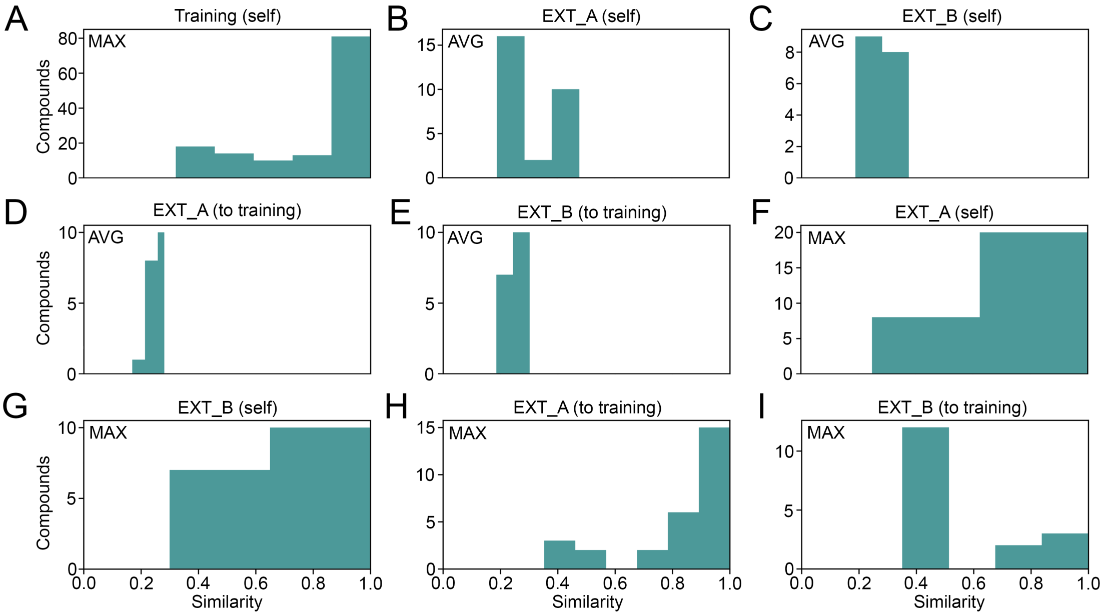


**Figure S 2** Similarity of the external sets to the training set. On the top left of each subfigure, it was indicated if maximal or average similarity was considered. On top of each subfigure, the corresponding compound set is indicated together with ”self” for similarity comparison within the library and ”to training” for comparison to the training set.

## Ligand-based metrics

**Table S 1** Performance of fragment descriptors.

| Set | Accuracy | AUC | MCC | Sensitivity | Specificity |
|---|---|---|---|---|---|
| Training | 0.90±0.05 | 0.94±0.05 | 0.79±0.12 | 0.85±0.17 | 0.93±0.02 |
| External A | 0.76 | 0.77 | 0.51 | 0.88 | 0.62 |
| External B | 0.76 | 0.86 | 0.35 | 0.50 | 0.85 |

Performance of a random forest model based on descriptors of the hydrolysis products (acyl and alcohol moieties). Here, we considered MW, number of atoms, volume, tPSA, and logP.

**Table S 2** Performance of eccentricity descriptors.

| Set | Accuracy | AUC | MCC | Sensitivity | Specificity |
|---|---|---|---|---|---|
| Training | 0.85±0.06 | 0.91±0.09 | 0.68±0.15 | 0.78±0.18 | 0.89±0.07 |
| External A | 0.76 | 0.76 | 0.57 | 1.00 | 0.46 |
| External B | 0.76 | 0.78 | 0.35 | 0.50 | 0.85 |

Performance of a random forest model using the molecular eccentricity of the whole compound, as well as the respective hydrolysis products.

**Table S 3** Performance of steric and QM descriptors.

| Set | Accuracy | AUC | MCC | Sensitivity | Specificity |
|---|---|---|---|---|---|
| Training | 0.82±0.05 | 0.88±0.05 | 0.57±0.13 | 0.65±0.19 | 0.90±0.02 |
| External A | 0.72 | 0.81 | 0.48 | 0.62 | 0.85 |
| External B | 0.71 | 0.68 | 0.25 | 0.50 | 0.77 |

Performance of a random forest model using TSEI and QM parameters. The features included $\text{TSEI}_{\text{carbonyl-O}}$, $\text{Length}_{\text{alcohol}}$, $\text{TSEI}_{\text{alcohol-O}}$, $\text{TSEI}_{\text{carbonyl-C}}$, $\text{Length}_{\text{carbonyl}}$, $\text{Charge}_{\text{carbonyl-C}}$, $E_{\text{LUMO}}$, Hardness, $D_N$, $\text{Charge}_{\text{carbonyl-C}}$, $\text{Charge}_{\text{carbonyl-O}}$, and $\Delta\text{Charge}_{\text{C-O}}$.

## Structure-based metrics

**Table S 4** Quality parameter of the generated homology models.

| Template | QMEAN | GMQE | Ramachandran (%) | MolProbity Score |
|---|---|---|---|---|
| 1MX1 | -2.25 | 0.73 | 92.80 | 1.75 |
| 2HRQ | -2.23 | 0.72 | 91.62 | 1.78 |
| 5A7G | -2.93 | 0.69 | 91.34 | 1.43 |

The selected template structures of hCE-1 are shown together with quality estimates of the resulting models. The estimates include Qualitative Model Energy Analysis (QMEAN), Global Model Quality Estimation (GMQE), percentage of Ramachandran favored conformations, as well as the MolProbity score [2, 3, 4].

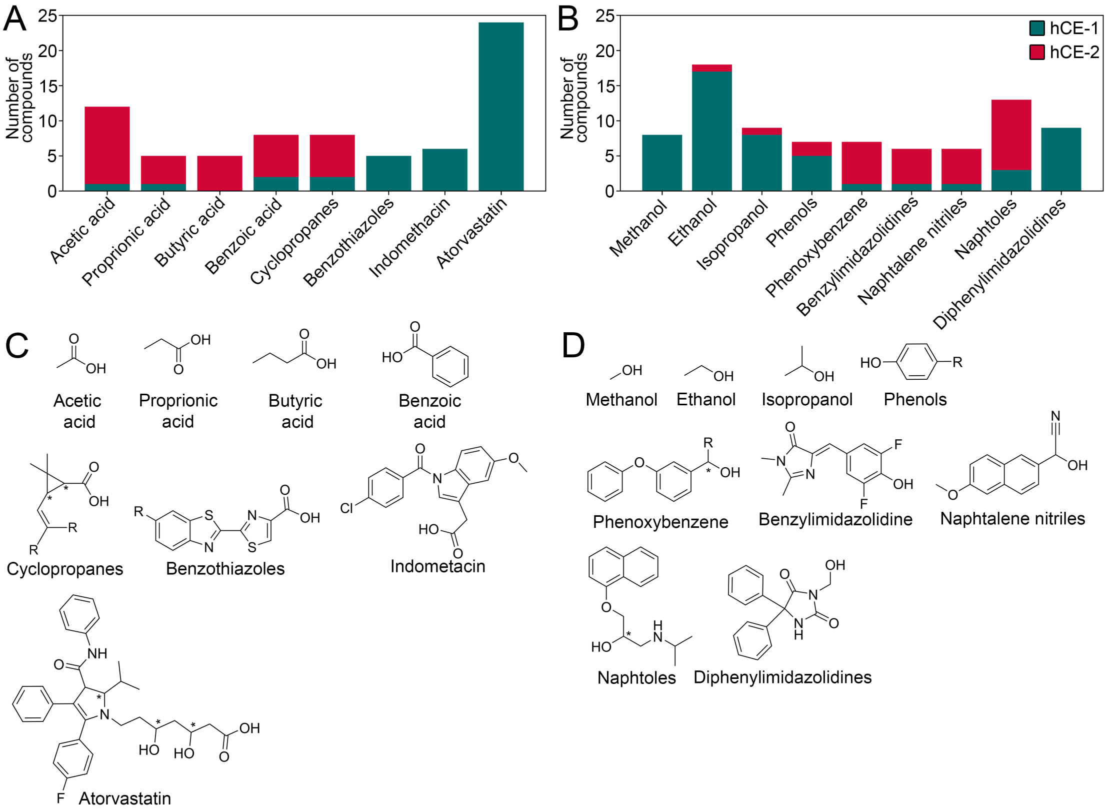


**Figure S 3** Substituent analysis. (A) Distribution of different acyl moieties regarding esterase specificity. (B) Distribution of different alcohol moieties regarding esterase specificity. (C) Structures of acyl moieties. (D) Structures of alcohol moieties. Variable regions are indicated with R-groups.

**Table S 5** Results from decoy docking with smina.

| **Protein** | **Structure** | **ROC AUC** |
|---|---|---|
| hCE-1 | 1YA4 | 0.524 |
|  | 2HRQ | 0.502 |
| hCE-2 | 1MX1 | 0.476 |
|  | 1MX1[a] | 0.704 |
|  | 1MX1[b] | 0.548 |
|  | 2HRQ | 0.407 |
|  | 2HRQ[a] | 0.501 |
|  | 5AG7 | 0.553 |

[a] Representative structure obtained from clustering of an MD simulation. [b] Structure obtained with MODELLER.

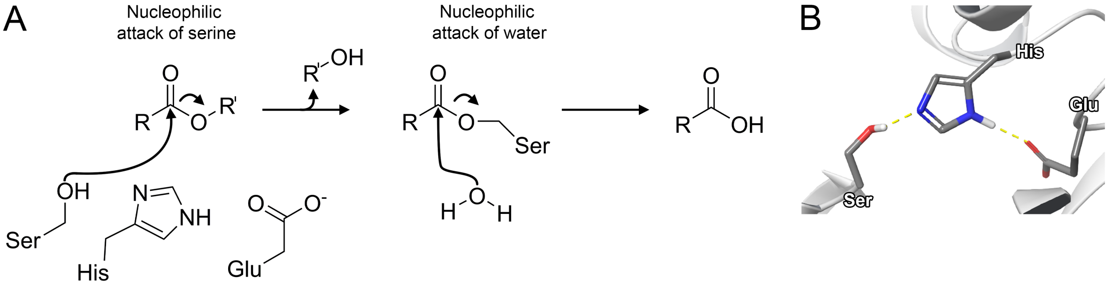
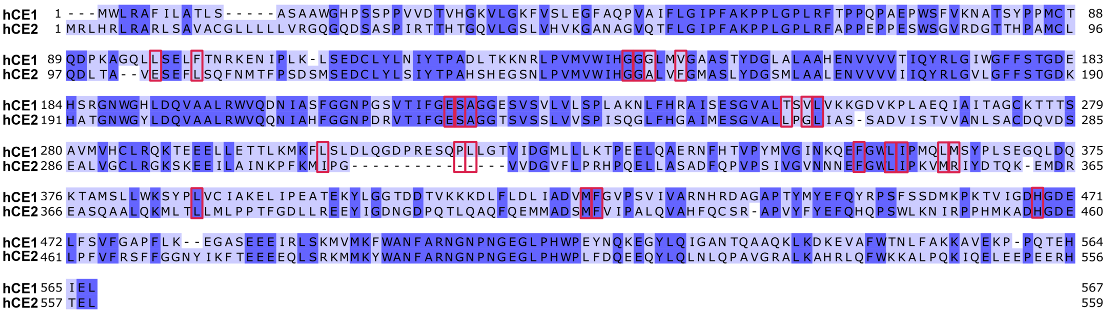


**Figure S 4** Catalytic cycle. (A) Simplified scheme of the catalytic cycle of hCEs according to Hosokawa *et al.*[1] (B) Resting state of the catalytic triad.



```
hCE1   1 ---MWLRAFILATLS-----ASAAWGHPSSPPVVDTVHGKVLGKFVSLEGFAQPVAIFLGIPFAKPPLGPLRFTPPQPAEPWSFVKNATSYPPMCT  88
hCE2   1 MRLHRLRARLSAVACGLLLLLVRGQGQDSASPIRTTHTGQVLGSLVHVKGANAGVQTFLGIPFAKPPLGPLRFAPPEPPESWSGVRDGTTHPAMCL  96

hCE1  89 QDPKAGQLLSELFTNRKENIPLK-LSEDCLYLNIYTPADLTKKNRLPVMVWIHGGGLMVGAASTYDGLALAAHENVVVVTIQYRLGIWGFFSTGDE 183
hCE2  97 QDLTA--VESEFLSQFNMTFPSDSMSEDCLYLSIYTPAHSHEGSNLPVMVWIHGGALVFGMASLYDGSMLAALENVVVVIIQYRLGVLGFFSTGDK 190

hCE1 184 HSRGNWGHLDQVAALRWVQDNIASFGGNPGSVTIFGESAGGESVSVLVLSPLAKNLFHRAISESGVALTSVLVKKGDVKPLAEQIAITAGCKTTTS 279
hCE2 191 HATGNWGYLDQVAALRWVQQNIAHFGGNPDRVTIFGESAGGTSVSSLVVSPISQGLFHGAIMESGVALLPGLIAS-SADVISTVVANLSACDQVDS 285

hCE1 280 AVMVHCLRQKTEEELLETTLKMKFLSLDLQGDPRESQPLLGTVIDGMLLLKTPEELQAERNFHTVPYMVGINKQEFGWLIPMQLMSYPLSEGQLDQ 375
hCE2 286 EALVGCLRGKSKEEILAINKPFKMIPG---------------VVDGVFLPRHPQELLASADFQPVPSIVGVNNNEFGWLIPKVMRIYDTQK-EMDR 365

hCE1 376 KTAMSLLWKSYPLVCIAKELIPEATEKYLGGTDDTVKKKDLFLDLIADVMFGVPSVIVARNHRDAGAPTYMYEFQYRPSFSSDMKPKTVIGDHGDE 471
hCE2 366 EASQAALQKMLTLLMLPPTFGDLLREEYIGDNGDPQTLQAQFQEMMADSMFVIPALQVAHFQCSR-APVYFYEFQHQPSWLKNIRPPHMKADHGDE 460

hCE1 472 LFSVFGAPFLK--EGASEEEIRLSKMVMKFWANFARNGNPNGEGLPHWPEYNQKEGYLQIGANTQAAQKLKDKEVAFWTNLFAKKAVEKP-PQTEH 564
hCE2 461 LPFVFRSFFGGNYIKFTEEEEQLSRKMMKYWANFARNGNPNGEGLPHWPLFDQEEQYLQLNLQPAVGRALKAHRLQFWKKALPQKIQELEEPEERH 556

hCE1 565 IEL                                                                                              567
hCE2 557 TEL                                                                                              559
```

**Figure S 5** Sequence alignment of hCE-1 and hCE-2 with active site residues marked in red.

**Table S 6** Structures selected for docking procedures.

| Protein | smina | Covalent docking |
|---|---|---|
| hCE-1 | 1YA4 | 1MX1, 1MX5, 1MX9, 1YA4, 3K9B |
| hCE-2 | 1MX1[a] | 1MX1[a], 1MX1[b], 5AG7[a] |

[a] Representative structure of MD-evolved homology model. [b] Structure obtained from MODELLER.

**Table S 7** Results from decoy docking with Glide.

| Protein | Structure | ROC AUC |
|---|---|---|
| hCE-1 | 1MX5 | 0.486 |
| | 1MX9 | 0.497 |
| | 1YA4 | 0.486 |
| | 3K9B | 0.485 |
| hCE-2 | 1MX1[a] | 0.469 |
| | 1MX1[b] | 0.465 |
| | 2HRQ[a] | 0.476 |
| | 5AG7 | 0.444 |

[a] Representative structure obtained from clustering of an MD simulation. [b] Structure obtained with MODELLER.

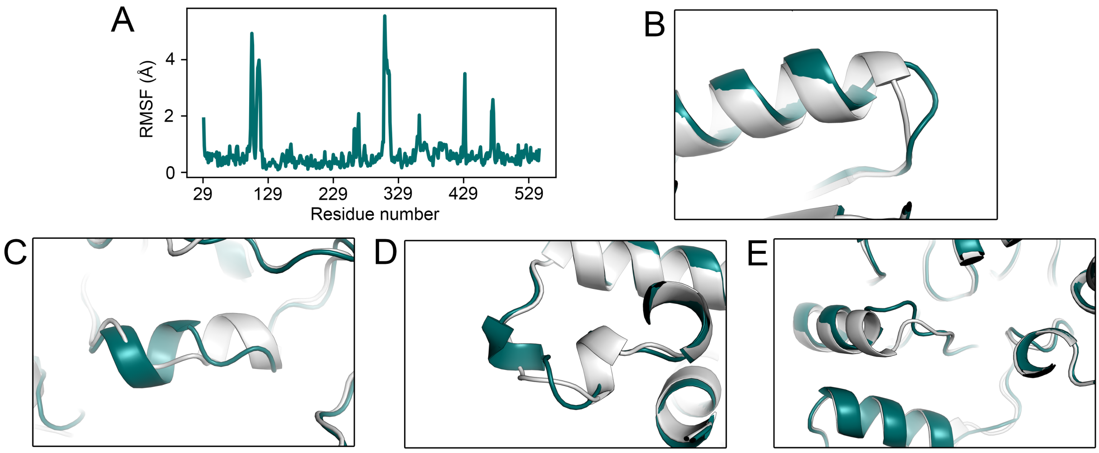


**Figure S 6** Homology model comparison. (A) Backbone RMSF between homology models generated by SWISS-MODEL and Modeller from the same template structure (PDB ID: 1MX1). Structural differences between homology models around (B) residue 429, (C) residue 105, (D) residue 475, and (E) residue 309. The model obtained from SWISS-MODEL is shown in pine green.

**Table S 8** Performance of structure-based descriptors.

| **Set** | **Accuracy** | **AUC** | **MCC** | **Sensitivity** | **Specificity** |
|---|---|---|---|---|---|
| Training | 0.77±0.05 | 0.85±0.05 | 0.47±0.12 | 0.54±0.14 | 0.89±0.6 |
| External A | 0.76 | 0.75 | 0.46 | 0.75 | 0.77 |
| External B | 0.66 | 0.71 | 0.32 | 0.62 | 0.69 |

Performance of a random forest model using structure-based descriptors from docking. The features included scores from conventional and covalent docking, distances (carbonyl carbon to serine oxygen) and GSEI values from conventional docking, and the boolean parameter for the outcome of covalent docking.

## Predictive model

**Table S 9** Metrics of different machine learning classification algorithms during internal validation.

| **Method** | **Accuracy** | **AUC** | **MCC** |
|---|---|---|---|
| Random Forest | 0.92 | 0.97 | 0.82 |
| XGBoost | 0.93 | 0.95 | 0.84 |
| KNN | 0.90 | 0.93 | 0.79 |
| SVM | 0.90 | 0.96 | 0.80 |
| Logistic regression | 0.90 | 0.97 | 0.82 |
| LDA | 0.89 | 0.96 | 0.77 |

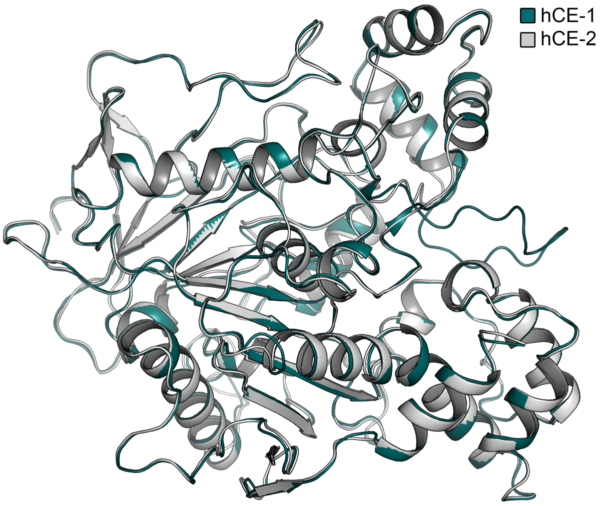


**Figure S 7** Global structure alignment of hCE-2 homology model (Template PDB ID: 1MX1) and template crystal structure.

**Table S 10** Metrics of different machine learning classification algorithms during external validation.

| Method | Accuracy | AUC | MCC |
|---|---|---|---|
| Random Forest | 0.86 | 0.85 | 0.73 |
| XGBoost | 0.79 | 0.81 | 0.56 |
| KNN | 0.82 | 0.91 | 0.67 |
| SVM | 0.86 | 0.92 | 0.73 |
| Logistic regression | 0.89 | 0.85 | 0.79 |
| LDA | 0.71 | 0.67 | 0.41 |

**Table S 13** Sensitivity and specificity of our final model.

| Set | Sensitivity | Specificity |
|---|---|---|
| Training | 0.89±0.14 | 0.93±0.05 |
| External A | 1.00 | 0.67 |
| External B | 0.50 | 0.85 |

**Table S 15** Misclassified compounds in the external set A.

| Compound | ID[a] | Prediction | Explanation | $S_{max}$[b] |
|---|---|---|---|---|
| (R)-Propanolol-heptyl | 147 | hCE-2 | stereospecificity | 0.48 |
| (R)-Propanolol-cyclohexyl | 148 | hCE-2 | stereospecificity | 0.48 |
| Eslicarbazepine acetate | 89 | hCE-2 | acyl:alcohol ratio | 0.35 |
| Fosinopril | 120 | hCE-2 | acyl:alcohol ratio | 0.45 |

[a] Internal compound ID in database. [b] Maximal similarity to training set.

**Table S 11** Features considered features in our final model.

| **Features** | **Description** |
|---|---|
| Score hCE-1 | Scores obtained from conventional docking to hCE-1 |
| Score hCE-2 | Scores obtained from conventional docking to hCE-2 |
| CovScore hCE-1 | Scores obtained from covalent docking to hCE-1 |
| CovDock hCE-1 | Boolean variable describing if pose was found in covalent docking to hCE-1 |
| CovDock hCE-2 | Boolean describing if pose was found in covalent docking to hCE-2 |
| Distance hCE-1 | Minimal distance between serine oxygen and carbonyl carbon |
| Distance hCE-2 | Minimal distance between serine oxygen and carbonyl carbon |
| GSEI hCE-1 | Geometric steric effect index of the carbonyl atom from hCE-1 docking poses |
| GSEI hCE-2 | Geometric steric effect index of the carbonyl atom from hCE-2 docking poses |
| $\mathrm{TSEI_{carbonyl}}$ | Topological steric effect index of the carbonyl carbon atom |
| $\mathrm{TSEI_{acyl}}$ | TSEI of the acyl carbon atom[a] |
| $\mathrm{\#rot_{acyl}}$ | Number of rotatable bonds of acyl moiety |
| $\mathrm{\#rot_{alc}}$ | Number of rotatable bonds of alcohol moiety |
| $\mathrm{Ratio_{MW}}$ | Ratio between $\mathrm{MW_{acyl}}$ and $\mathrm{MW_{alc}}$ |
| $\mathrm{Volume_{acyl}}$ | Volume of the acyl moiety |
| $\mathrm{Volume_{alc}}$ | Volume of the alcohol moiety |
| $\mathrm{Ratio_{Volume}}$ | Ratio between $\mathrm{Volume_{acyl}}$ and $\mathrm{Volume_{alc}}$ |
| $\mathrm{Ratio_{tPSA}}$ | Ratio between $\mathrm{tPSA_{acyl}}$ and $\mathrm{tPSA_{alc}}$ |
| $\mathrm{logP_{acyl}}$ | logP of acyl moiety |
| $\mathrm{logP_{alc}}$ | logP of alcohol moiety |
| $\mathrm{Ratio_{logP}}$ | Ratio between $\mathrm{logP_{acyl}}$ and $\mathrm{logP_{alc}}$ |
| Eccentricity | Topological eccentricity of whole compound |
| $\mathrm{E_{LUMO}}$ | LUMO energy of the substrate |
| $\mathrm{Bond_{C\text{-}O}}$ | Bond length of the alcohol bond |
| $\mathrm{Bond_{C=O}}$ | Bond length of the carbonyl bond |
| Hardness | Hardness of the substrate |
| $\mathrm{Charge_{Carbon}}$ | Net atomic charge of carbonyl carbon atom |
| $\Delta$Charge | Charge difference between $\mathrm{Charge_{Carbon}}$ and $\mathrm{Charge_{Oxygen}}$ |

[a] Nitrogen atom if carbamate.

**Table S 12** Remaining features considered during feature selection.

| Features | Description |
|---|---|
| $\text{TSEI}_{\text{alcohol}}$ | TSEI of the oxygen of the alcohol moiety |
| $\text{TSEI}_{\text{carbonyl-ox}}$ | TSEI of the carbonyl oxygen atom |
| CovScore hCE-2 | Score obtain from covalent docking to hCE-2 |
| MW | Molecular weight of whole compound |
| $\text{MW}_{\text{acyl}}$ | Molecular weight of acyl moiety |
| $\text{MW}_{\text{alc}}$ | Molecular weight of alcohol moiety |
| tPSA | Topological polar surface area of whole compound |
| $\text{tPSA}_{\text{acyl}}$ | tPSA of acyl moiety |
| $\text{tPSA}_{\text{alc}}$ | tPSA of alcohol moiety |
| logP | logP of whole compound |
| Volume | Volume of whole compound |
| LabutASA | Labute's Approximate Surface Area of whole compound |
| $\text{LabutASA}_{\text{acyl}}$ | Labute's Approximate Surface Area of acyl moiety |
| $\text{LabutASA}_{\text{alc}}$ | Labute's Approximate Surface Area of alcohol moiety |
| #Atoms | Number of atoms of whole compound |
| $\text{\#Atoms}_{\text{acyl}}$ | Number of atoms of acyl moiety |
| $\text{\#Atoms}_{\text{alc}}$ | Number of atoms of alcohol moiety |
| $\text{Ratio}_{\text{\#Atoms}}$ | Ratio between $\text{\#Atoms}_{\text{acyl}}$ and $\text{\#Atoms}_{\text{alc}}$ |
| $\text{Eccentricity}_{\text{acyl}}$ | Topological eccentricity of acyl moiety |
| $\text{Eccentricity}_{\text{alc}}$ | Topological eccentricity of alcohol moiety |
| $D_N$ | Nucleophilic delocalizability |
| $\text{Charge}_{\text{Oxygen}}$ | Net atomic charge of carbonyl oxygen atom |

**Table S 14** Misclassified compounds in the training set.

| Compound | ID[a] | Prediction | Explanation |
|---|---|---|---|
| 2-(1,3-Benzothiazol-2-yl)-6-methoxyphenyl benzoate | 83 | hCE-2 | acyl:alcohol ratio |
| (1RS)-cis-Bifenthrin | 141 | hCE-1 | tPSA contradicts volume, similarly sized acyl:alcohol |
| Bioresmethrin | 30 | hCE-2 | tPSA contradicts volume, similarly sized acyl:alcohol |
| (S)-Permethrin | 32 | hCE-1 | stereospecificity |
| 4-(4Z)-1,2-dimethyl-5-oxo-4,5-dihydro-1H-imidazol-4-ylidene methyl-2,6-difluorophenyl 1,1'-biphenyl-4-carboxylate | 156 | hCE-2 | similarly sized acyl:alcohol |
| Dabigatran etexilate | 5 | hCE-1 | acyl:alcohol ratio |
| (R)-Permethrin | 31 | hCE-2 | stereospecificity |
| Procaine | 43 | hCE-1 | similarly sized acyl:alcohol |
| Flupirtine | 117 | hCE-1 | acyl:alcohol ratio |
| (S,S)-A4 | 25 | hCE-2 | stereospecificity |
| Isovaleryl-(R)-Propranolol | 128 | hCE-2 | stereospecificity |

[a] Internal compound ID in database.

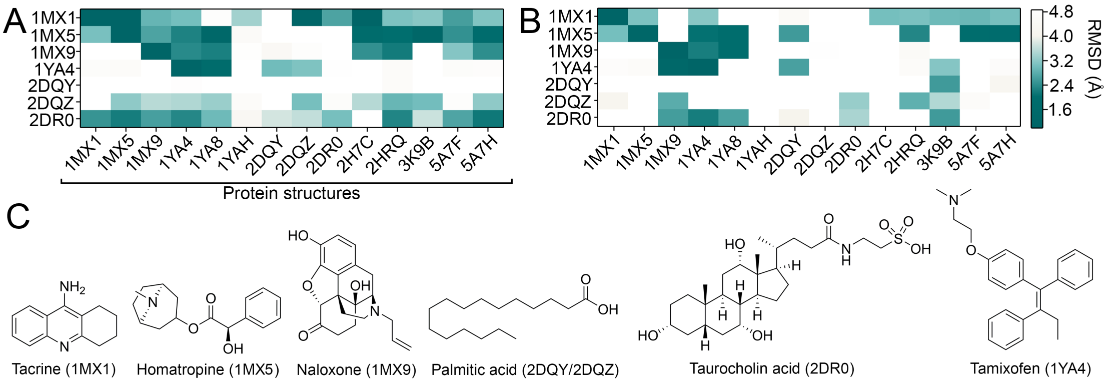


**Figure S 8** Cross docking evaluation. RMSD values of cross docking 7 cocrystallized ligands (vertical column) to 14 crystal structures for (A) smina and (B) Glide SP. (C) Structures of the respective cocrystallized ligands highlighting their structural diversity.

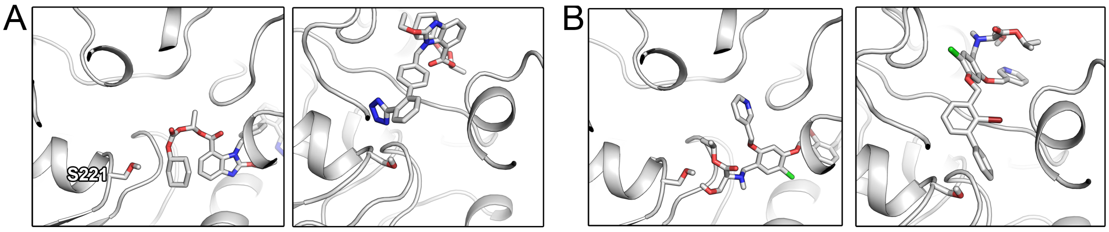


**Figure S 9** Binding modes and spatial relationship between ester and catalytic serine of hCE-1 substrates. (A) Binding modes of candesartan cilexetil in hCE-1 (left) and hCE-2 (right). (B) Binding modes of IMMH-010 in hCE-1 (left) and hCE-2 (right).

# Supporting Materials and Methods

## Library generation

## Homology modeling and molecular docking

The obtained homology models were subjected to MD simulations followed by the determination of representative structures to improve performance of hCE-2 structure-based considerations. The MD simulations were conducted using the Desmond (v2019-1) simulation engine [5] with the OPLS_2005 force field in an NPT ensemble at a temperature of 310 K maintained by the Nose–Hoover thermostat and atmospheric pressure regulated by the Martyna–Tobias–Klein barostat, both with a relaxation time of 2.0 ps. The structures were placed in orthorhombic periodic boundary systems solvated with TIP3P water molecules with counter-ions neutralizing the systems. Short-range interactions were cut off at 9 Å and long-range interactions were treated with the u-series algorithm [6]. The M-SHAKE algorithm was used to constrain bonds to hydrogen atoms and the time step of the RESPA integrator was set to 2.0 fs. The simulations were terminated after 25 ns and snapshots with atomic coordinates were collected at an

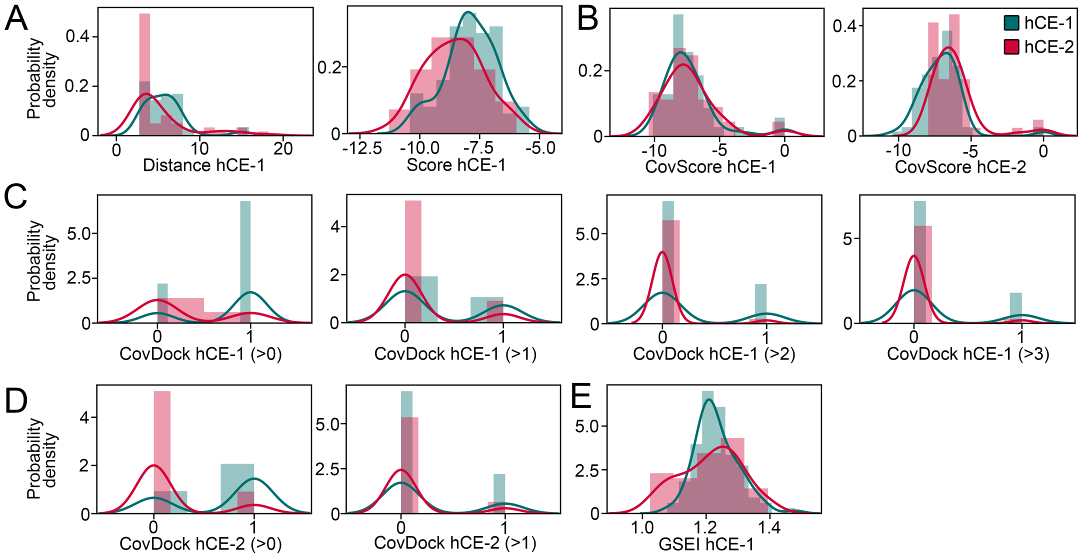


**Figure S 10** Additional structure-based features. (A) Readouts from conventional docking to hCE-1. (B) Scores obtained from covalent docking. Different thresholds for number of unsuccessful covalent docking attempts for (C) hCE-1 and (D) hCE-2. (E) GSEI of docking poses bound to hCE-1.

interval of 25 ps. The simulations were conducted for models from SWISS-MODEL (Template PDB IDs: 1MX1, 2H7C, 2HRQ, 5AG7) and MODELLER (Template PDB ID: 1MX1). The RMSDs of the simulations indicated acceptable convergence (Figure S14). Clustering of the last 200 frames of the trajectories to obtain representative structures was done using the `trj_cluster.py` script that comes with Maestro. The number of output clusters was limited to five per simulation and we retained the most populated ones. The number of frames represented by the selected structures amounted to 11, 11, 16, 11, and 13 for 1MX1 (SWISS-MODEL) 1MX1 (MODELLER), 2H7C, 2HRQ, and 5AG7, respectively.

**Table S 16** Statistics of compounds for decoy docking.

| **Protein** | **Actives** | **Decoys** | **Ratio** |
|---|---|---|---|
| hCE-1 | 235 | 14330 | 71.0 |
| hCE-2 | 55 | 3110 | 56.5 |

The number of actives and decoys is given together with their ratio.

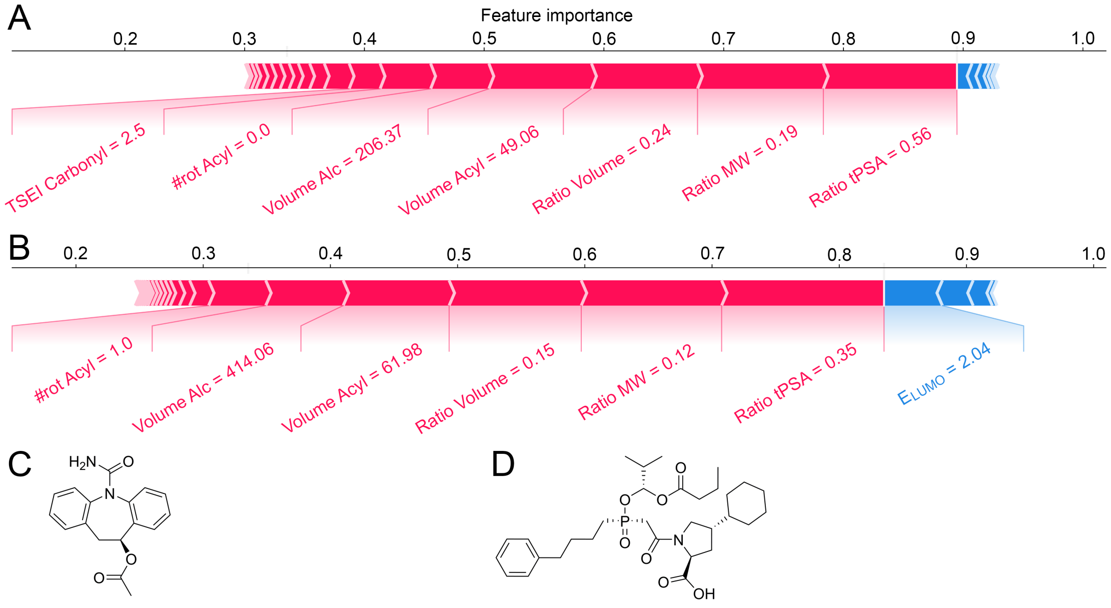


**Figure S 11** Outlier analysis. (A) SHAP analysis of eslicarbazepine acetate. (B) SHAP analysis of fosinopril. (C) Structure of elsicarbazepine acetate. (D) Structure of fosinopril.

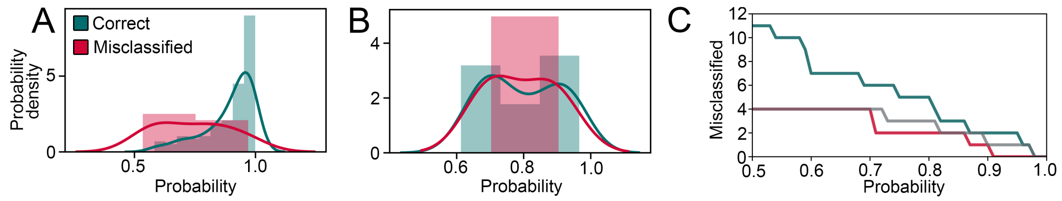


**Figure S 12** Prediction probabilities. Distribution of the prediction probabilities between correct predictions and misclassified compounds for (A) the training set and (B) the external set A. (C) Number of misclassified compounds by varying the probability threshold (training set in pine green, external set A in red, external set B in gray).

## Ligand-based metrics

**Table S 17** RDKit commands used during metrics computation.

| Command | Description |
|---|---|
| GetFormalCharge | Computation of absolute charge of a compound |
| GetSubstructMatch | Find substructure in a compound |
| MolFromSmarts | Define a SMARTS pattern |
| FragmentOnBonds | Fragmentation of a bond within a compound |

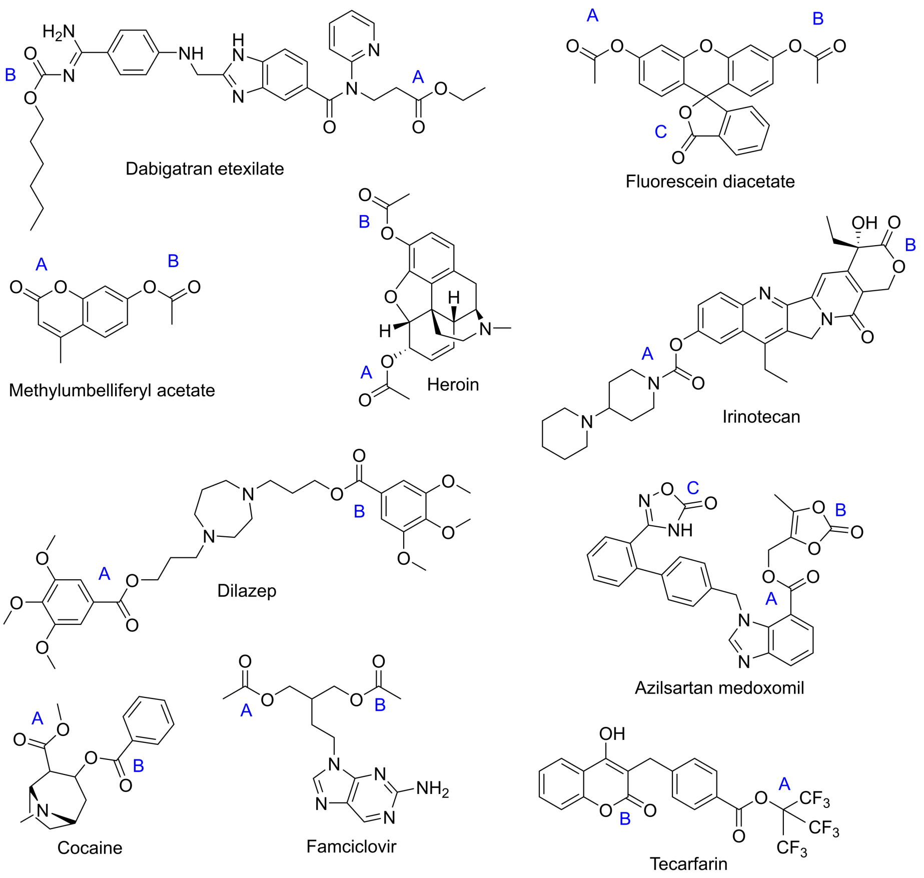


**Figure S 13** Multiple ester or carbamate groups in substrates. The assignment of ester groups for our computations is given.

## Machine learning

**Table S 18** Classification algorithms.

| **Method** | **Classifier** |
|---|---|
| Random forest (RF) | RandomForestClassifier |
| XGBoost | XGBClassifier |
| Support vector machine (SVM) | SVC |
| Linear discriminant analysis (LDA) | LinearDiscriminantAnalysis |
| k-Nearest neighbors (kNN) | KNeighborsClassifier |
| Logistic regression | LogisticRegression |

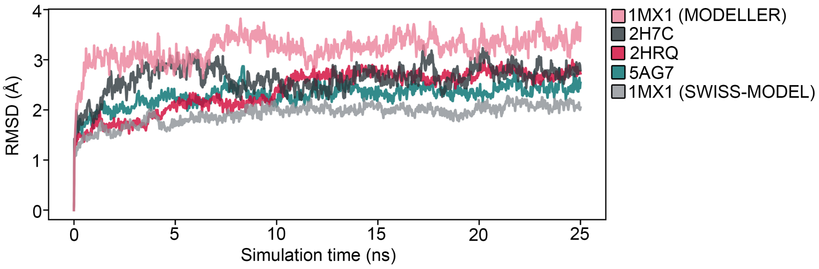


**Figure S 14** RMSDs of MD simulations in this study.

**Table S 19** Hyperparameters selected for machine learning.

| RF | XGBoost | SVM | kNN |
|---|---|---|---|
| bootstrap: true | learning_rate: 0.3 | kernel: rbf | K = 5 |
| max_depth: 50 | colsample_bytree: 0.6 | C: 1 | |
| max_features: sqrt | eval_metric: logloss | | |
| min_samples_leaf: 4 | gamma: 0 | | |
| min_samples_split: 5 | max_depth: 3 | | |
| n_estimators: 200 | min_child_weight: 5 | | |
| | objective: binary:logistic: 0 | | |
| | subsample: 1.0 | | |
| | nthread: 1 | | |

## References

[1] Masakiyo Hosokawa. Structure and catalytic properties of carboxylesterase isozymes involved in metabolic activation of prodrugs. *Molecules (Basel, Switzerland)*, 13(2):412–431, 2 2008.

[2] Gabriel Studer, Christine Rempfer, Andrew M Waterhouse, Rafal Gumienny, Juergen Haas, and Torsten Schwede. QMEANDisCo—distance constraints applied on model quality estimation. *Bioinformatics*, 36(6):1765–1771, 3 2020.

[3] Marco Biasini, Stefan Bienert, Andrew Waterhouse, Konstantin Arnold, Gabriel Studer, Tobias Schmidt, Florian Kiefer, Tiziano Gallo Cassarino, Martino Bertoni, Lorenza Bordoli, Torsten Schwede, Tiziano Gallo Cassarino, Martino Bertoni, Lorenza Bordoli, and Torsten Schwede. SWISS-MODEL: Modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Research*, 42(Web Server issue):252–8, 7 2014.

[4] Vincent B Chen, W Bryan 3rd Arendall, Jeffrey J Headd, Daniel A Keedy, Robert M Immormino, Gary J Kapral, Laura W Murray, Jane S Richardson, and David C Richardson. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta crystallographica. Section D, Biological crystallography*, 66(Pt 1):12–21, 1 2010.

[5] Kevin Bowers, Edmond Chow, Huafeng Xu, Ron Dror, Michael Eastwood, Brent Gregersen, John Klepeis, Istvan Kolossvary, Mark Moraes, Federico Sacerdoti, John Salmon, Yibing Shan, and David Shaw. Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters. *ACM/IEEE SC 2006 Conference (SC'06)*, (November): 43, 2006.

[6] David E. Shaw, J. P. Grossman, Joseph A. Bank, Brannon Batson, J. Adam Butts, Jack C. Chao, Martin M. Deneroff, Ron O. Dror, Amos Even, Christopher H. Fenton, Anthony Forte, Joseph Gagliardo, Gennette Gill, Brian Greskamp, C. Richard Ho, Douglas J. Ierardi, Lev Iserovich, Jeffrey S. Kuskin, Richard H. Larson, Timothy Layman, Li Siang Lee, Adam K. Lerer, Chester Li, Daniel Killebrew, Kenneth M. Mackenzie, Shark Yeuk Hai Mok, Mark A. Moraes, Rolf Mueller, Lawrence J. Nociolo, Jon L. Peticolas, Terry Quan, Daniel Ramot, John K. Salmon, Daniele P. Scarpazza, U. Ben Schafer, Naseer Siddique, Christopher W. Snyder, Jochen Spengler, Ping Tak Peter Tang, Michael Theobald, Horia Toma, Brian Towles, Benjamin Vitale, Stanley C. Wang, and Cliff Young. Anton 2: Raising the Bar for Performance and Programmability in a Special-Purpose Molecular Dynamics Supercomputer. *International Conference for High Performance Computing, Networking, Storage and Analysis, SC*, 2015-Janua(January):41–53, 2014.

# CHAPTER 5

# Review: Ligand Pathways in Nuclear Receptors

Besides drug-metabolizing enzymes, NRs were the second central protein family addressed in this thesis. Motivated by the work on ligand tunnels in CYPs and several reviews written on the topic, this review aimed at summarizing findings on ligand pathways in NRs in the literature. As discussed, the uptake of ligands to buried binding pockets constitutes a key component of their molecular recognition process. This work also served as a foundation for the study on ligand pathways in estrogen-related receptors presented in Chapter 6.

---

**Author contributions:** Conceptualization, A.F.; formal analysis, A.F.; writing and original draft preparation, A.F.; writing, review and editing, A.F., M.S.; visualization, A.F.; supervision, M.A., M.S.

---

*Based on published review article:*

Fischer, A.; Smieško, M. Ligand Pathways in Nuclear Receptors. *J. Chem. Inf. Model.* 2019, 59, 3100–3109.

## Abstract

Nuclear receptors (NRs) are ligand-inducible transcription factors that play an essential role in a multitude of physiological processes as well as diseases rendering them attractive drug targets. Crystal structures revealed the binding site of NRs to be buried in the core of the protein with no obvious route for ligands to access this cavity. The process of ligand binding is known to be an often-neglected contribution to the efficacy of drug candidates and is thought to influence the selectivity and specificity of NRs. While experimental methods generally fail to highlight the dynamic processes of ligand access or egress on the atomistic scale, computational methods have provided fundamental insight into the pathways connecting the buried binding pocket to the surrounding environment. Methods based on molecular dynamics (MD) and Monte Carlo simulations have been applied to identify pathways and quantify their capability to transport ligands. Here, we systematically review findings of more than 20 years of research in the field including the applied methodology and controversies. Further, we establish a unified nomenclature to describe the pathways in respect to their location relative to protein secondary structure elements and summarize findings relevant drug design. Lastly, we discuss the effect of NR interaction partners such as coactivators and corepressors, as well as mutations on the pathways.

## Introduction

Nuclear receptors (NRs) are ligand-inducible transcription factors that translocate to the nucleus and directly regulate gene transcription. Due to their involvement in important physiological processes such as cell proliferation, development, immunity, metabolism, and reproduction, they are of major interest to the field of life sciences [1, 2]. Naturally, some of them are involved in diseases such as cancer and diabetes rendering them attractive drug targets. Since hundreds of crystal structures of NRs have been deposited in the Protein Data Bank until today, the binding mode of a multitude of ligands and related structural implications on the receptors could be investigated to ultimately develop or optimize drug molecules [3, 4, 5]. NRs share a common structural architecture consisting of three main domains including the highly variable N-terminal domain (NTD), the relatively conserved DNA-binding domain (DBD), and the ligand-binding domain (LBD) as shown in Figure 1A. Ligand binding to the LBD induces conformational changes in the receptors that, in most cases, lead to the dissociation of auxiliary proteins such as corepressors and allow for the association of coactivator proteins at the so-called activation function 2 (AF-2) located on the surface of the LBD. In the next step, most NRs form either homo- or heterodimers and the newly formed complex then directly regulates gene transcription in the nucleus [2, 6]. Most of the currently available therapeutics exploit the hormone binding site of the LBD, which is buried in the core of the receptor (Figure 1B). Since crystal structures provide no obvious path for ligands to enter or leave this cavity, it is accepted that pathways must connect it to the surrounding solvent environment [7, 8, 9, 10]. Obviously, dynamic protein motions need to occur in order for these pathways to open and transport ligands to their respective binding site. Current experimental methods have only limited applicability to qualitatively and quantitatively detect such ligand pathways on an atomic level and there is no established laboratory method to investigate these pathways [8, 11, 12, 13]. On the other hand, computational methods such as molecular dynamics (MD) simulations and related techniques can give a detailed and atomistic model on the protein motions responsible for transporting ligands through the pathways and the resulting ligand-protein interactions [8, 14, 15]. In this regard, it is of major importance to understand the atomic mechanism of ligand binding to rationally develop and optimize therapeutics [7, 9, 11, 16, 17, 18]. For example, detailed knowledge on the ligand binding mechanism and its kinetics can be used to modify the residence time of drugs by optimizing off-rates [7, 16, 19, 20, 21]. Commonly used docking methodologies could be supplemented, since they neglect the access to buried binding pockets, which might constitute a high energy barrier for the ligand [12, 22]. This matter was discussed in a recent review on the interplay of docking and MD simulations [23]. Further, binding pathways are thought to impact the ligand specificity of NRs since differences in single amino acids in the binding pocket often fail to deliver a complete picture of ligand preference [24, 25]. This is also supported by the relatively conserved helical architecture of different NRs as shown in Figure 1C [26]. The main structural differences among NRs are located in the proximity of the ligand which includes the region where H3, H7, and H11 meet as well as the vicinity of H2 (cf. Figure 1C) [7]. For these reasons, over

26 studies focused on the topic of ligand binding pathways in 15 different NRs by employing various computational methods. The statistics on the most intensively studied receptors are displayed in Figure 1D.

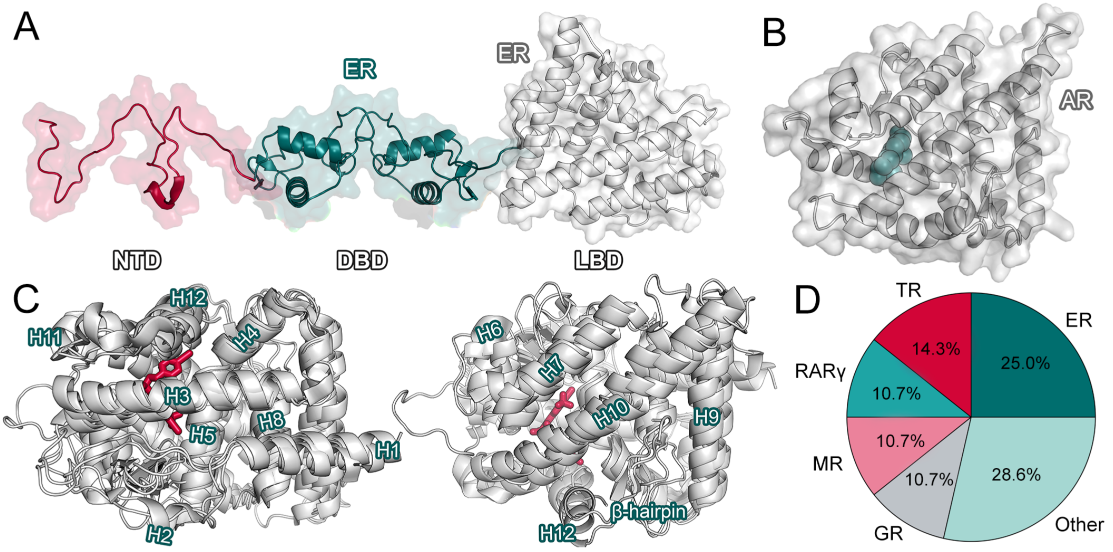


**Figure 1** Structure of nuclear receptors and study focus. (A) The three main domains of NRs at the example of the estrogen receptor (PDB IDs: 3OS8 and 1HCQ). Since the structure for the disordered [27] NTD of NRs was not yet determined, a similarly disordered structure of the protein At2g23090 (PDB ID: 1WVK) was used for a schematic visualization. (B) The buried binding pocket of the androgen receptor (AR) (PDB ID: 2PIV) is shown with the ligand colored in pine green. (C) Structural alignment of NRs and secondary structure elements. Two perspectives on the alignment of AR, glucocorticoid receptor (GR), retinoic acid receptor $\gamma$ (RAR$\gamma$), thyroid receptor $\beta$ (TR$\beta$) (PDB IDs: 3RLJ, 5NFP, 2LBD, 1NAX) are shown. The alignment was performed in PyMol [28]. Secondary structure assignments were made based on the structure of the AR with the exception of H2 that does not occur in the AR [29]. (D) Pie chart showing the main receptors for which ligand pathways have been determined.

Several review papers have highlighted computational methods to study NRs in general [30], the architecture of NRs including the mousetrap mechanism [26, 31], the in silico evaluation of NR binders [32], as well as the methods that can be universally applied to explore biological transport events [33]. In contrast to previous publications, we provide an in-depth overview on two decades of research in the specific field of ligand pathways in NRs. We first introduce the topic, then describe and compare all computational methods that were used in the field, and emphasize general experimental insights into the pathways. We then introduce a nomenclature for the pathways to standardize the heterogenous terminology and systematically summarize their spatial location within the receptors. Lastly, we discuss the results and conclusions of various studies regarding drug design as well as the influence of protein interaction partners and mutations.

## Applied Methodology

**Computational Methods.** Several computational methods were applied to study ligand pathways within NRs [9, 12, 19, 25, 34, 35], as well as other proteins such as Cytochrome P450 enzymes (CYPs) [36, 37]. In the majority of cases, the used techniques were related to MD simulations which allow to determine the spatial location of pathways and their capability to transport ligands. Due to the relatively slow timescale (microseconds to minutes) of small molecule binding and unbinding from NRs exceeding the available computational power of conventional simulations, most studies employed some form of accelerated simulation protocol involving the addition of biasing potentials to the regular force field [9, 15, 35]. Kosztin and colleagues pioneered the field by applying steered MD (SMD) simulations to study the unbinding pathways of retinoic acid from the RAR$\gamma$ in 1999 [9]. In SMD simulations an external force is applied to a molecule in the form of a directional vector that is attached via a harmonic spring [38, 39]. In the case of ligand pathways, small molecules can be pulled out of the receptor through a predefined region, which is at the same time one of its main disadvantages compared to other methods [11] that do not depend on the definition of the spatial region prior to simulation. Based on the force profile or the maximal force, different pathways or compounds can be ranked and it was suggested that SMD simulations could supplement conventional molecular docking methods to prioritize potential drug compounds, especially in the case of equally well-scored ligands [11, 39, 40].
Interestingly, a modified SMD protocol was applied to study the access of multiple compounds to the TRs [41]. In the same year as Kosztin and colleagues, Blondel *et al.* applied locally enhanced sampling MD (LES-MD) simulations to the same model system of RAR$\gamma$. In LES-MD, the protein is subjected to multiple high-temperature ligand copies to increase the probability of observing rare molecular events such as intramolecular ligand diffusion at a low computational expense [15, 34]. Compared to SMD, a clear advantage of the LES-MD methodology is that there is no need to predefine a path for the ligand [34], as it is also the case for random accelerated MD (RAMD) simulations. Initially introduced as random expulsion MD (REMD) simulations by Lüdemann

and colleagues [42], the RAMD technique allows the determination of ligand pathways while minimizing the introduced bias. Similar to SMD, this algorithm is based on the addition of an artificial force to the default force field, but the direction of the force is randomly adapted and it is only applied if the molecule does not cover a given distance in a predefined number of simulation steps. It was suggested to use RAMD simulations to determine the relevant pathways in a system, followed by SMD simulations to precisely determine their capability to transport ligands by comparing the potential of mean force (PMF) [16, 19, 42]. Disadvantages of the RAMD methodology include the unnatural deformation of the protein in short simulations as well as the dependence on the atom to which the additional force is applied [11, 19]. Similar to RAMD and LES-MD, the targeted MD (TMD) methodology is not dependent on the prior knowledge of potential ligand pathways [12].

**Table 1** Hallmarks of all 26 articles considered for this review.

| *Receptor* | *Focus* | *Method* | *Ligand* | *Year* |
|---|---|---|---|---|
| RAR$\gamma$ | Egress | SMD | all-trans retinoic acid | 1999 [9] |
| RAR$\gamma$ | Egress | LES-MD | all-trans retinoic acid | 1999 [34] |
| TR$\alpha$, TR$\beta$ | Egress | LES-MD | triiodothyronine, tiratricol, IH-5[a], GC-24[a] | 2005 [15] |
| TR$\alpha$, TR$\beta$ | Egress | SMD | triiodothyronine, tiratricol, IH-5[a], GC-24[a] | 2006 [14] |
| RAR$\gamma$ | Egress | RAMD | all-trans retinoic acid | 2006 [19] |
| AR | Egress | SMD | ethylated cyanonilutamide | 2007 [43] |
| ER$\alpha$ | Egress | LES-MD | 17$\beta$-estradiol, raloxifene | 2008 [20] |
| TR$\alpha$, TR$\beta$ | Access | SMD[b] | triiodothyronine, tiratricol, IH-5[a], GC-24[a] | 2008 [41] |
| PPAR$\gamma$ | Egress | rTMD, TDR | GW0072[c] | 2008 [12] |
| VDR | Egress | RAMD, SMD, TMD | calcitriol | 2009 [11] |
| ER$\alpha$, ER$\beta$ | Egress | RAMD, SMD, CAVER | 17$\beta$-estradiol, genistein, 4-hydroxytamoxifen | 2009 [25] |
| ER$\alpha$, ER$\beta$ | Egress | SMD | genistein, Way-244[c] | 2009 [17] |
| PPAR$\gamma$ | Access | TMD | GW0072[c] | 2011 [22] |
| FXR | Egress | RAMD, SMD | GW4064[c] | 2012 [18] |
| MR | Acess | PELE | LD1[a] | 2013 [35] |
| TR$\alpha$, TR$\beta$ | Egress | RMSD | triiodothyronine | 2013 [13] |
| GR | Egress | SMD | dexamethasone, fluticasone furoate, fluticasone propionate | 2013 [40] |
| ER$\alpha$ | Egress | SMD, CAVER | salpichrolide analog | 2015 [44] |
| RXR$\alpha$ | Access | Docking, cMD | 9-cis retinoic acid | 2015 [45] |
| ER$\alpha$, ER$\beta$ | - | CAVER | - | 2015 [46] |
| GR, MR | Access, Egress | PELE | dexamethasone, desisobutyrylciclesonide | 2015 [7] |
| ER$\beta$ | Access | PELE | 272[a], 797[a] | 2016 [47] |
| RXR$\alpha$ | Egress | SMD, CAVER | bexarotene, GVD[c], 3gxl[a] | 2017 [39] |
| AR, ER$\alpha$ , GR, MR, PR | Access | PELE | testosterone, estradiol, cortisol, aldosterone, progesterone | 2017[10] |
| EcR | Egress | SMD, CAVER | HWG[a] | 2018 [16] |
| PXR | Access | MD-binding | SRL[a] | 2018 [8] |

[a] Compound name according to Protein Data Bank [48]. [b] Modified simulation protocol.

[c] Compound name according to PubChem [49].

TMD simulations can be classified into direct TMD and reversed TMD (rTMD). In direct TMD simulations, the root mean square deviation (RMSD) between an initial structure and a target structure is decreased step by step, while rTMD increases the RMSD from an initial reference structure. Therefore, the rTMD protocol exerts less constraints on the atoms since no target structure for the search is given [12, 22, 33]. Similarly, the time-dependent distance-restraint (TDR) method increases the distance between the centers of a ligand and its respective binding site from an initial structure to observe an egress event [12]. The MD-binding method uses an additive bias based on electrostatic-like forces to enhance the probability of the ligand passing through an access pathway. This method was recently applied to study access pathways in the PXR [8, 50]. While previously discussed methods can be classified as biased simulation techniques, the CAVER method detects pathways from either static crystal structures or ensembles from conventional MD simulations [51]. In CAVER, pathways are determined from a predefined starting point, usually within the presumed binding site, from which the algorithm detects the cheapest paths towards the protein surface based on a cost function accounting for diameter and length. To perform the calculation, the protein atoms are approximated by a Voronoi diagram. In a single study, a ligand was docked into the entrance of the presumed binding pathway and a subsequent conventional MD simulation lead to the spontaneous binding of a ligand to RXR$\alpha$ [45]. Only recently, the protein energy landscape exploration (PELE) method was introduced and applied to study ligand pathways in NRs [35, 47]. In contrast to the above-mentioned MD protocols, PELE relies on a combination of protein structure prediction and Monte Carlo sampling allowing to study access pathways in an unbiased matter. Such unbiased simulations with NR ligands allowed to reproducibly determine the ligand binding site of several NR LBDs without prior knowledge on its location. Additionally, relative binding affinities showing a good correlation to experimentally determined values could be obtained from the simulations [7]. Together with its low computational cost, this has made PELE an attractive approach to study ligand access pathways. As a main disadvantage, the PELE method is limited by its ineffectiveness to capture larger, global changes in the protein secondary structure [47]. The most prominent computational methods applied to study pathways in NRs are shown in Figure 2A.

**Experimental Methods.** As stated above, experimental methods have only a limited applicability to study access or egress pathways in NRs and other proteins, especially if atomic detail is desired [8, 11, 12, 13]. However, crystal structures that inherently build the foundation for structure-based computational methods have also provided insights into ligand pathways despite the static nature of the structural information [52]. For example, they revealed the presence of peripheral sites on the surface of the FXR [53] and the MR [54] that could be connected to an access pathway (Figure 2B and C) as discussed in detail later. Further, structural differences between NRs in distinct regions, could contribute to the understanding of receptor specificity that is closely related to entry and exit pathways (cf. Figure 2D and E). On the other hand, tryptophan fluorescence studies can provide insight into the superficial accessibility of receptor regions and have suggested major conformational changes associated with ligand binding [55]. The role of specific residues involved in the access or egress process can be probed with site-directed mutagenesis experiments. This was for example shown on the example of the ER$\alpha$ where a mutation affected the association rate but not the dissociation rate of ligands indicating different pathways for either process [56]. Interestingly, researchers were able to show that the association of a coactivator to the ER slows down ligand dissociation in a fluorescence-based assay [57]. Finally, a study focusing on binding kinetics of the ER suggested different binding and unbinding pathways to be used between agonists and antagonists [58].

## The Pathways in Nuclear Receptors

**Nomenclature of Pathways.** In the past, a rather heterogenic terminology was used to describe and name pathways in the field of NRs and proteins in general. The pathways connecting the buried binding cavity of LBDs to the surrounding environment were described as pathways, paths, tunnels, gates, or channels. The term channel is typically used to describe a path leading throughout a protein, with no interruption by a larger cavity, as it is the case with ion channels, while the term tunnel is often used to describe the connecting path between a buried cavity and the surface of the protein [37, 51, 59]. The term gate is generally applied to describe a structural motif that regulates the access to a protein and therefore only describes a part of a pathway, tunnel, or channel [60].The related terms path and pathway were used most frequently (68%) in all reviewed studies and we therefore suggest their use in future studies in order to standardize the terminology for NRs. Similar to the terms used for the pathways, the terminology to distinguish them individually varied throughout the 26 studies considered for this review (Table 1). While some groups preferred numerals in both roman and arabic form, others decided to use alphabetic characters in both upper and lower case. In early studies, as well as the majority of other articles, the terminology with roman numerals was favored and we therefore suggest its use to refer to pathways in NRs.

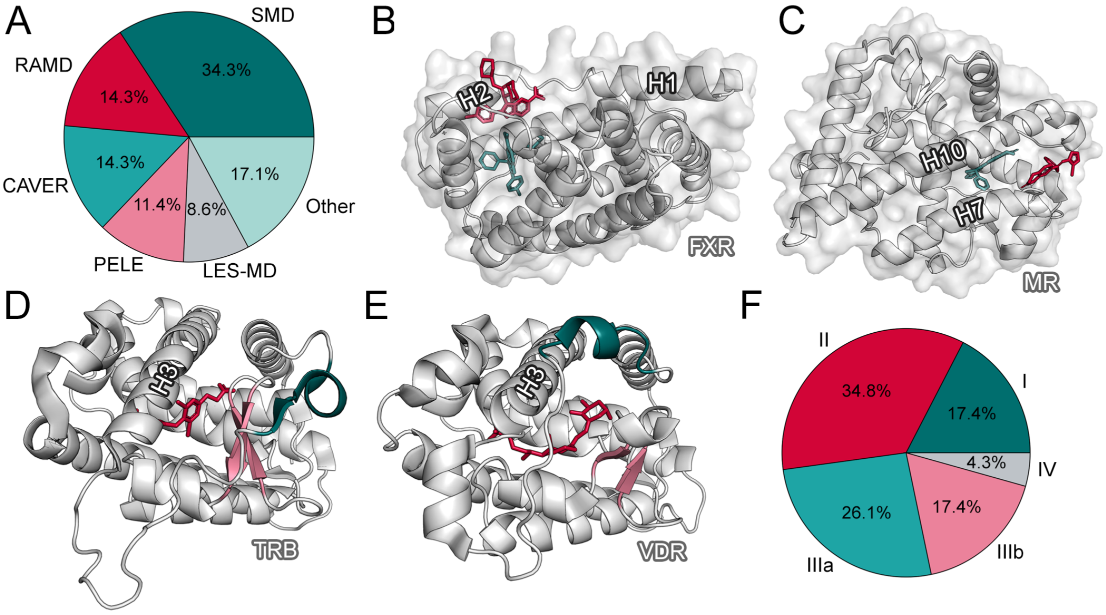


**Figure 2** Methods, crystallographic insights, and pathway statistics. (A) Methods used to determine ligand pathways in NRs in all considered studies. (B) Ligand bound to a perihperal site of FXR (PDB ID: 3OKH). The ligand in the binding pocket is shown in pine green while the peripheral ligand is shown in red. (C) Ligand bound to a perihperal site of MR (PDB ID: 3VHV). The ligand in the binding pocket is shown in pine green while the peripheral ligand is shown in red. (D) Structure of the TR$\beta$ (PDB ID: 2J4A) with the ligand (red), the two $\beta$-sheets (pink), and helix 2 (pine green). (E) Structure of VDR (PDB ID: 1IE9) with the ligand (red), the two $\beta$-sheets (pink), and helix 2 (pine green). (F) Chart showing the percentage of all studies, in which the respective pathway (either I, II, IIIa, IIIb, or IV) was described as most favorable.

**Location of Pathways.** The pathways characterized as the favored ones in all considered studies are shown in Figure 2F. Due to differences in the nomenclature of secondary structural elements that mainly arise from to distinct structural features of NR structures (cf. Figure 1C), we aligned the respective receptor structures to the AR and determined the location of the pathways based on its secondary structure (cf. Table 2). The first pathway (pathway I, cf. Figure 3A) was reported based on crystallographic data as the unliganded LDB of RAR$\gamma$ showed an extended conformation of helix 12 (H12) while the ligand-bound structure showed it tightly packed to the body of the receptor as presented in Figure 3B [52]. Based on these findings, the so-called mousetrap mechanism, according to which the ligand is trapped by H12 after binding to the receptor, was postulated and widely discussed in the literature [14, 18, 20, 34, 41]. Later, other structures of unliganded NRs without an extended conformation of H12 were determined and it was found that the orientation of this helix in RAR$\gamma$ was imposed by crystal packing effects pointing towards a misinterpretation of the data [31, 34]. Therefore, ligand binding does not directly lead to the entrapment of the ligand even though the binding is coupled to particular conformational changes of H12 that rather depend on the either agonistic or antagonistic nature of the compound [61]. In the following years, computational methods were applied to study pathways in atomic detail leading to the discovery of several new regions involved in ligand access or egress. Pathway II was reported in the largest share of NRs and it was identified as most the favorable pathway in a large number of studies. It protrudes the protein surface among the H6-H7 loop, the C-terminal region of H3, and the H11-H12 loop in a region that is characterized by a conserved plasticity among NRs which is thought to be connected to the opening of this region for ligand (un-)binding [7]. Additionally, pathway II was commonly identified in AR, ER$\alpha$, GR, MR, and PR by unbiased PELE simulations [10] and crystal structures show ligands bound to the entrance of the pathway supporting its relevance (cf. Figure 2C). Two paths located in spatial proximity of the $\beta$-hairpin, the H1-H2 loop, and H3 (cf. Figure 3A) were summarized as pathways IIIa and IIIb since they describe a similar region. The translocation through this pathway was associated with the deformation of the protein during biased simulations [9, 34]. Similar to pathway II, a ligand bound close to the entrance of pathway III indicates the importance of this path. This region is of special interest for the ligand specificity of the receptors since it is comparably variable among NRs [7] that otherwise share a common helical fold. In the TR, for example, the distinct location of H2 does not allow the discrimination between IIIa and IIIb. Pathway IV, which was detected in eight studies, protrudes through the H6-H7 loop. The last two pathways V and VI were both only described once and are therefore unlikely to be common pathways for NR ligands.

**Table 2** Description of the pathway location and receptor system in which they were described.

| *Pathway* | *Location*[a] | *Receptors*[b] |
|---|---|---|
| I | between H10, H11, and H12 | AR, ER$\alpha$, ER$\beta$, FXR, RAR$\gamma$, RXR$\alpha$, TR$\alpha$, TR$\beta$, VDR |
| II | between the H6-H7 loop, the H11-H12 loop, and the C-terminal part of H3 | AR, EcR, ER$\alpha$, ER$\beta$, FXR, GR, MR, PR, PXR, RAR$\gamma$, PPAR$\gamma$, RXR$\alpha$, TR$\alpha$, TR$\beta$, VDR |
| IIIa | between the H1-H3 loop and the $\beta$-hairpin | EcR, ER$\alpha$, ER$\beta$, FXR, PPAR$\gamma$, RAR$\gamma$, TR$\alpha$, TR$\beta$, VDR |
| IIIb | between the H1-H3 loop and the central part of H3 | AR, ER$\alpha$, ER$\beta$, RAR$\gamma$, TR$\alpha$, TR$\beta$, VDR |
| IV | through the H6-H7 loop close to the $\beta$-hairpin | EcR, ER$\alpha$, ER$\beta$, FXR, PXR, RAR$\gamma$, TR$\alpha$, TR$\beta$, VDR |
| V | between the H1-H3 loop and H6 | FXR |
| VI | between H12, H4, and the N-terminal part of H3 | ER$\beta$ |

[a] Secondary structural elements in close proximity to the pathways. The description was made based on the structure of the androgen receptor after alignment of the respective receptor. The secondary structure nomenclature was obtained from a recent review [29].

[b] The receptors, in which the respective pathway was identified are shown. In some NRs, the pathways were identified multiple times.

**Pathways in General.** Depending on the method, different numbers of pathways were identified in the corresponding NRs. While studies using RAMD, LES-MD, CAVER, or rTMD reported multiple pathways [12, 13, 14, 34, 46, 44], one SMD study [39] and one TMD study [22], as well as all PELE studies [7, 10] reported only a single pathway. Furthermore, as mentioned in the section on the applied computational methodology, most MD-based technologies are limited to either study access or egress [22, 43] which lead researchers to discuss if their results are applicable to the reverse process. While several studies came to the conclusion that NRs possess a common pathway used for both ligand binding and unbinding [7, 10, 19, 46], others argued that different pathways would be used for these processes [9, 22, 25, 40]. The latter standpoint was supported by studies using site-directed mutagenesis in combination with kinetic measurements

[16, 25, 56]. Further, it was suggested that the preferred pathway is dependent on the properties of the ligand [15, 17, 20, 25]. Properties to characterize pathways include width, length, bottleneck radius, bottleneck residues, as well as the physicochemical properties of residues forming the pathway [25, 51].

Another point of discussion were the conformational changes associated with ligand access or egress. An experimental study has suggested that large conformational changes are required for ligand binding [62]. In contrast, the majority of computational studies did not observe any massive conformational changes on the level of the protein backbone associated with either access or egress and indicate that the ligands mainly exploit the intrinsic protein flexibility to access the binding cavity [7, 12, 25, 41].Two studies discussed significant conformational rearrangements [15, 46].

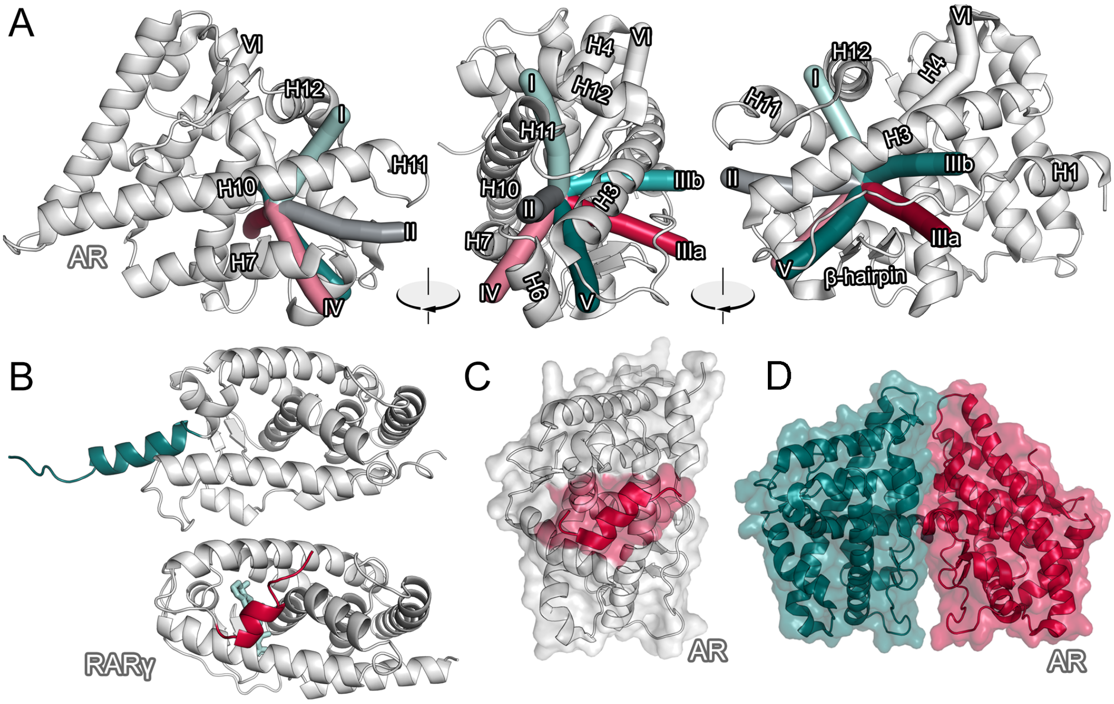


**Figure 3** Location of pathways and protein partners. (A) The location of all pathways (I, II, IIIa, IIIb, IV, V, and VI) that were characterized in NRs is shown at the example of the AR (PDB ID: 2PIV). The secondary structural elements [29] in proximity are indicated. (B) The crystal structures of apo RAR$\gamma$ (PDB ID: 1LBD, top) and holo RAR$\gamma$ (PDB ID: 2LBD, bottom) that were the basis for the mousetrap mechanism are shown. The location of H12 is indicated by specific coloring. (C) A fragment of a coactivator peptide (red) is shown on the surface of the AR (PDB ID: 1T63). (D) The dimerization interface of the AR homodimer (PDB ID: 5JJM) is shown.

Interestingly, it was suggested that peripheral sites on the protein surface could serve as recognition regions for ligands in order to capture them at the entrance of the pathway. The suggestion was based on the finding of a peripheral site on the surface of the receptors in a study with multiple NRs [7]. This supports the notion that ligands bound to the protein surface at the entry of pathways II and III in crystal structures indicate their functional relevance (cf. Figures 2B and C). Similar phenomena were observed in CYPs where it was suggested that such a two-step process consisting of recognition on the protein surface followed by the translocation through the pathway would allow kinetically efficient ligand uptake [36, 63, 64].

As mentioned above, insights from the investigation of ligand binding pathways can be used in rational structure-based drug design. While the unbiased determination of binding poses combined with compound ranking by using the PELE method is one of the newer applications, researchers used SMD simulations to investigate the binding mode of an antiandrogen linked to colchicine designed to inhibit both tubulin and the AR in prostate cancer in an earlier study [43]. On the basis of ligand-protein interactions that were observed during the translocation through ligand pathways, possibilities to increase the binding affinity of the agonist GW4064 for the FXR were recommended [18]. Martínez and colleagues offered valuable insights for the rational design of novel TR ligands with higher binding affinity based on their SMD simulations [14]. They also reviewed the results of a screening performed by Pfizer laboratories and concluded that the reduced affinity due to the removal of the phenolic group in triiodothyronine analogues was caused by the loss of an important hydrophilic interaction during ligand unbinding. Shen *et al.* suggested the removal of the polar functional group of the ligands genistein and Way-244 to improve ER$\beta$ selectivity based on SMD simulations. In addition, they determined secondary structure elements along the pathways, with the H7-H8 loop as main gatekeeper, that contribute to the selectivity between ER$\alpha$ and ER$\beta$ [17]. In a similar fashion, it was suggested that pathways determine the selectivity of genistein for ER$\beta$, since pathway IIIa offered less steric hindrance in ER$\beta$ compared to ER$\alpha$ [25]. In another study, the TR$\beta$ selectivity of GC-24 was related to its preference for pathway IIIa/b [41]. These studies demonstrate how simulations of ligands passing through NR pathways can give valuable insight for the rational design and the modification of drug compounds. For other proteins, including Src kinase, $\beta$2-adrenergic receptor, trypsin, Hsp90$\alpha$, and bifunctional epoxide hydrolase 2 it was shown that simulations of ligand binding can be used to determine binding rates and residence times of drug-like ligands in good correlation to experiments [21, 65, 66, 67].

**Influence of Protein Interaction Partners and Mutations.** The binding of coactivators and corepressors is essential for NR signaling [2, 6]. According to the currently accepted mechanism, coactivator proteins associate upon ligand binding and dissociate from the LBD after the ligand. This implies that the superficial AF-2 site, where coactivators bind, is occupied during ligand egress (cf. Figure 3C) [18, 20, 25, 68]. Laboratory experiments revealed a lower dissociation rate from the ER due to coactivator binding, which suggests a relationship to ligand pathways [62]. The effects of coactivator binding on ligand pathways was investigated in various NRs [14, 15, 69]. For example, it was suggested that pathway I would be blocked in the presence of a coactivator protein in the ER [20]. Together with the fact that coactivator binding stabilizes H12 in the agonistic position which potentially reduces the H12 plasticity mandatory for the translocation through pathway I, this adds more evidence against the postulated mousetrap mechanism [11]. Studies investigating the effect of coactivator or corepressor binding were limited to consider peptide fragments of the coactivator proteins [18, 25], due to the lack of a complete structure. Two studies came to the conclusion that the translocation through the studied ligand pathways is not affected by the presence of a coactivator protein [7, 18]. Experimental data suggests heat shock protein 90 (Hsp90), which is thought to structurally stabilize steroidal NRs in their unliganded form, to also bind at the AF-2 site. Since ligand binding is thought to induce the dissociation of Hsp90 from the receptor, it is likely that the AF-2 region would be occluded during ligand access and prevent the ligand translocation through pathway I [7, 70]. Interestingly, interactions of Hsp90 with the GR have been suggested to support the accessibility of the binding pocket and therefore promote ligand binding [7, 71]. Until today, only a minor number of all possible interaction partners was considered in the examination of ligand pathways and receptor conformations in general due to the high number of partners and missing structural information [72].

Since NR dimerization takes place after ligand binding (cf. Figure 3D) [68], its effect on egress pathways was discussed in a handful of studies [11, 13, 20, 25]. For example, dimerization hindered the translocation through certain pathways in the TRs and it was suggested that both the dimeric and monomeric states should be considered to study ligand binding [13]. Elsewhere, it was reported that several egress routes of raloxifene from the ER$\alpha$ were suppressed by dimerization [20]. Therefore, pathways leading to the dimerization interface are likely not favored for the egress from NRs [11]. Experiments revealed that the dimerization increases the half-life of the protein-ligand complex in agreement with the computational results [56]. Due to the dissimilarity of dimerization interfaces among NRs, it is likely that different results will be obtained from receptor to receptor.

Mutations in NRs are associated with several diseases and have been shown to influence the efficacy of drug therapy [61, 68, 73]. For example, MD simulations recently revealed the atomic mechanism, how single amino acid mutations in the AR can invert the effect of antagonists and promote the progression of prostate cancer [61, 74]. Since the associated conformational changes occur after ligand translocation to the binding site, this limits their relationship to ligand pathways. Nevertheless, the effect of such mutations on the pathways cannot be excluded. In addition, splice variants of the AR lacking the LDB lead to a constitutively active receptor, which is resistant to classical antiandrogen therapy and independent of ligand binding through pathways [68]. Unfortunately, the influence of amino acid mutations on ligand pathways was only characterized in two studies. In affected individuals, the I747M mutation in the GR causes glucocorticoid resistance which is associated with a variety of symptoms [40, 75]. In their simulations, Capelli and colleagues showed that the dissociation rate of dexamethasone was increased in the presence of this specific mutation and proposed this as the cause for its clinical implications [40]. This provides additional evidence for the relevance of the ligand binding mechanism for the efficacy of drug compounds and underlines the sensitivity of such simulations. The second study considered mutations in the TR$\beta$ that cause thyroid hormone resistance in their models that were subjected to LES-MD simulations. It was found that the favored egress pathway was influenced by single amino acid mutations [15]. In similar protein systems with buried active sites, such as CYPs and G protein-coupled receptors (GPCRs), it was also proposed that mutations influence properties of ligand pathways [37, 76, 77].

## Conclusions and Outlook

In the past two decades, 26 studies characterized ligand pathways in NRs and highlighted their role with computational methods. We aimed to unify the heterogenous nomenclature for pathways that was used in the past and suggest the use of roman numerals to distinguish individual pathways. The spatial location of all described pathways was summarized and revealed the most commonly described pathway located between the H6-H7 loop, H11, and H3. Recent studies report a common pathway for ligand translocation, as opposed to earlier studies reporting multiple pathways for either access or egress. Future studies should consider that the number of detected pathways might depend on the applied simulation protocol and the studied receptors. Like others, we suggest the inclusion of all protein interaction partners such as corepressors, coactivators, and dimerization partners into the simulations for optimal results since they showed to affect ligand binding through the pathways. Multiple studies focusing on ligand (un-)binding offered valuable recommendations for the design of ligands with higher affinity or improved selectivity. Further, MD simulations of ligand pathways have been shown to be sensitive enough to detect the fine effects of single amino acid mutations on the binding kinetics of ligands. However, more studies will have to consider the effect of changes in the protein sequence. Since the use of graphics processing units (GPUs) already proved to massively advance the simulation performance, sophisticated hardware will advance the field and allow researchers to study intramolecular ligand diffusion in an unbiased manner. In this regard, the PELE method to study ligand access was already proven to be an attractive alternative to classical molecular docking, since it does not require definition of the binding site, allows to determine binding energies, and incorporates flexibility, which is often neglected or simplified in docking.

## References

[1] Vineet K. Dhiman, Michael J. Bolt, and Kevin P. White. Nuclear receptors in cancer - Uncovering new and evolving roles through genomic analysis. *Nature Reviews Genetics*,

19(3):160–174, 2018.

[2] Richard Sever and Christopher K. Glass. Signaling by nuclear receptors. *Cold Spring Harbor Perspectives in Biology*, 5(3):1–4, 2013.

[3] Andrea Guerrini, Anna Tesei, Claudia Ferroni, Giulia Paganelli, Alice Zamagni, Silvia Carloni, Marzia Di Donato, Gabriella Castoria, Carlo Leonetti, Manuela Porru, Michelandrea De Cesare, Nadia Zaffaroni, Giovanni Luca Beretta, Alberto Del Rio, and Greta Varchi. A new avenue toward androgen receptor pan-antagonists: C2 sterically hindered substitution of hydroxy-propanamides. *Journal of Medicinal Chemistry*, 57(17):7263–7279, 2014.

[4] Marcella Bassetto, Salvatore Ferla, Fabrizio Pertusati, Sahar Kandil, Andrew D. Westwell, Andrea Brancale, and Christopher McGuigan. Design and synthesis of novel bicalutamide and enzalutamide derivatives as antiproliferative agents for the treatment of prostate cancer. *European Journal of Medicinal Chemistry*, 118:230–243, 2016.

[5] Chuangxing Guo, Susan Kephart, Martha Ornelas, Javier Gonzalez, Angelica Linton, Mason Pairish, Asako Nagata, Samantha Greasley, Jeff Elleraas, Natilie Hosea, Jon Engebretsen, and Andrea N. Fanjul. Discovery of 3-aryloxy-lactam analogs as potent androgen receptor full antagonists for treating castration resistant prostate cancer. *Bioorganic and Medicinal Chemistry Letters*, 22(2):1230–1236, 2012.

[6] Ni Ai, Matthew D Krasowski, William J Welsh, and Sean Ekins. Understanding nuclear receptors using computational methods. *Drug Discovery Today*, 14(9-10):486–494, 2009.

[7] Karl Edman, Ali Hosseini, Magnus K Bjursell, Anna Aagaard, Lisa Wissler, Anders Gunnarsson, Tim Kaminski, Christian Köhler, Stefan Bäckström, Tina J Jensen, Anders Cavallin, Ulla Karlsson, Ewa Nilsson, Daniel Lecina, Ryoji Takahashi, Christoph Grebner, Stefan Geschwindner, Matti Lepistö, Anders C Hogner, and Victor Guallar. Ligand Binding Mechanism in Steroid Receptors: From Conserved Plasticity to Differential Evolutionary Constraints. *Structure*, 23(12):2280–2291, 2015.

[8] Stefano Motta, Lara Callea, Sara Giani Tagliabue, and Laura Bonati. Exploring the PXR ligand binding mechanism with advanced Molecular Dynamics methods. *Scientific Reports*, 8(1):1–12, 2018.

[9] Dorina Kosztin, Sergei Izrailev, and Klaus Schulten. Unbinding of retinoic acid from its receptor studied by steered molecular dynamics. *Biophysical Journal*, 76(1 I):188–197, 1999.

[10] Christoph Grebner, Daniel Lecina, Victor Gil, Johan Ulander, Pia Hansson, Anita Dellsen, Christian Tyrchan, Karl Edman, Anders Hogner, and Victor Guallar. Exploring Binding Mechanisms in Nuclear Hormone Receptors by Monte Carlo and X-ray-derived Motions. *Biophysical Journal*, 112(6):1147–1156, 2017.

[11] Mikael Peräkylä. Ligand unbinding pathways from the vitamin D receptor studied by molecular dynamics simulations. *European Biophysics Journal*, 38(2):185–198, 2009.

[12] D Genest, N Garnier, A Arrault, C Marot, L Morin-Allory, and M Genest. Ligand-escape pathways from the ligand-binding domain of PPARgamma receptor as probed by molecular dynamics simulations. *European biophysics journal : EBJ*, 37(4):369–379, 4 2008.

[13] Shulin Zhuang, Lingling Bao, Apichart Linhananta, and Weiping Liu. Molecular modeling revealed that ligand dissociation from thyroid hormone receptors is affected by receptor heterodimerization. *Journal of Molecular Graphics and Modelling*, 44:155–160, 2013.

[14] Leandro Martínez, Paul Webb, Igor Polikarpov, and Munir S Skaf. Molecular dynamics simulations of ligand dissociation from thyroid hormone receptors: Evidence of the likeliest escape pathway and its implications for the design of novel ligands. *Journal of Medicinal Chemistry*, 49(1):23–26, 2006.

[15] Leandro Martínez, Milton T Sonoda, Paul Webb, John D Baxter, Munir S Skaf, and Igor Polikarpov. Molecular dynamics simulations reveal multiple pathways of ligand dissociation from thyroid hormone receptors. *Biophysical Journal*, 89(3):2011–2023, 2005.

[16] Xueping Hu, Song Hu, Jiazhe Wang, Yawen Yanhong Dong, Li Zhang, and Yawen Yanhong Dong. Steered molecular dynamics for studying ligand unbinding of ecdysone receptor. *Journal of Biomolecular Structure and Dynamics*, 1102:1–10, 2017.

[17] Jie Shen, Weihua Li, Guixia Liu, Yun Tang, and Hualiang Jiang. Computational insights into the mechanism of ligand unbinding and selectivity of estrogen receptors. *Journal of Physical Chemistry B*, 113(30):10436–10444, 2009.

[18] Weihua Li, Jing Fu, Feixiong Cheng, Mingyue Zheng, Jian Zhang, Guixia Liu, and Yun Tang. Unbinding pathways of GW4064 from human farnesoid X receptor as revealed by molecular dynamics simulations. *Journal of Chemical Information and Modeling*, 52(11): 3043–3052, 2012.

[19] Peter Carlsson, Sofia Burendahl, and Lennart Nilsson. Unbinding of Retinoic Acid from the Retinoic Acid Receptor by Random Expulsion Molecular Dynamics. *Biophysical Journal*, 91(9):3151–3161, 2006.

[20] Milton T Sonoda, Leandro Martínez, Paul Webb, Munir S Skaf, and Igor Polikarpov. Ligand dissociation from estrogen receptor is mediated by receptor dimerization: evidence from molecular dynamics simulations. *Mol. Endocrinol.*, 22(7):1565–1578, 2008.

[21] Yibing Shan, Eric T. Kim, Michael P. Eastwood, Ron O. Dror, Markus A. Seeliger, and David E. Shaw. How does a drug molecule find its target binding site? *Journal of the American Chemical Society*, 133(24):9181–9183, 2011.

[22] Samia Aci-Sèche, Monique Genest, and Norbert Garnier. Ligand entry pathways in the ligand binding domain of PPAR$\gamma$ receptor. *FEBS letters*, 585(16):2599–2603, 8 2011.

[23] Veronica Salmaso and Stefano Moro. Bridging molecular docking to molecular dynamics in exploring ligand-protein recognition process: An overview. *Frontiers in Pharmacology*, 9(AUG):1–16, 2018.

[24] Jing Fang, Jie Shen, Feixiong Cheng, Zhejun Xu, Guixia Liu, and Yun Tang. Computational insights into ligand selectivity of estrogen receptors from pharmacophore modeling. *Molecular Informatics*, 30(6-7):539–549, 2011.

[25] Sofia Burendahl, Cristian Danciulescu, and Lennart Nilsson. Ligand unbinding from the estrogen receptor: A computational study of pathways and ligand specificity. *Proteins: Structure, Function and Bioinformatics*, 77(4):842–856, 2009.

[26] Emily R. Weikum, Xu Liu, and Eric A. Ortlund. The nuclear receptor superfamily: A structural perspective. *Protein Science*, 27(11):1876–1892, 2018.

[27] Iain J. McEwan, Derek Lavery, Katharina Fischer, and Kate Watt. Natural Disordered Sequences in the Amino Terminal Domain of Nuclear Receptors: Lessons from the Androgen and Glucocorticoid Receptors. *Nuclear Receptor Signaling*, 5(1):nrs.05001, 2009.

[28] Schrodinger LLC. The PyMOL Molecular Graphics System, Version 2.1.1. 2018.

[29] Mh Eileen Tan, Jun Li, H. Eric Xu, Karsten Melcher, and Eu Leong Yong. Androgen receptor: Structure, role in prostate cancer and drug discovery. *Acta Pharmacologica Sinica*, 36(1):3–23, 2015.

[30] Fraydoon Rastinejad, Pengxiang Huang, Vikas Chandra, and Sepideh Khorasanizadeh. Understanding nuclear receptor form and function using structural biology. *Journal of molecular endocrinology*, 51(3):T1–T21, 12 2013.

[31] Fraydoon Rastinejad, Vincent Ollendorff, and Igor Polikarpov. Nuclear receptor full-length architectures: Confronting myth and illusion with high resolution. *Trends in Biochemical Sciences*, 40(1):16–24, 2015.

[32] Qinchang Chen, Haoyue Tan, Hongxia Yu, and Wei Shi. Activation of steroid hormone receptors: Shed light on the in silico evaluation of endocrine disrupting chemicals. *Science of the Total Environment*, 631-632:27–39, 2018.

[33] Tomasz Pieńko and Joanna Trylska. Computational Methods Used to Explore Transport Events in Biological Systems. *Journal of Chemical Information and Modeling*, page acs.jcim.8b00974, 2019.

[34] Arnaud Blondel, Jean Paul Renaud, Stefan Fischer, Dino Moras, and Martin Karplus. Retinoic acid receptor: A simulation analysis of retinoic acid binding and the resulting conformational changes. *Journal of Molecular Biology*, 291(1):101–115, 1999.

[35] Armin Madadkar-Sobhani and Victor Guallar. PELE web server: atomistic study of biomolecular systems at your fingertips. *Nucleic acids research*, 41(Web Server issue): 322–328, 2013.

[36] Philippe Urban, Thomas Lautier, Denis Pompon, and Gilles Truan. Ligand Access Channels in Cytochrome P450 Enzymes: A Review. *Int J Mol Sci.*, 19(6), 5 2018.

[37] André Fischer, Charleen G. Don, and Martin Smieško. Molecular Dynamics Simulations Reveal Structural Differences among Allelic Variants of Membrane-Anchored Cytochrome P450 2D6. *Journal of Chemical Information and Modeling*, 58(9):1962–1975, 2018.

[38] Phuc Chau Do, Eric H. Lee, and Ly Le. Steered Molecular Dynamics Simulation in Rational Drug Design. *Journal of Chemical Information and Modeling*, 58(8):1473–1482, 2018.

[39] Nguyen Quoc Thai, Hoang Linh Nguyen, Huynh Quang Linh, and Mai Suan Li. Protocol for fast screening of multi-target drug candidates: Application to Alzheimer's disease. *Journal of Molecular Graphics and Modelling*, 77:121–129, 2017.

[40] Anna Maria Capelli, Agostino Bruno, Antonio Entrena Guadix, and Gabriele Costantino. Unbinding pathways from the glucocorticoid receptor shed light on the reduced sensitivity of glucocorticoid ligands to a naturally occurring, clinically relevant mutant receptor. *Journal of Medicinal Chemistry*, 56(17):7003–7014, 2013.

[41] Leandro Martínez, Igor Polikarpov, and Munir S. Skaf. Only subtle protein conformational adaptations are required for ligand binding to thyroid hormone receptors: Simulations using a novel multipoint steered molecular dynamics approach. *Journal of Physical Chemistry B*, 112(34):10741–10751, 2008.

[42] Susanna K. Lüdemann, Valère Lounnas, and Rebecca C. Wade. How do substrates enter and products exit the buried active site of cytochrome P450cam? 2. Steered molecular dynamics and adiabatic mapping of substrate pathways. *Journal of Molecular Biology*, 303(5):813–830, 2000.

[43] Nima Sharifi, Ernest Hamel, Markus A Lill, Prabhakar Risbood, Charles T Kane, Md Tafazzal Hossain, Amanda Jones, James T Dalton, and William L Farrar. A bifunctional colchicinoid that binds to the androgen receptor. *Molecular cancer therapeutics*, 6 (8):2328–2336, 2007.

[44] Lautaro D. Alvarez, Adriana S. Veleiro, and Gerardo Burton. Exploring the molecular basis of action of ring D aromatic steroidal antiestrogens. *Proteins: Structure, Function and Bioinformatics*, 83(7):1297–1306, 2015.

[45] Motonori Tsuji. A ligand-entry surface of the nuclear receptor superfamily consists of the helix H3 of the ligand-binding domain. *Journal of molecular graphics and modelling*, 62: 262–275, 2015.

[46] Atif Zafar, Sabahuddin Ahmad, and Imrana Naseem. Insight into the structural stability of coumestrol with human estrogen receptor {$\alpha{\$}} and {$\beta{\$}} subtypes: A combined approach involving docking and molecular dynamics simulation studies. *RSC Advances*, 5(99):81295–81312, 2015.

[47] Christoph Grebner, Jessica Iegre, Johan Ulander, Karl Edman, Anders Hogner, and Christian Tyrchan. Binding Mode and Induced Fit Predictions for Prospective Computational Drug Design. *Journal of Chemical Information and Modeling*, 56(4):774–787, 2016.

[48] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, T N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 1 2000.

[49] Sunghwan Kim, Paul A Thiessen, Evan E Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A Shoemaker, Jiyao Wang, Bo Yu, Jian Zhang, and Stephen H Bryant. PubChem Substance and Compound databases. *Nucleic acids research*, 44(D1):1202–13, 1 2016.

[50] Andrea Spitaleri, Sergio Decherchi, Andrea Cavalli, and Walter Rocchia. Fast Dynamic Docking Guided by Adaptive Electrostatic Bias: The MD-Binding Approach. *Journal of Chemical Theory and Computation*, 14(3):1727–1736, 2018.

[51] Eva Chovancova, Antonin Pavelka, Petr Benes, Ondrej Strnad, Jan Brezovsky, Barbora Kozlikova, Artur Gora, Vilem Sustr, Martin Klvana, Petr Medek, Lada Biedermannova, Jiri Sochor, and Jiri Damborsky. CAVER 3.0: A Tool for the Analysis of Transport Pathways in Dynamic Protein Structures. *PLoS Computational Biology*, 8(10):23–30, 2012.

[52] Renaud-Nature-1995.

[53] Hans G F Richter, Gregory M Benson, Denise Blum, Evelyne Chaput, Song Feng, Christophe Gardes, Uwe Grether, Peter Hartman, Bernd Kuhn, Rainer E Martin, Jean Marc Plancher, Markus G Rudolph, Franz Schuler, Sven Taylor, and Konrad H Bleicher. Discovery of novel and orally active FXR agonists for the potential treatment of dyslipidemia and diabetes. *Bioorganic and Medicinal Chemistry Letters*, 21(1):191–194, 2011.

[54] Tomoaki Hasui, Nobuyuki Matsunaga, Taiichi Ora, Norio Ohyabu, Nobuhiro Nishigaki, Yoshimi Imura, Yumiko Igata, Hideki Matsui, Takashi Motoyaji, Toshimasa Tanaka, Noriyuki Habuka, Satoshi Sogabe, Midori Ono, Christopher S Siedem, Tony P Tang, Cassandra Gauthier, Lisa A De Meese, Steven A Boyd, and Shoji Fukumoto. Identification of benzoxazin-3-one derivatives as novel, potent, and selective nonsteroidal mineralocorticoid receptor antagonists. *Journal of Medicinal Chemistry*, 54(24):8616–8631, 2011.

[55] A C Gee and J A Katzenellenbogen. Probing conformational changes in the estrogen receptor: evidence for a partially unfolded intermediate facilitating ligand binding and release. *Molecular endocrinology (Baltimore, Md.)*, 15(3):421–428, 3 2001.

[56] L. Zhong and D. F. Skafar. Mutations of tyrosine 537 in the human estrogen receptor-$\alpha$ selectively alter the receptor's affinity for estradiol and the kinetics of the interaction. *Biochemistry*, 41(13):4209–4217, 2002.

[57] A C Gee, K E Carlson, P G Martini, B S Katzenellenbogen, and J A Katzenellenbogen. Coactivator peptides have a differential stabilizing effect on the binding of estrogens and antiestrogens with the estrogen receptor. *Molecular endocrinology (Baltimore, Md.)*, 13 (11):1912–1923, 11 1999.

[58] Rebecca L Rich, Lise R Hoth, Kieran F Geoghegan, Thomas A Brown, Peter K LeMotte, Samuel P Simons, Preston Hensley, and David G Myszka. Kinetic analysis of estrogen receptor/ligand interactions. *Proceedings of the National Academy of Sciences of the United States of America*, 99(13):8562–7, 2002.

[59] Petr Jeřábek, Jan Florián, Václav Martínek, P Jerabek, J Florian, and V Martinek. Lipid molecules can induce an opening of membrane-facing tunnels in cytochrome P450 1A2. *Phys. Chem. Chem. Phys.*, 18(44):30344–30356, 2016.

[60] Artur Gora, Jan Brezovsky, and Jiri Damborsky. Gates of enzymes. *Chemical Reviews*, 113(8):5871–5923, 2013.

[61] Na Liu, Wenfang Zhou, Yue Guo, Junmei Wang, Weitao Fu, Huiyong Sun, Dan Li, Mojie Duan, and Tingjun Hou. Molecular Dynamics Simulations Revealed the Regulation of Ligands to the Interactions between Androgen Receptor and its Coactivator. *Journal of Chemical Information and Modeling*, 58:1652–1661, 2018.

[62] Arvin C. Gee and John A. Katzenellenbogen. Probing Conformational Changes in the Estrogen Receptor: Evidence for a Partially Unfolded Intermediate Facilitating Ligand Binding and Release. *Molecular Endocrinology*, 15(3):421–428, 2001.

[63] Yao Huili, McCullough Christopher R., Costache Aurora D., Pullela Phani Kumar, and Sem Daniel S. Structural evidence for a functionally relevant second camphor binding site in P450cam: Model for substrate entry into a P450 active site. *Proteins.*, 69(1):125–138, 6 2007.

[64] Emre M. Isin and F. Peter Guengerich. Kinetics and thermodynamics of ligand binding by cytochrome P450 3A4. *Journal of Biological Chemistry*, 281(14):9127–9136, 2006.

[65] R. O. Dror, A. C. Pan, D. H. Arlow, D. W. Borhani, P. Maragakis, Y. Shan, H. Xu, and D. E. Shaw. Pathway and mechanism of drug binding to G-protein-coupled receptors. *Proceedings of the National Academy of Sciences*, 108(32):13118–13123, 2011.

[66] Daria B. Kokh, Marta Amaral, Joerg Bomke, Ulrich Grädler, Djordje Musil, Hans Peter Buchstaller, Matthias K. Dreyer, Matthias Frech, Maryse Lowinski, Francois Vallee, Marc Bianciotto, Alexey Rak, and Rebecca C. Wade. Estimation of Drug-Target Residence Times by $\tau$-Random Acceleration Molecular Dynamics Simulations. *Journal of Chemical Theory and Computation*, 14(7):3859–3869, 2018.

[67] Samuel D. Lotz and Alex Dickson. Unbiased Molecular Dynamics of 11 min Timescale Drug Unbinding Reveals Transition State Stabilizing Interactions. *Journal of the American Chemical Society*, 140(2):618–628, 2018.

[68] Peter E Lonergan and Donald J Tindall. Androgen receptor signaling in prostate cancer development and progression. *Journal of Carcinogenesis*, 10:20, 8 2011.

[69] Sofia Burendahl and Lennart Nilsson. Computational studies of LXR molecular interactions reveal an allosteric communication pathway. *Proteins: Structure, Function and Bioinformatics*, 80(1):294–306, 2012.

[70] Tommi Manninen, Sami Purmonen, and Timo Ylikomi. Interaction of nuclear receptors with hsp90 in living cells. *Journal of Steroid Biochemistry and Molecular Biology*, 96(1): 13–18, 2005.

[71] D. Ricketson, U. Hostick, L. Fang, K. R. Yamamoto, and B. D. Darimont. A Conformational Switch in the Ligand-binding Domain Regulates the Dependence of the Glucocorticoid Receptor on Hsp90. *Journal of Molecular Biology*, 368(3):729–741, 2007.

[72] Sepideh Khorasanizadeh and Fraydoon Rastinejad. Visualizing the architectures and interactions of nuclear receptors. *Endocrinology*, 157(11):4212–4221, 2016.

[73] John C. Achermann, John Schwabe, Louise Fairall, and Krishna Chatterjee. Genetic disorders of nuclear receptors. *Journal of Clinical Investigation*, 127(4):1181–1192, 2017.

[74] Ye Jin, Mojie Duan, Xuwen Wang, Xiaotian Kong, Wenfang Zhou, Huiyong Sun, Hui Liu, Dan Li, Huidong Yu, Youyong Li, and Tingjun Hou. Communication between the Ligand-Binding Pocket and the Activation Function-2 Domain of Androgen Receptor Revealed by Molecular Dynamics Simulations. *Journal of Chemical Information and Modeling*, 59 (2):842–857, 2019.

[75] Evangelia Charmandari, Tomoshige Kino, Takamasa Ichijo, and George P Chrousos. Generalized glucocorticoid resistance: clinical aspects, molecular mechanisms, and implications of a rare genetic disorder. *The Journal of clinical endocrinology and metabolism*, 93 (5):1563–1572, 5 2008.

[76] Parker W. De Waal, Kyle F. Sunden, and Laura Lowe Furge. Molecular dynamics of CYP2D6 polymorphisms in the absence and presence of a mechanism-based inactivator reveals changes in local flexibility and dominant substrate access channels. *PLoS ONE*, 9 (10), 2014.

[77] Henriette Stoy and Vsevolod V Gurevich. How genetic errors in GPCRs affect their function: Possible therapeutic strategies. *Genes and Diseases*, 2(2):108–132, 2015.

# CHAPTER 6

# Ligand Pathways in Estrogen-Related Receptors

While working on the information retrieval for the review presented in Chapter 5, I realized a lack of structural studies on estrogen-related receptors (ERRs). Due to by recent developments regarding the understanding of the relevance of ERRs for health and disease, this work was focused on elucidating the recognition process of their ligands through pathways connecting their active site to the surrounding solvent. Detailed atomic knowledge on molecular recognition in these receptors may support the rational design and prediction of modulators.

---

**Author contributions:** Conceptualization, A.F.; formal analysis, A.F., F.B.; writing and original draft preparation, A.F.; writing, review and editing, A.F., M.A., M.S.; visualization, A.F.; supervision, M.A., M.S.

---

*Based on manuscript submitted to J. Biomol. Struct. Dyn.*

Fischer, A.; Ferhat, B.; Sellner, M.; Lill, M.A.; Smieško, M. Ligand Pathways in Estrogen-Related Receptors.

## Abstract

The three subtypes of estrogen-related receptors ERR$\alpha$, ERR$\beta$, and ERR$\gamma$ are nuclear receptors mediating metabolic processes in various tissues such as the skeletal muscle, fat tissue, bone, and liver. Although the knowledge on their physiological ligands is limited, they have been implicated as drug targets for important indications including diabetes, cardiovascular diseases, and osteoporosis. As in other nuclear receptors, their ligand binding pocket is buried within the core of the receptor and connected to its surrounding by ligand pathways. Here, we investigated these pathways with conventional molecular dynamics as well as metadynamics simulations to reveal their distribution and their capability to facilitate ligand translocation. Dependent on the ERR subtype and the conformational state of the receptor, we could detect different pathways to be favored. Overall, the results suggested pathways IIIa and IIIb to be favored in the agonistic conformation, while antagonists preferred pathways I, II, and V. Along the pathways, the ligands passed different gating mechanisms of the receptor, including groups of protein residues as well as whole secondary structure elements, to leave the binding site. Even though these pathways are suggested to influence ligand specificity of the receptors and their elucidation might advance rational drug design, they have not yet been studied in ERRs.

## Introduction

Estrogen related receptors (ERR) belong to the superfamily of nuclear receptors (NRs). Despite their nomenclature, which originates from their sequence homology to estrogen receptors (ERs), estrogens have not been found to strongly interact with ERRs [1, 2]. However, the three subtypes ERR$\alpha$, ERR$\beta$, and ERR$\gamma$ interfere with ER signaling as they share transcriptional targets. While the ERs have been intensively studied, less is known about ERRs. They are mainly expressed in skeletal muscle, fat, bone, liver, as well as the brain tissue and are of particular interest due to their involvement in metabolic processes [3]. For example, ERR$\alpha$ regulates mitochondrial biogenesis and muscle regeneration, while ERR$\gamma$ modulates oxidative phosphorylation and angiogenesis in the skeletal muscle. ERR$\beta$ is involved in maintaining embrionic stem cell pluripotency. Other activities of ERRs relate to the immune system, bone physiology, absorption of lipids in the gastrointestinal tract, and lipid metabolism [4]. Due to their involvement in various physiological processes, ERRs have been proposed as therapeutic targets for multiple indications such as diabetes, cardiovascular diseases, osteoporosis, and muscle atrophy [4, 5, 6]. Originally, ERRs were designated as orphan receptors, as no endogenous ligands were known regulating their signaling, probably because they are constitutively active [3, 4]. However, recent evidence suggests that cholesterol might be a natural ligand of ERR$\alpha$ [7]. Interestingly, multiple small molecules with therapeutic potential have been discovered to influence ERR signaling. In the comparatively small binding pocket of ERRs, agonists can increase the basal signaling of the receptors, while antagonists can decrease it [4, 8, 9]. Several drug discovery programs investigated compounds modulating ERR activity, but none of them have reached market approval [3]. Thus, there remains a medical need for novel ERR modulators to treat human diseases [4].

In general, NRs present a multi-domain organization with an N-terminal domain modulating protein-protein interactions, a DNA-binding domain mediating the interaction with the DNA, and a ligand-binding domain (LBD) with a small-molecule binding pocket primarily involved in the regulation of receptor signaling by small molecules [2, 3, 10]. Due to its buried character, ligands need to translocate through pathways within the protein to reach the binding pocket. Depending on the studied NR, different pathways might facilitate this translocation and contribute to the ligand-specificity of the receptor. Knowledge on the mechanism and occurrence of ligand pathways can facilitate rational structure-based design of novel ligands [11, 12, 13]. As translocation typically requires conformational changes of the protein, such pathways are rarely observed in static crystal structures. Computational techniques such as molecular dynamics (MD) simulations can model the inherent protein flexibility and, therefore, be used to localize the pathways and evaluate their capability to translocate ligands. Ligand pathways can be explored independent of a ligand with small spherical probes, or in the context of an actual ligand propagating through the pathway [11]. The use of conventional MD simulations to study ligand translocation is highly demanding with regards to computational resources and often not very efficient due to the long time scale of such molecular events. In contrast, specific sampling techniques such as accelerated MD, steered MD, or metadynamics simulations have been successfully applied to various protein systems [11, 14, 15, 16]. Previous computational work on ERRs was limited to the analysis of ligand-induced conformational changes and ligand-protein interactions [17, 18, 19]. While several studies have been conducted to examine other NRs, the characteristics of ligand pathways in ERRs has, to the best of our knowledge, not been addressed until today. For example, Capelli and colleagues have shown that the dissociation rate of a ligand binding to the glucocorticoid receptor is increased by a mutation along a ligand pathway leading to decreased ligand efficacy [11, 13]. In the field of ERs, rational modification of a ligand was used to improve its selectivity for ER$\beta$ based on steered MD simulations along a pathway [11, 12]. On the other hand, metadynamics simulations have not been applied to study ligand pathways in NRs [11]. In this simulation technique, so-called collective variables (CVs) need to be predefined for the algorithm to bias the system towards sampling CV space. The applied bias is history-dependent in order to sample new states of the system and, therefore, allows to sample rare molecular events such as ligand-protein association and dissociation as well as to obtain a free energy surface of the process [20, 21].

In this study, we conducted conventional MD and metadynamics simulations to elucidate the location and functionality of pathways for ligands to translocate to and from the buried binding pocket in ERRs. By carefully analyzing over 9 $\mu$s of MD trajectories, we highlighted the distribution and opening of pathways in the ERR subtypes independent of a ligand, and brought this into relation with the preferred pathways determined in the metadynamics simulations. We analyzed the associated conformational changes on the residue as well as the secondary structure level and map the deposited biasing potential during ligand translocation. Even though these pathways are known to influence ligand specificity of the receptors, they have not previously been investigated in ERRs despite their therapeutic relevance.

## Results and Discussion

**Model building and validation.** Of the four available crystal structures of ERR$\alpha$, two structures presented an agonistic conformation without any ligand bound, while the remaining two structures were bound to an antagonist with a characteristic displacement of helix-12 (H12) towards the coactivator binding site [22, 23]. As our primary goal was to investigate ligand-dependent phenomena, we did not consider ERR$\alpha$ in its agonistic constitutively active conformation due to the absence of a ligand and focused on antagonist-bound structures. The available crystal structures indicate the agonistic and apo conformation of ERR$\gamma$ to be highly similar (PDB IDs: 2P7G and 2ZBS). In the case of the selected structure for ERR$\alpha$ (Table 1), the ligand Q27455709 (PubChem ID: 49866529) was cocrystallized (Figure 1A). For ERR$\gamma$, structures with both agonists and antagonists were available [24] leading us to select one structure each (Table 1). While the inverse agonist DN200434 (PubChem ID: 377642864) was bound to the active site in the structure resembling the antagonist conformation with H12 displacement (Figures 1B-D), the other structure was cocrystallized with the agonist bisphenol A.

**Table 1** Overview of the structures considered in this study.

| System | Structure | Ligand | Mode of action[a] |
|---|---|---|---|
| ERR$\alpha$ | 3K6P | Q27455709 | antagonist |
| ERR$\beta$ | 5YSO [b] | diethylstilbestrol [c] | antagonist |
| ERR$\gamma$ ago | 2P7G | bisphenol A | agonist |
| ERR$\gamma$ antago | 5YSO | DN200434 | antagonist |

[a] The mode of action also represents the conformational state of the receptors. [b] Homology model based on ERR$\gamma$ crystal structure. [c] Ligand orientation obtained by docking.

While ligand-bound structures were available in the Protein Data Bank for ERR$\alpha$ and ERR$\gamma$, there were no crystal structures of the LBD published for ERR$\beta$ when we initiated this project. However, only recently, two structures (PDB IDs: 6LIT and 6LN4) resembling an agonistic conformation were deposited in the Protein Data Bank (PDB) [25]. Nevertheless, we constructed a homology model of ERR$\beta$ in an antagonistic conformation using the SWISS-MODEL [26] web server with an ERR$\gamma$ crystal structure (PDB ID: 5YSO) serving as template as it has a higher sequence identity to ERR$\beta$ than to ERR$\alpha$ (Figure S1). The template structure presented a sequence identity of 79.2% and the resulting model presented a good Global Model Quality Estimation (GMQE) index of 0.85 and a Qualitative Model Energy Analysis (QMEAN) score of -0.26 (Figure 1B). The latter metric describes the absolute quality of a protein structure and the obtained value indicated a high nativeness of the model [26, 27, 28]. As mentioned above, we were able to retrospectively validate the correctness of our homology model based on a newly available crystal structure of the receptor that, however, lacks the N-terminal region of H1 and H2 and represents an agonistic conformation of the receptor as indicated by the co-crystallized coactivator fragment and the orientation of H12 [25]. After removing the N-terminus of our homology model, as well as both C-termini, we compared both structures and observed an impressive similarity. While the backbone root-mean square deviation (RMSD) of superposition amounted to only 0.70 Å, the overall heavy-atom RMSD was 1.51 Å, clearly below the fluctuations we observed in the corresponding MD simulations (Figure S2). In conclusion, these metrics justified the use of our model for further procedures.

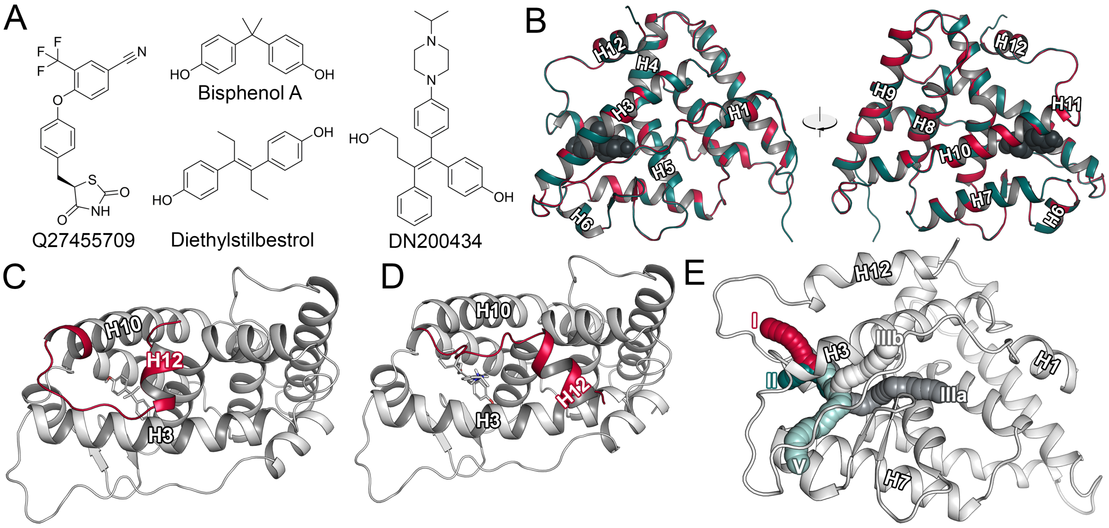


**Figure 1** Structural overview. (A) Structures of the ligands considered in this work. (B) Alignment between our ERR$\beta$ homology model and the selected ERR$\gamma$ crystal structure from two different orientations. Secondary structure elements were assigned according to our previous work [11]. (C) Agonistic conformation of ERR$\gamma$ (PDB ID: 2P7G) with H12 colored in red. (D) Antagonistic conformation of ERR$\gamma$ (PDB ID: 5YSO) with H12 colored in red. (E) Ligand pathways in ERR$\alpha$.

As the active site was in an apo state after generating the homology model, we aimed to dock a known antagonist to the model. The prior validation of the smina [29] docking protocol (Figure S3A-E) displayed acceptable to high accuracy (RMSDs between 0.3 to 2.2 Å) in reproducing cocrystallized binding modes in all three ERRs. Interestingly, the binding mode of the docked antagonist diethylstilbestrol overlapped excellently with the cocrystallized pose in the template of ERR$\gamma$ structure providing additional confidence into the obtained pose (Figure S3F).

The RMSD values of all MD simulations generally indicated good convergence with two exceptions: one replica of the conventional simulations of ERR$\gamma$ bound to an agonist and one metadynamics simulation of ERR$\beta$ displayed a spike in the diagram towards the end of the trajectory (Figure S2). Both RMSD spikes were caused by a distortion of the N-terminal region of H3, as we will elaborate on in the following sections. In the metadynamics simulation the change took place after ligand egress. As our analysis of ligand translocation ended with the ligand egress, this RMSD peak has no effect on our overall conclusions. The fact that the receptors underwent this conformational change in both conventional and biased simulations reduces the chance that it was an artifact imposed by the biasing potential.

**Different patterns of ligand pathways in ERRs.** While crystal structures of NRs generally do not allow an immediate detection of routes for ligand access or egress, protein flexibility introduced by MD simulations may trigger conformational fluctuations that help to unmask the ligand pathways. Using various computational protocols based on MD simulations, multiple possible egress routes for ligands were reported in NRs [11]. Here, we first applied the CAVER protocol allowing to determine and visualize pathways independent of a ligand molecule based on conventional MD simulations [30]. In a ligand pathway, the most narrow point is referred to as its bottleneck, where gating residues typically act as molecular filters contributing to ligand specificity [15, 11]. The bottleneck radius describes the width of the pathway at this particular location. Interestingly, our analysis of average bottleneck radii during the simulations showed different pathways to be present in the three ERR subtypes (Figures 1E and 2A). In ERR$\alpha$, for example, we could detect a comparatively large opening of pathway II (pw-II) located among the H6-H7 loop, the H11-H12 loop, and the C-terminal region of H3 with an average bottleneck radius of up to 2.4 Å. Also in the other receptor systems, pw-II was present in all simulation replicas and, hence, the most abundant pathway among the studied receptors. In the related proteins ER$\alpha$ and ER$\beta$, pw-II has been previously described as major access pathway for estradiol with a Monte Carlo based technique termed protein energy landscape exploration [31] in agreement with our results.

Another study on ERs using the same MD-based methodology we used here, found pw-IIIa to be the most likely access and egress route for ligands. In our simulations, pw-IIIa presented the largest bottleneck radii in the agonistic state of ERR$\gamma$. As the mentioned

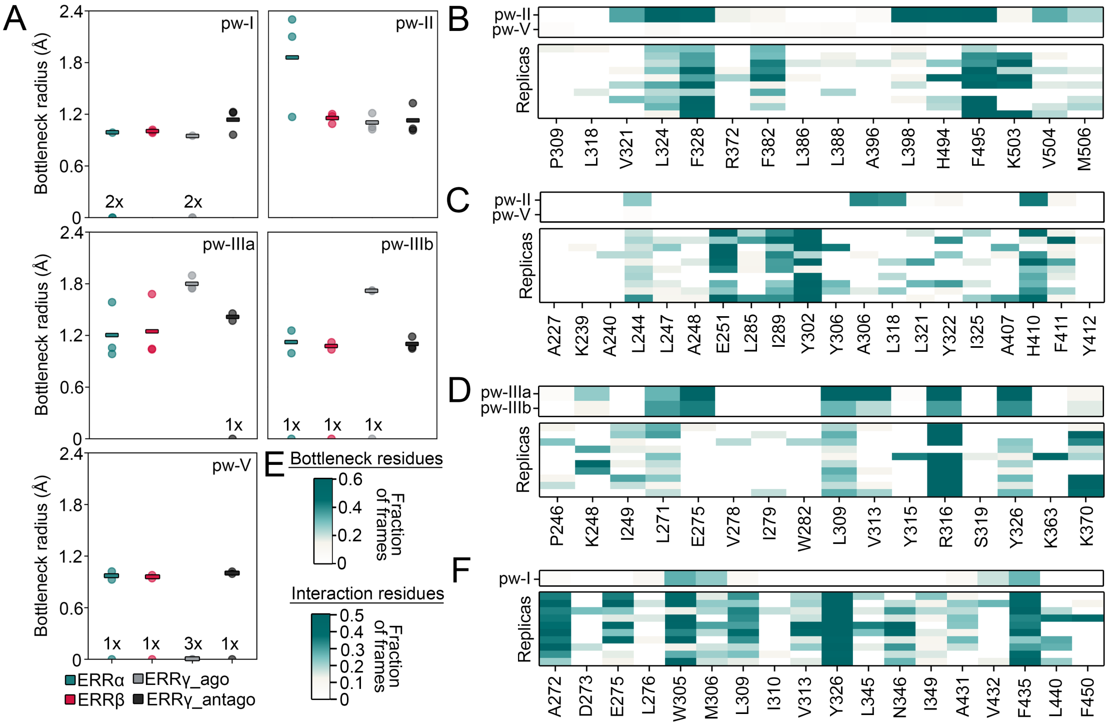


**Figure 2** Characteristics of the pathways. (A) Average bottleneck radii for pathways I, II, IIIa, IIIb, and V in the three ERR subtypes. The average of the three replicas (null values not included in the average) is indicated by a horizontal line. The number of null values was indicated if any were present. (B) On top, the occurrence of a residue as bottleneck in a pathway is shown as a heat map, while on the bottom ligand-protein interactions in the ten metadynamics replicas are illustrated. The data is given for ERR$\alpha$, (C) ERR$\beta$, (D) ERR$\gamma$ with agonist, and (F) ERR$\gamma$ with antagonist. (E) Legends for the heat maps.

study also evaluated the agonistic conformation of ER$\alpha$ and ER$\beta$, our results indicate that this preference for pw-IIIa in ERs is also applicable for ERRs, as we will further underline in the following sections. As their nomenclature suggests, the two pathways pw-IIIa and pw-IIIb are only separated by the highly flexible region between H1 and H3 [11]. Similar to pw-IIIa, pw-IIIb was most open in the agonist-bound ERR$\gamma$ indicating this structural region to be relevant for ligand translocation if H12 rests in an orientation that allows the interaction with coactivators. In the same complex, we could not detect pw-V in any of the replica simulations. In the antagonistic conformation of ERR$\gamma$, pw-I presented the highest bottleneck radius compared to the remaining receptors (Tables S1 and S2). As we highlight later on, this also was the only pathway selected by the ligand in our metadynamics simulations. Moreover, pw-V, which we observed in metadynamics simulations of both ERR$\alpha$ and ERR$\beta$, only displayed a small opening in all receptors. In all studied systems, pw-IV presented only a minor degree of opening (Figure S4). In conclusion, the data from the ligand-independent analysis of the pathways in ERRs indicate pw-I, pw-II, pw-IIIa, and pw-IIIb as possible access or egress routes for ligands.

**Pathways I, II, III and V are favored by ERR ligands.** After characterizing the pathways independent of a particular ligand, we aimed to study the egress process of ligands from ERRs using metadynamics simulations. Both ERR$\alpha$ and ERR$\beta$ ligands dissociated through either pw-II or pw-V (Tables 2 and S3-S4). These pathways are separated by the N-terminus of H3 and H6 and, dependent on the conformational state of the receptor, can merge and form one large pathway. As the previous ligand-independent analysis suggested pw-II as the most feasible pathway for ERR$\alpha$, the distribution of favored pathways in the metadynamics simulations provide additional evidence thereof. The list of bottleneck residues, obtained from the previous analysis, allowed us to verify if they engaged in interactions with the ligand while it was leaving the binding site (Figure 2B-F). Interestingly, several bottleneck residues along pw-II, either hydrophobic or aromatic, interacted with the ligand in ERR$\alpha$ and ERR$\beta$ to a high degree during the translocation. Especially the presence of aromatic amino acids at bottlenecks is a well-known phenomenon [15, 32]. In ERR$\alpha$, we were not able to detect a classical gate consisting of two phenylalanine residues, but rather a cluster of hydrophobic and aromatic residues consisting of L324, L398, H494, and F495 (Figure 3A). Based on its topology, this gate likely acts in a so-called swinging door fashion [15]. As already discussed earlier, pw-V could be identified in ERR$\alpha$ and ERR$\beta$ in the ligand-independent procedures, although the opening was narrow. Thus, the ligands bound to these two receptors shared a preference for the translocation through pw-V besides pw-II.

**Table 2** Preferred pathways during metadynamics simulations.

| System | pw-I | pw-II | pw-IIIa | pw-IIIb | pw-V | Outcome[a] | p($\Delta$T)[b] | p(P$_{max}$)[b] |
|---|---|---|---|---|---|---|---|---|
| ERR$\alpha$ | 0 | 5 | 0 | 0 | 5 | pw-II | 0.001 | 0.691 |
| ERR$\beta$ | 0 | 5 | 1 | 0 | 4 | pw-II | 0.171 | 0.918 |
| ERR$\gamma$ ago | 0 | 0 | 6 | 4 | 0 | pw-IIIb | 0.850 | 0.815 |
| ERR$\gamma$ antago | 10 | 0 | 0 | 0 | 0 | pw-I | n/a | n/a |

[a] The mode of action also represents the conformational state of the receptors. [b] Homology model based on ERR$\gamma$ crystal structure. [c] Ligand orientation obtained by docking.

In ERR$\gamma$ bound to an agonist, pw-IIIa and pw-IIIb were the egress routes chosen by the ligand. Interestingly, both of these pathways presented the largest bottleneck radii in this complex. As previously mentioned, these two pathways are located in close vicinity to each other indicating this specific region among H3, the $\beta$-hairpin, and the H1-H3 loop to be of particular importance. Indeed these pathways seem to be preferred if the receptor is present in an agonistic conformation. In the antagonistic ERR$\gamma$ complex, on the other hand, the ligand left the binding site only through pw-I. This complex presented the highest degree of opening (statistically significant) in the previous analysis, establishing additional consensus by the two different methods (Tables S1 and S2). In the ERR$\gamma$-agonist complex, there was an apparently high occurrence of charged residues such as E275 and R316 at the bottleneck of both pw-IIIa and pw-IIIb. After inspecting the respective structures, we could identify a salt bridge at a distance of approximately 4.5 Å between these two residues (Figure 3B). As these two residues were among the most abundant bottleneck residues along with L309, V313, and Y326, the gate formed by them is likely involved in controlling ligand translocation. The fact that the ligand frequently interacted with R316 suggested that it might competitively weaken the ionic interaction at the gate in order to pass it. Visual inspection of several trajectories indicated that this was facilitated by the phenolate moiety of the ligand.

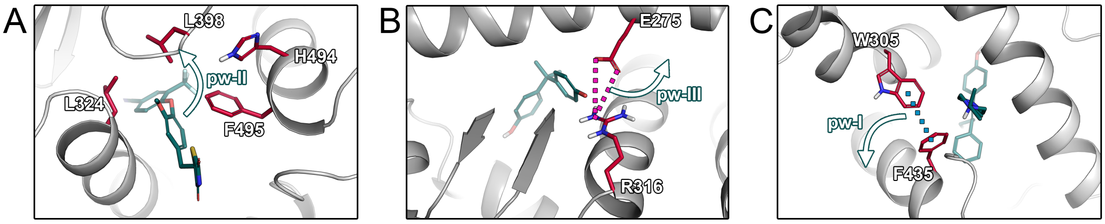


**Figure 3** Gating residues. Molecular gates in (A) ERR$\alpha$, (B) ERR$\gamma$ bound to an agonist, and (C) ERR$\gamma$ bound to an antagonist. Ionic interactions and $\pi$-$\pi$ contacts are highlighted along with the direction of the pathway.

When ERR$\gamma$ was bound to an antagonist, there were several highly consistent ligand-protein interactions among the ten replica simulations including ones with A272, L309, Y326 and F435, of which only the latter residue was involved in forming the bottleneck of pw-I. Visual inspection of the trajectory revealed F435 to form a gate with W305 (Figure 3C), which was involved in ligand-protein interactions in eight of the replica simulations. To conclude, we could identify three molecular gates in ERR$\alpha$ and ERR$\gamma$ based on our analysis of ligand-protein interactions during the egress process and bottleneck residues obtained from the pathway analysis.

As the metadynamics protocol introduces a biasing potential in order to sample rare molecular events such as ligand translocation, the methodology allows to compute the free energy landscape of the respective system based on the sum of Gaussian potentials deposited [20, 21]. As we selected a single CV describing the distance of the ligand and the binding site, we obtained a one-dimensional free energy surface. Based on the biasing potential that was accumulated to sample the ligand translocation process, we computed the maximal cumulative biasing potential deposited during ligand egress (Figure 4A, Tables S5 and S6). Additionally, we documented the time needed for the ligand to egress from the binding site (Figure 4B, Tables S7 and S8). Both the maximal potential and simulation time of ligand dissociation were correlated with residence times of ligands in previous work [33, 34]. The authors described the applicability of both readouts from multiple replica simulations even for a non-converged free energy landscape. Thus, our analysis allowed us to estimate which pathway might be more favorable for ligand translocation (Table 2). In ERR$\alpha$ and ERR$\beta$, pw-II presented the lowest maximal potential among the ten replicas. However, ERR$\beta$ also presented one trajectory, in which the translocation through either pw-II was comparatively unfavorable, suggesting both pathways to be feasible routes for the ligand in respect to the maximal potential. This was also reflected in the respective p-values (Table 2), which did not present statistically significant differences between their potentials in ERR$\alpha$ or ERR$\beta$. However, the simulation times observed in ERR$\alpha$ displayed a statistically significant preference for pw-II. Together with the maximal potential and the bottleneck radii, the results suggest pw-II to be the major route in ERR$\alpha$. In the agonist-bound complex of ERR$\gamma$, the data indicated pw-IIIb to be slightly preferred, although the differences to pw-IIIa were not statistically significant regarding both maximal potential and simulation time. These two pathways are only separated by a highly flexible region without a clearly defined secondary structure. Therefore, the results, including the ligand-independent analysis, suggest both pw-IIIa and pw-IIIb to be feasible egress routes in the agonistic conformation of ERR$\gamma$. As only pw-I was sampled in ERR$\gamma$ bound to an antagonist, the data did not allow to compare different pathways. In comparison to the remaining receptors, we observed the highest maximal potentials in this system. Potentially, pw-I might be the only feasible egress route in this particular complex. The different pathways selected by the ligands between the two ERR$\gamma$ complexes was most likely caused by the ligand-associated conformational changes in H12 as pw-I leads through the gap among H10, H11, and H12. In the agonist conformation, this region was less open as indicated by the bottleneck radii (Figure 2A) due to the tight packing of H12.

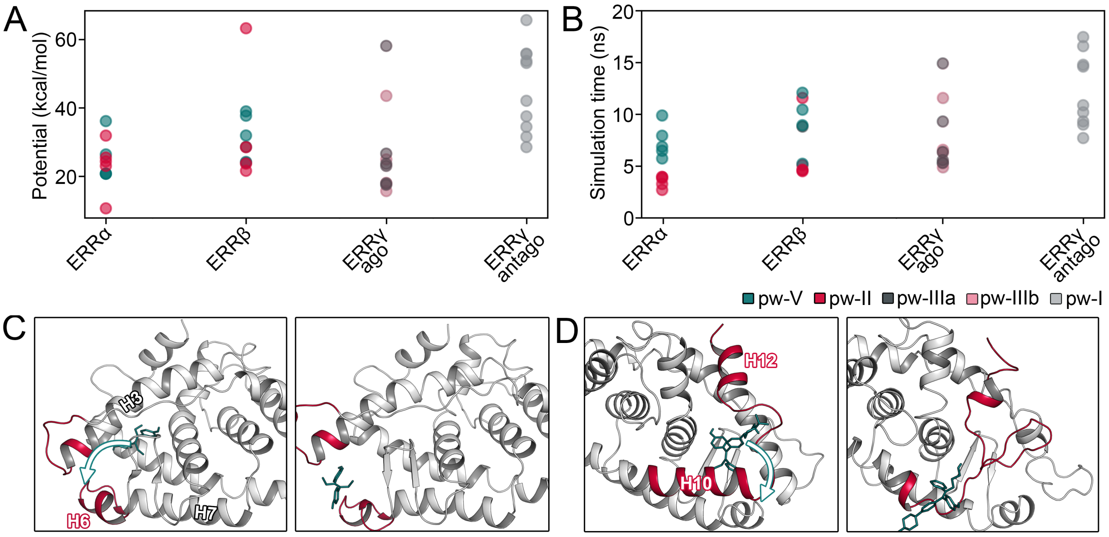


**Figure 4** Pathway preference and structural adaptation. (A) Maximal cumulative biasing potential observed in all simulations. The data points of different pathways are presented in individual colors. (B) Simulation times until the ligand dissocatied from the binding pocket. The data points of different pathways are presented in individual colors. (C) Ligand-induced conformational adaptation of H3 and H6 in ERR$\beta$ before (left) and after (right) the ligand passed through pw-II. (D) Ligand-induced conformational adaptation of H3 and H6 in ERR$\beta$ before (left) and after (right) the translocation of the ligand through pw-V. (E) Ligand-induced conformational adaptation of H10, H11, and H12 before (left) and after (right) the ligand translocated through pw-I in ERR$\gamma$ bound to an antagonist.

**ERRs structurally adapt during ligand translocation.** Pathways to buried binding pockets are often only accessible after conformational changes of the protein, which are often provoked by the ligand in an induced-fit mechanism. Such adaptations can occur at the level of protein side chains, as described above, as well as based on larger changes of secondary structural elements [15, 32]. Here, we determined the root-mean square fluctuation (RMSF) between the input structure and the part of the trajectory during which the ligand egressed from the binding site augmented by the visual inspection of the trajectories. To highlight the changes, we compared the RMSF between the unbiased and metadynamics simulations.

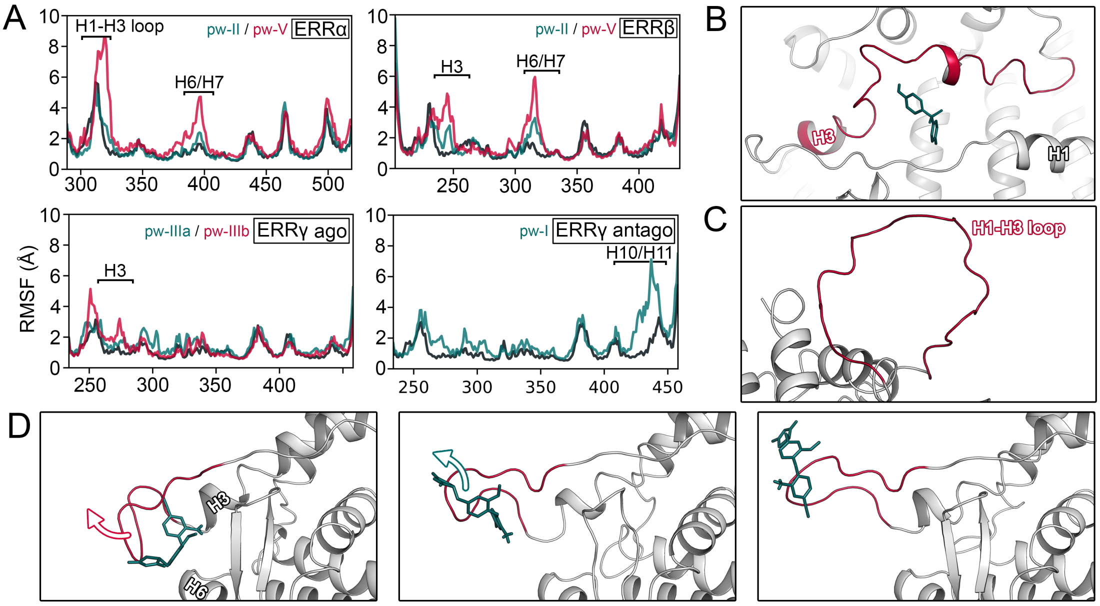


**Figure 5** Structural adaptation of the protein during ligand egress. (A) RMSF diagrams of conventional MD (black, average of three replicas) and metadynamics simulations (red or blue, maximum of respective replicas) of all studied receptor systems. Regions with substantial differences were indicated in the diagrams. (B) Distortion of H3 during egress through pw-IIIb in ERR$\gamma$ bound to an antagonist. (C) Highly extended H1-H3 loop in ERR$\alpha$. (D) Transition of H1-H3 loop together with the ligand in ERR$\alpha$.

In ERR$\alpha$, the RMSF indicated increased fluctuations of the H1-H3 loop, H6, and H7 (Figure 5A). Indeed, the visualization of the trajectories revealed considerable structural adaptation of the H1-H3 loop (Figures 5C and 5D). In one replica simulation, there was a large-scale change with a dissociation of the H1-H3 loop from the globular part of the protein as well as partial unwinding of H3 (Figure 5C). In all cases, the described changes led to the fusion of pw-II and pw-V as discussed in the previous sections. This explains the distribution of pw-II and pw-V observed in the metadynamics trajectories for ERR$\alpha$, as the spatial distinction between these two pathways became challenging and highly dependent on the individual trajectory if they were merged. Moreover, we observed an intriguing interaction between the ligand and the H1-H3 loop, during which the loop transitioned together with the ligand moving away from the protein in three replica simulations (Figure 5D). Visually, this process resembled the ligand being steered away from the protein. In ERR$\beta$, conformational adaptation of the protein was more subtle when the ligand translocated through pw-V (Figure 4C).

While H3 remained intact throughout the simulation, there was a slight adaptation of H6. Remarkably, the highlighted simulation (replica 4) also presented a low maximal potential, indicating this route to be favorable. In ERR$\gamma$ bound to an agonist, we observed a deformation of H3 in one simulation (Figure 5B), which also displayed the lowest maximal potential among the ten replicas. This was visible in the RMSF spike around residue 275 (Figure 5A). In the remaining simulations, the flexible unstructured region between H1 and H3 presented changes, probably associated with less biasing potential needed for ligand translocation. As the ligand preferred pw-I in the antagonistic conformation of ERR$\gamma$, the associated changes were different than in the agonistic complex. In all simulations, we could observe a restructuring of the terminal region of H10, H11, and H12, as also indicated by the large changes in the RMSF comparison to the conventional simulations. In conclusion, the receptors presented various structural adaptations along the pathways selected by the ligand, and the translocation was more favorable if the adaptations were smaller.

## Materials and Methods

**Model building and homology modeling.** We selected crystal structures in an antagonistic conformation for both ERR$\alpha$ and ERR$\gamma$ based on the resolution, bound ligands, completeness, and correctness of the sequence. Only for ERR$\gamma$, there was a structure available with an agonist bound to the LBP (Table 1). While there were multiple crystal structures available for ERR$\alpha$ and ERR$\gamma$, there were none for ERR$\beta$ when we started to work on the project. However, an agonist-bound structure (PDB ID: 6LIT) was published later on, allowing us to validate our modeling procedures [25]. ERR$\beta$ presents an acceptable sequence identity of 79.2% to ERR$\gamma$ (Figure S1). Thus, a homology model was generated using the SWISS-MODEL web server based on an antagonistic ERR$\gamma$ crystal structure (PDB ID: 5YSO) [26]. Retrospectively, we compared the ERR$\beta$ model to the novel crystal structure, after removing the missing N-terminal region from the model as well as the C-terminal region of both proteins. The C-terminal region was removed due to the antagonist-induced conformational changes [25] and the similarity between the structures (backbone and heavy atoms) was evaluated using the rmsd.py routine that comes with the Maestro Small-Molecule Drug Discovery Suite [35]. All crystal structures as well as the homology model were pre-processed using the Protein Preparation Wizard [36] within Maestro. For multimeric crystal structures, only chain A was retained for further procedures. We added hydrogen atoms, assigned bond orders, predicted protonation states at pH 7.4, and completed missing loops and side chains with the Prime routine. We used an in-house python routine to check if there were any residues missing or mutated in the respective structures, and if present, we resolved them in the 3D Builder in Maestro. Cocrystallized water molecules were retained, while ions and cosolvents were removed. An acetate cap was added to the N-terminus due to downstream connection with the DNA-binding domain of the receptors, while the C-terminus was modeled as free carboxylic acid. Next, the hydrogen bonding network was refined and the structures was subjected to a restrained minimization using the OPLS_2005 force field and the maximum heavy-atom displacement limit of 0.3 Å.

**Molecular docking and validation.** To obtain an antagonist-bound structure of ERR$\beta$, we docked the known inhibitor diethylstilbestrol [37] to its orthosteric binding site using the smina docking protocol [29]. Beforehand, we verified the capability of the docking protocol to reproduce crystallographic binding modes by re-docking cocrystallized ligands of all ERRs. The heavy atom RMSD of the obtained poses was determined using the Superposition panel within Maestro. Further, we compared the obtained binding mode to the cocrystallized pose of the native ligand within the template structure. Due to the antagonist-induced conformational changes between pairs of the selected structures, cross-docking was not considered.

**Conventional MD and metadynamics simulations.** All structures in this work were placed in orthorhombic periodic boundary systems solvated with TIP4P water molecules with counter-ions to neutralize the systems using the Maestro System Builder panel. The MD simulations were conducted using the Desmond (v2019-1) simulation engine [38] with the OPLS_2005 force field in an NPT ensemble at a temperature of 310 K maintained by the Nose–Hoover thermostat and atmospheric pressure regulated by the Martyna–Tobias–Klein barostat, both with a relaxation time of 2.0 ps. While short-range interactions were cut off at 9 Å, long-range interactions were treated with the u-series algorithm [39]. The M-SHAKE algorithm was used to constrain bonds to hydrogen atoms and the time step of the RESPA integrator was set to 2.0 fs. For the 600 ns conventional simulations conducted in triplicates, snapshots with atomic coordinates were collected at an interval of 60 ps. To ensure a unique course of the replica simulations, the random seed to compute the initial velocities of the simulations was set to either 2007, 3007, or 4007 respectively.

The metadynamics simulations were also conducted using the Desmond simulation engine. The collective variables for the systems were defined as the distance between the ligand and its binding site. To compute an approximate center of the binding site, three residues in its vicinity were selected based on the similarity of their mass center to the mass center of the ligand (Table S9). We set a wall for the CV of 40 Å and left the height of the Gaussian at 0.03 kcal/mol as well as the width of 0.05 Å on default. The simulations were conducted with ten replicas of 50 ns with atomic coordinates deposited at an interval of 5 ps. To ensure a unique course of the replica simulations the random seeds for the initial velocities were set to 2000, 2007, 2507, 3000, 3007, 3507, 4000, 4007, 5007, and 6007 respectively. The remaining parameters were left as described above for the conventional MD simulations.

**Evaluation of the MD trajectories.** Metrics such as RMSD, RMSF, and ligand-protein interactions were computed in the Simulation Interaction Diagram panel within Maestro for both conventional and metadynamics simulations. To compute the ligand pathways within the structures, we used the CAVER (v3.0) command-line program [30]. The coordinates of the starting points for this procedure were obtained based on the respective cocrystallized ligands. For the calculation, we extracted 500 frames of the last 120 ns of the conventional MD trajectories at an interval of 240 ps while only retaining the protein structure without solvent or ligands. Besides a clustering threshold of 4.0, we retained default settings for the CAVER calculations. The pathways were visually deduced from the obtained clusters according to our recent review article [11]. From the computed data, we determined the average bottleneck radius and the corresponding bottleneck residues of the pathways. To rate the importance of a bottleneck residue, we computed the total number of occurrences of a residue for each pathway per receptor system by concatenating the output of the three replicas. Statistical significance was evaluated using ttest_ind_from_stats routine in the python-scipy module using the concatenated output of bottleneck radii with the number of objects defined as the number of frames. For the metadynamics simulations, the deposited potentials were directly obtained from the simulation workflow. The maximal potential was derived using the metadynminer toolkit based on R scripting language [40]. Statistical significance was evaluated as described above with the number of replicas defined as number of objects for the average values. The trajectories were visualized to determine the time point of ligand egress and then truncated to only represent the time span until the ligand left the binding site using the trj_parch.py routine that comes with Maestro. Further, the pathways used by the ligand to egress the binding site were obtained from visualization. To quantify the conformational changes induced by ligand egress, we computed the RMSF for the truncated metadynamics simulations, representing only the egress process, and the conventional simulations. To adjust for the increased flexibility imposed by the larger time scale of the conventional MD simulations and the increased number of frames compared to the truncated metadynamics trajectories, we also truncated the conventional simulations for the RMSF calculations. The number of retained frames from the conventional MD trajectories was determined by the average number of frames the ligand took to egress from the binding site in the ten metadynamics replicas. Ultimately, to determine the RMSF, the first frame of the simulations was used as reference structure. As mentioned above, the ligand-protein interactions were computed using the Simulation Interaction Diagram panel in Maestro. For each truncated replica simulation of the metadynamics runs, we documented interactions if they occurred in at least 10% of the MD frames. We considered hydrogen bonding, hydrophobic, ionic, $\pi$-cation, and $\pi$-$\pi$ interactions for this analysis.

## Conclusion

ERRs have been recently highlighted as drug targets for diabetes, cardiovascular diseases, and osteoporosis. As they harbor a buried binding site, pathways within the receptor are used by ligands to translocate to and from it. As these pathways are not visible in static crystal structures, we conducted conventional MD and metadynamics simulations to elucidate their existence and estimate the most favorable routes for the ligand egress. The analysis independent of a translocating ligand revealed pw-I, pw-II, pw-IIIa, and pw-IIIb, depending on the present ligand-protein complex and starting conformation, to be open to a considerable degree. In the subsequent metadynamics trajectories we could confirm the previous observations, but we additionally observed pw-V in ERR$\alpha$ and ERR$\beta$ to be of relevance. In ERR$\gamma$, the conformational state of the receptor influenced the preference of the ligand. If the structure was bound to an agonist, the ligand translocated through pw-IIIa or pw-IIIb, while the studied antagonist preferred pw-I. In pw-IIIa and pw-IIIb, we could detect a gating mechanism of the two aromatic residues W305 and F435 held together by $\pi$-interactions. Further, pw-I seemed to be regulated by an ionic interaction between E275 and R316. During the translocation through pw-II, ligands had to pass through a gate consisting of multiple aromatic and hydrophobic side chains (L324, L394, H494, F495). By comparing the RMSF between conventional MD and metadynamics simulations, we could reproducibly deduce various conformational changes associated with the translocation offering additional mechanistic insight into the process. Further on, due to the fact that we could retrospectively analyze the performance of our homology model because of a newly released crystal structure, we could show that such procedures perform well for NRs. Overall, due to the emerging potential of ERRs as drug targets, our work offers insights into the functionality and topology of their ligand pathways, which can ultimately be used to guide the rational design of selective modulators.

## References

[1] Madhulika Tripathi, Paul Michael Yen, and Brijesh Kumar Singh. Estrogen-related receptor alpha: An under-appreciated potential target for the treatment of metabolic diseases. *International Journal of Molecular Sciences*, 21(5), 2020.

[2] B. Horard and J. M. Vanacker. Estrogen receptor-related receptors: Orphan receptors desperately seeking a ligand. *Journal of Molecular Endocrinology*, 31(3):349–357, 2003.

[3] Kenji Saito and Huxing Cui. Emerging roles of estrogen-related receptors in the brain: Potential interactions with estrogen signaling. *International Journal of Molecular Sciences*, 19(4), 2018.

[4] Janice M. Huss, Wojciech G. Garbacz, and Wen Xie. Constitutive activities of estrogen-related receptors: Transcriptional regulation of metabolism by the ERR pathways in health and disease. *Biochimica et Biophysica Acta - Molecular Basis of Disease*, 1852(9):1912–1927, 2015.

[5] Marlène Gallet and Jean Marc Vanacker. ERR receptors as potential targets in osteoporosis. *Trends in Endocrinology and Metabolism*, 21(10):637–641, 2010.

[6] Sharon Cresci, Janice M. Huss, Amber L. Beitelshees, Philip G. Jones, Matt R. Minton, Gerald W. Dorn, Daniel P. Kelly, John A. Spertus, and Howard L. McLeod. A ppar$\alpha$ promoter variant impairs ERR-dependent transactivation and decreases mortality after acute coronary ischemia in patients with diabetes. *PLoS ONE*, 5(9):1–9, 2010.

[7] Wei Wei, Adam G. Schwaid, Xueqian Wang, Xunde Wang, Shili Chen, Qian Chu, Alan Saghatelian, and Yihong Wan. Ligand activation of ERR$\alpha$ by cholesterol mediates statin and bisphosphonate effects. *Cell Metabolism*, 23(3):479–491, 2016.

[8] Caitlin Lynch, Jinghua Zhao, Srilatha Sakamuru, Li Zhang, Ruili Huang, Kristine L. Witt, B. Alex Merrick, Christina T. Teng, and Menghang Xia. Identification of compounds that inhibit estrogen-related receptor alpha signaling using high-throughput screening assays. *Molecules*, 24(5), 2019.

[9] Masatomo Suetsugi, Leila Su, Kimberly Karlsberg, Yate Ching Yuan, and Shiuan Chen. Flavone and Isoflavone Phytoestrogens Are Agonists of Estrogen-Related Receptors. *Molecular Cancer Research*, 1(13):981–991, 2003.

[10] Michal Pawlak, Philippe Lefebvre, and Bart Staels. General molecular biology and architecture of nuclear receptors. *Current topics in medicinal chemistry*, 12(6):486–504, 2012.

[11] André Fischer and Martin Smieško. Ligand Pathways in Nuclear Receptors. *Journal of Chemical Information and Modeling*, 59(7):3100–3109, 7 2019.

[12] Jie Shen, Weihua Li, Guixia Liu, Yun Tang, and Hualiang Jiang. Computational insights into the mechanism of ligand unbinding and selectivity of estrogen receptors. *Journal of Physical Chemistry B*, 113(30):10436–10444, 2009.

[13] Anna Maria Capelli, Agostino Bruno, Antonio Entrena Guadix, and Gabriele Costantino. Unbinding pathways from the glucocorticoid receptor shed light on the reduced sensitivity of glucocorticoid ligands to a naturally occurring, clinically relevant mutant receptor. *Journal of Medicinal Chemistry*, 56(17):7003–7014, 2013.

[14] Markéta Paloncýova, Veronika Navrátilova, Karel Berka, Alessandro Laio, and Michal Otyepka. Role of Enzyme Flexibility in Ligand Access and Egress to Active Site: Bias-Exchange Metadynamics Study of 1,3,7-Trimethyluric Acid in Cytochrome P450 3A4. *Journal of Chemical Theory and Computation*, 12(4):2101–2109, 2016.

[15] Artur Gora, Jan Brezovsky, and Jiri Damborsky. Gates of enzymes. *Chemical Reviews*, 113(8):5871–5923, 2013.

[16] Laura J. Kingsley and Markus A. Lill. Including ligand-induced protein flexibility into protein tunnel prediction. *Journal of Computational Chemistry*, 35(24):1748–1756, 2014.

[17] Fuxing Li, Xianqiang Sun, Yingchun Cai, Defang Fan, Weihua Li, Yun Tang, and Guixia Liu. Computational investigation of the interaction mechanism between the estrogen related receptor $\alpha$ and its agonists. *RSC Advances*, 6(96):94119–94127, 2016.

[18] Dongping Li, Yingchun Cai, Dan Teng, Weihua Li, Yun Tang, and Guixia Liu. Computational insights into the interaction mechanisms of estrogen-related receptor alpha with endogenous ligand cholesterol. *Chemical Biology and Drug Design*, 94(1):1316–1329, 2019.

[19] Dongping Li, Yingchun Cai, Dan Teng, Zengrui Wu, Weihua Li, Yun Tang, and Guixia Liu. Insights into the interaction mechanisms of estrogen-related receptor alpha (ERR$\alpha$) with ligands via molecular dynamics simulations. *Journal of Biomolecular Structure and Dynamics*, 38(13):3867–3878, 2020.

[20] Vittorio Limongelli, Massimiliano Bonomi, and Michele Parrinello. Funnel metadynamics as accurate binding free-energy method. *Proceedings of the National Academy of Sciences of the United States of America*, 110(16):6358–6363, 2013.

[21] Giovanni Bussi, Alessandro Laio, and Pratyush Tiwary. Metadynamics: A Unified Framework for Accelerating Rare Events and Sampling Thermodynamics and Kinetics BT - Handbook of Materials Modeling : Methods: Theory and Modeling. pages 1–31. Springer International Publishing, Cham, 2018. ISBN 978-3-319-42913-7.

[22] André Fischer, Gabriela Frehner, Markus A Lill, and Martin Smieško. Conformational Changes of Thyroid Receptors in Response to Antagonists. *Journal of Chemical Information and Modeling*, 2021.

[23] A K Shiau, D Barstad, P M Loria, L Cheng, P J Kushner, D A Agard, and G L Greene. The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen. *Cell*, 95(7):927–937, 12 1998.

[24] Marta C. Abad, Hossein Askari, John O'Neill, Alexandra L. Klinger, Cynthia Milligan, Frank Lewandowski, Barry Springer, John Spurlino, and Dionisios Rentzeperis. Structural determination of estrogen-related receptor $\gamma$ in the presence of phenol derivative compounds. *Journal of Steroid Biochemistry and Molecular Biology*, 108(1-2):44–54, 2008.

[25] Benqiang Yao, Shuchi Zhang, Yijuan Wei, Siyu Tian, Zhou Lu, Lihua Jin, Ying He, Wen Xie, and Yong Li. Structural Insights into the Specificity of Ligand Binding and Coactivator Assembly by Estrogen-Related Receptor $\beta$. *Journal of Molecular Biology*, 432(19): 5460–5472, 2020.

[26] Marco Biasini, Stefan Bienert, Andrew Waterhouse, Konstantin Arnold, Gabriel Studer, Tobias Schmidt, Florian Kiefer, Tiziano Gallo Cassarino, Martino Bertoni, Lorenza Bordoli, Torsten Schwede, Tiziano Gallo Cassarino, Martino Bertoni, Lorenza Bordoli, and

Torsten Schwede. SWISS-MODEL: Modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Research*, 42(Web Server issue):252–8, 7 2014.

[27] Pascal Benkert, Marco Biasini, and Torsten Schwede. Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics*, 27(3):343–350, 2011.

[28] Gabriel Studer, Christine Rempfer, Andrew M Waterhouse, Rafal Gumienny, Juergen Haas, and Torsten Schwede. QMEANDisCo—distance constraints applied on model quality estimation. *Bioinformatics*, 36(6):1765–1771, 3 2020.

[29] David Ryan Koes, Matthew P. Baumgartner, and Carlos J. Camacho. Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. *Journal of Chemical Information and Modeling*, 53(8):1893–1904, 2013.

[30] Eva Chovancova, Antonin Pavelka, Petr Benes, Ondrej Strnad, Jan Brezovsky, Barbora Kozlikova, Artur Gora, Vilem Sustr, Martin Klvana, Petr Medek, Lada Biedermannova, Jiri Sochor, and Jiri Damborsky. CAVER 3.0: A Tool for the Analysis of Transport Pathways in Dynamic Protein Structures. *PLoS Computational Biology*, 8(10):23–30, 2012.

[31] Christoph Grebner, Daniel Lecina, Victor Gil, Johan Ulander, Pia Hansson, Anita Dellsen, Christian Tyrchan, Karl Edman, Anders Hogner, and Victor Guallar. Exploring Binding Mechanisms in Nuclear Hormone Receptors by Monte Carlo and X-ray-derived Motions. *Biophysical Journal*, 112(6):1147–1156, 2017.

[32] André Fischer and Martin Smieško. Spontaneous Ligand Access Events to Membrane-Bound Cytochrome P450 2D6 Sampled at Atomic Resolution. *Scientific Reports*, 9(1): 16411, 2019.

[33] Andrea Bortolato, Francesca Deflorian, Dahlia R Weiss, and Jonathan S Mason. Decoding the Role of Water Dynamics in Ligand–Protein Unbinding: CRF1R as a Test Case. *Journal of Chemical Information and Modeling*, 55(9):1857–1866, 9 2015.

[34] Huiyong Sun, Youyong Li, Mingyun Shen, Dan Li, Yu Kang, and Tingjun Hou. Characterizing Drug–Target Residence Time with Metadynamics: How To Achieve Dissociation Rate Efficiently without Losing Accuracy against Time-Consuming Approaches. *Journal of Chemical Information and Modeling*, 57(8):1895–1906, 8 2017.

[35] Schrodinger LCC. Maestro Small-Molecular Drug Discovery Suite 2019-4. 2019.

[36] G. Madhavi Sastry, Matvey Adzhigirey, Tyler Day, Ramakrishna Annabhimoju, and Woody Sherman. Protein and ligand preparation: Parameters, protocols, and influence on virtual screening enrichments. *Journal of Computer-Aided Molecular Design*, 27(3): 221–234, 2013.

[37] Shailaja D Divekar, Deanna M Tiek, Aileen Fernandez, and Rebecca B Riggins. Estrogen-related Receptor $\beta$ (ERR$\beta$) – Renaissance Receptor or Receptor Renaissance? *Nuclear Receptor Signaling*, 14(1):nrs.14002, 1 2016.

[38] Kevin Bowers, Edmond Chow, Huafeng Xu, Ron Dror, Michael Eastwood, Brent Gregersen, John Klepeis, Istvan Kolossvary, Mark Moraes, Federico Sacerdoti, John Salmon, Yibing Shan, and David Shaw. Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters. *ACM/IEEE SC 2006 Conference (SC'06)*, (November):43, 2006.

[39] David E. Shaw, J. P. Grossman, Joseph A. Bank, Brannon Batson, J. Adam Butts, Jack C. Chao, Martin M. Deneroff, Ron O. Dror, Amos Even, Christopher H. Fenton, Anthony Forte, Joseph Gagliardo, Gennette Gill, Brian Greskamp, C. Richard Ho, Douglas J. Ierardi, Lev Iserovich, Jeffrey S. Kuskin, Richard H. Larson, Timothy Layman, Li Siang Lee, Adam K. Lerer, Chester Li, Daniel Killebrew, Kenneth M. Mackenzie, Shark Yeuk Hai Mok, Mark A. Moraes, Rolf Mueller, Lawrence J. Nociolo, Jon L. Peticolas, Terry Quan, Daniel Ramot, John K. Salmon, Daniele P. Scarpazza, U. Ben Schafer, Naseer Siddique, Christopher W. Snyder, Jochen Spengler, Ping Tak Peter Tang, Michael Theobald, Horia Toma, Brian Towles, Benjamin Vitale, Stanley C. Wang, and Cliff Young. Anton 2: Raising the Bar for Performance and Programmability in a Special-Purpose Molecular Dynamics Supercomputer. *International Conference for High Performance Computing, Networking, Storage and Analysis, SC*, 2015-Janua(January):41–53, 2014.

[40] Dalibor Trapl and Vojtěch Spiwok. Analysis of the Results of Metadynamics Simulations by metadynminer and metadynminer3d. *arXiv.org*, XX:1–11, 2020.

## 6.1 Supporting Information

## Supporting Results and Discussion

### Model building and validation

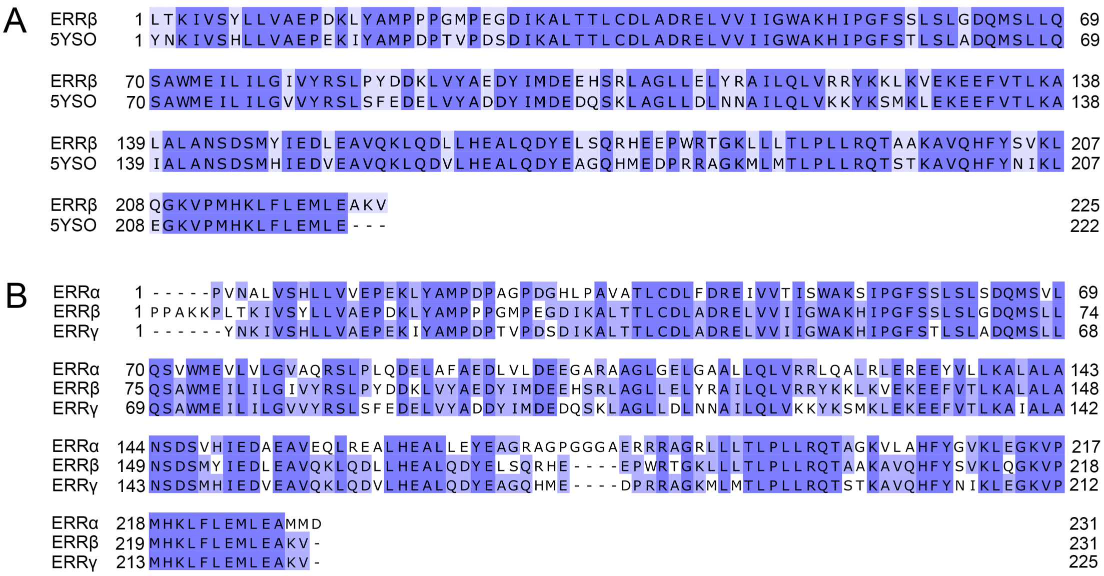

A
ERRβ 1 LTKIVSYLLVAEPDKLYAMPPPGMPEGDIKALTTLCDLADRELVVIIGWAKHIPGFSSLSLGDQMSLLQ 69
5YSO 1 YNKIVSHLLVAEPEKIYAMPDPTVPDSDIKALTTLCDLADRELVVIIGWAKHIPGFSTLSLADQMSLLQ 69
ERRβ 70 SAWMEILILGIVYRSLPYDDKLVYAEDYIMDEEHSRLAGLLELYRAILQLVRRYKKLKVEKEEFVTLKA 138
5YSO 70 SAWMEILILGVVYRSLSFEDELVYADDYIMDEDQSKLAGLLDLNNAILQLVKKYKSMKLEKEEFVTLKA 138
ERRβ 139 LALANSDSMYIEDLEAVQKLQDLLHEALQDYELSQRHEEPWRTGKLLLTLPLLRQTAAKAVQHFYSVKL 207
5YSO 139 IALANSDSMHIEDVEAVQKLQDVLHEALQDYEAGQHMEDPRRAGKMLMTLPLLRQTSTKAVQHFYNIKL 207
ERRβ 208 QGKVPMHKLFLEMLEAKV 225
5YSO 208 EGKVPMHKLFLEMLE--- 222
B
ERRα 1 -----PVNALVSHLLVVEPEKLYAMPDPAGPDGHLPAVATLCDLFDREIVVTISWAKSIPGFSSLSLSDQMSVL 69
ERRβ 1 PPAKKPLTKIVSYLLVAEPDKLYAMPPPGMPEGDIKALTTLCDLADRELVVIIGWAKHIPGFSSLSLGDQMSLL 74
ERRγ 1 ------YNKIVSHLLVAEPEKIYAMPDPTVPDSDIKALTTLCDLADRELVVIIGWAKHIPGFSTLSLADQMSLL 68
ERRα 70 QSVWMEVLVLGVAQRSLPLQDELAFAEDLVLDEEGARAAGLGELGAALLQLVRRLQALRLEREEYVLLKALALA 143
ERRβ 75 QSAWMEILILGIVYRSLPYDDKLVYAEDYIMDEEHSRLAGLLELYRAILQLVRRYKKLKVEKEEFVTLKALALA 148
ERRγ 69 QSAWMEILILGVVYRSLSFEDELVYADDYIMDEDQSKLAGLLDLNNAILQLVKKYKSMKLEKEEFVTLKAIALA 142
ERRα 144 NSDSVHIEDAEAVEQLREALHEALLEYEAGRAGPGGGAERRRAGRLLLTLPLLRQTAGKVLAHFYGVKLEGKVP 217
ERRβ 149 NSDSMYIEDLEAVQKLQDLLHEALQDYELSQRHE----EPWRTGKLLLTLPLLRQTAAKAVQHFYSVKLQGKVP 218
ERRγ 143 NSDSMHIEDVEAVQKLQDVLHEALQDYEAGQHME----DPRRAGKMLMTLPLLRQTSTKAVQHFYNIKLEGKVP 212
ERRα 218 MHKLFLEMLEAMMD 231
ERRβ 219 MHKLFLEMLEAKV- 231
ERRγ 213 MHKLFLEMLEAKV- 225

**Figure S 1** Sequence alignments. (A) Sequence alignment of ERR$\beta$ to the sequence in the ERR$\gamma$ crystal structure (PDB ID: 5YSO). (B) Sequence alignment of all three ERRs.

### Different patterns of ligand pathways are present in ERRs.

**Table S 1** Statistics of bottleneck radii.

| System | Pathway | Mean | SD | Samples |
|---|---|---|---|---|
| ERR$\alpha$ | pw-I | 0.987 | 0.094 | 57 |
| ERR$\beta$ | pw-I | 1.005 | 0.108 | 250 |
| ERR$\gamma$ ago | pw-I | 0.951 | 0.027 | 9 |
| ERR$\gamma$ antago | pw-I | 1.204 | 0.332 | 505 |
| ERR$\alpha$ | pw-IIIa | 1.453 | 0.319 | 659 |
| ERR$\beta$ | pw-IIIa | 1.290 | 0.347 | 1288 |
| ERR$\gamma$ ago | pw-IIIa | 1.799 | 0.213 | 1460 |
| ERR$\gamma$ antago | pw-IIIa | 1.414 | 0.212 | 981 |

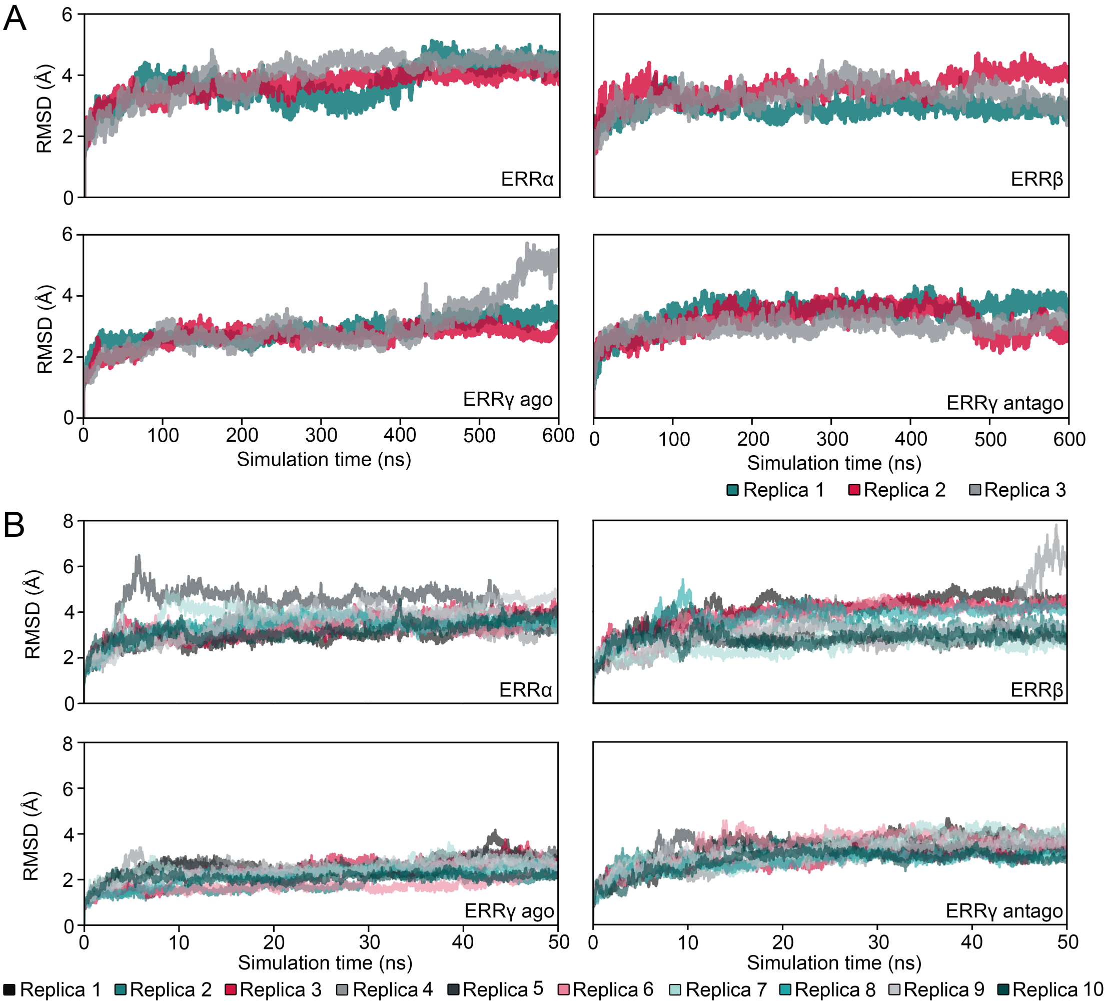


**Figure S 2** Backbone RMSD. (A) Backbone RMSD of conventional MD simulations. (B) Backbone RMSD of metadynamics simulations.

**Table S 2** Statistical evaluation of bottleneck radii.

| System 1 | System 2 | Pathway | Preference | Significance |
|---|---|---|---|---|
| ERR$\gamma$ antago | ERR$\alpha$ | pw-I | ERR$\gamma$ antago | yes |
| ERR$\gamma$ antago | ERR$\beta$ | pw-I | ERR$\gamma$ antago | yes |
| ERR$\gamma$ antago | ERR$\gamma$ ago | pw-I | ERR$\gamma$ antago | yes |
| ERR$\gamma$ ago | ERR$\alpha$ | pw-IIIa | ERR$\gamma$ ago | yes |
| ERR$\gamma$ ago | ERR$\beta$ | pw-IIIa | ERR$\gamma$ ago | yes |
| ERR$\gamma$ antago | ERR$\alpha$ | pw-IIIa | ERR$\gamma$ antago | yes |
| ERR$\gamma$ antago | ERR$\beta$ | pw-IIIa | ERR$\gamma$ antago | yes |

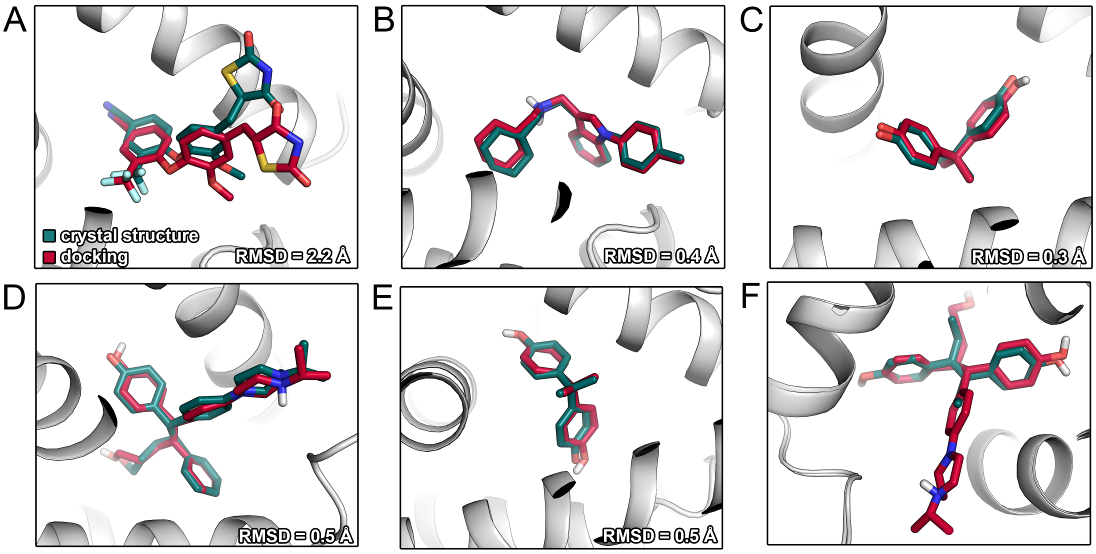


**Figure S 3** Docking pose of (A) of Q27455709 docked to ERR$\alpha$ aligned to the cocrystallized binding mode (PDB ID: 3K6P), (B) of CHEMBL478524 docked to ERR$\alpha$ aligned to to cocrystallized binding mode (PDB ID: 2PJL), (C) of bisphenol A docked to ERR$\gamma$ aligned to the cocrystallized binding mode (PDB ID: 2P7G), (D) of DN200434 docked to ERR$\gamma$ aligned to the cocrystallized binding mode (PDB ID: 5YSO), (E) of bisphenol A docked to ERR$\beta$ aligned to cocrystallized binding mode (PDB ID: 6LIT). The respective heavy atom RMSD values are indicated at the bottom of the figures. (F) Docking pose of diethylbestrol used for ERR$\beta$ simulations aligned to the cocrystallized ligand of the template structure (PDB ID: 5YSO).

**Pathways I, II, III and V are favored by ERR ligands.**

**Table S 3** Pathways selected by the ligand per replica simulation.

| **System** | **Replica 1** | **Replica 2** | **Replica 3** | **Replica 4** | **Replica 5** |
|---|---|---|---|---|---|
| ERR$\alpha$ | pw-V | pw-V | pw-V | pw-II | pw-V |
| ERR$\beta$ | pw-II | pw-II | pw-V | pw-V | pw-II |
| ERR$\gamma$ ago | pw-IIIb | pw-IIIa | pw-IIIb | pw-IIIb | pw-IIIa |
| ERR$\gamma$ antago | pw-I | pw-I | pw-I | pw-I | pw-I |

**Table S 4** Pathways selected by the ligand per replica simulation (continued).

| **System** | **Replica 6** | **Replica 7** | **Replica 8** | **Replica 9** | **Replica 10** |
|---|---|---|---|---|---|
| ERR$\alpha$ | pw-II | pw-V | pw-II | pw-II | pw-II |
| ERR$\beta$ | pw-IIIa | pw-II | pw-V | pw-V | pw-II |
| ERR$\gamma$ ago | pw-IIIa | pw-IIIa | pw-IIIa | pw-IIIa | pw-IIIb |
| ERR$\gamma$ antago | pw-I | pw-I | pw-I | pw-I | pw-I |

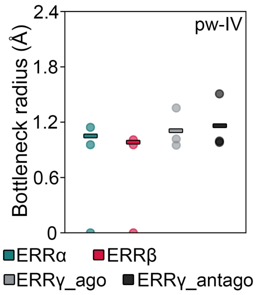


**Figure S 4** Bottleneck radii of pw-IV in all studied receptor systems. The average of the three replicas (null values not included in the average) is indicated by a horizontal line.

**Table S 5** Maximal potential until complete translocation.

| System | Replica 1 | Replica 2 | Replica 3 | Replica 4 | Replica 5 |
|---|---|---|---|---|---|
| ERR$\alpha$ | 20.75 | 26.38 | 20.74 | 25.45 | 20.73 |
| ERR$\beta$ | 23.94 | 63.26 | 38.97 | 24.17 | 23.70 |
| ERR$\gamma$ ago | 24.91 | 23.00 | 18.06 | 15.67 | 17.87 |
| ERR$\gamma$ antago | 53.71 | 55.70 | 37.50 | 34.44 | 42.05 |

**Table S 6** Maximal potential until complete translocation (continued).

| System | Replica 6 | Replica 7 | Replica 8 | Replica 9 | Replica 10 |
|---|---|---|---|---|---|
| ERR$\alpha$ | 24.33 | 36.11 | 23.00 | 31.89 | 10.59 |
| ERR$\beta$ | 28.57 | 21.58 | 31.91 | 37.70 | 28.46 |
| ERR$\gamma$ ago | 17.46 | 23.62 | 58.10 | 26.60 | 43.51 |
| ERR$\gamma$ antago | 53.13 | 28.50 | 55.85 | 65.65 | 31.54 |

**Table S 7** Times of complete translocation.

| System | Replica 1 | Replica 2 | Replica 3 | Replica 4 | Replica 5 |
|---|---|---|---|---|---|
| ERR$\alpha$ | 6.85 | 6.46 | 7.91 | 3.93 | 5.71 |
| ERR$\beta$ | 5.07 | 11.56 | 12.06 | 5.20 | 4.63 |
| ERR$\gamma$ ago | 6.55 | 5.26 | 6.31 | 4.87 | 5.30 |
| ERR$\gamma$ antago | 17.43 | 14.76 | 9.29 | 8.98 | 10.85 |

The simulation time after which the ligand completely translocated through the respective pathways. The values are given in nanoseconds (ns).

**Table S 8** Times of complete translocation (continued).

| System | Replica 6 | Replica 7 | Replica 8 | Replica 9 | Replica 10 |
|---|---|---|---|---|---|
| ERR$\alpha$ | 3.80 | 9.86 | 3.25 | 3.94 | 2.68 |
| ERR$\beta$ | 8.80 | 4.56 | 10.42 | 8.92 | 4.49 |
| ERR$\gamma$ ago | 5.54 | 6.33 | 14.87 | 9.28 | 11.56 |
| ERR$\gamma$ antago | 16.55 | 7.69 | 14.58 | 22.78 | 10.17 |

The simulation time after which the ligand completely translocated through the respective pathways. The values are given in nanoseconds (ns).

## Supporting Materials and Methods

### Conventional MD and metadynamics simulations

**Table S 9** Residues for collective variables.

| System | Residues |
|---|---|
| ERR$\alpha$ | V321, C325, M362 |
| ERR$\beta$ | C42, D43, Q199 |
| ERR$\gamma$ ago | I308, L309, L339 |
| ERR$\gamma$ antago | L271, A272, N437 |

Protein residues that were selected to define the CVs in the metadynamics simulations. The residues were selected to have a centroid close to the one of the binding site.

# CHAPTER 7

# Conformational Changes of Thyroid Receptors in Response to Antagonists

Conformational adaptation in response to ligand binding is one of the most complex phenomena observed in molecular recognition. The aim of this study was to elucidate conformational changes of thyroid receptors in response to antagonists. The interference with thyroid receptor signaling is relevant for the development of novel therapeutics as well as the estimation of the toxicity of environmental compounds such as pesticides. Among various structural changes, the ligand access pathways were altered dependent on the orthosteric ligand present in the structure, connecting this study to work highlighted in Chapters 2 and 3 on CYPs as well as the two previous chapters.

**Author contributions:** Conceptualization, A.F.; formal analysis, A.F., G.F.; writing and original draft preparation, A.F.; writing, review and editing, A.F., M.A., M.S.; visualization, A.F.; supervision, M.A., M.S.

*Based on the published research article:*

Fischer, A.; Frehner, G.; Lill, M. A.; Smieško, M. Conformational Changes of Thyroid Receptors in Response to Antagonists. *J. Chem. Inf. Model.* 2021, 61 (2), 1010-1019 .

## Abstract

Thyroid hormone receptors (TRs) play a critical role in human development, growth, and metabolism. Antagonists of TRs offer an attractive strategy to treat hyperthyroidism without the disadvantage of a delayed onset of drug action. While it is challenging to examine the atomistic behavior of TRs in a laboratory setting, computational methods such as molecular dynamics (MD) simulations have proven their value to elucidate ligand-induced conformational changes in nuclear receptors. Here, we performed MD simulations of TR$\alpha$ and TR$\beta$ complexed to their native ligand triiodothyronine (T3), as well as several antagonists. Based on the examination of 27 $\mu$s MD trajectories, we showed how binding of these compounds influences various structural features of the receptors including the helicity of helices 3 and 10, as well as the location of helix-12. Helices 3 and 12 are known to mediate coactivator association required for downstream signaling suggesting these changes to be the molecular basis for TR antagonism. A mechanistic analysis of the trajectories revealed an allosteric pathway between H3 and H12 responsible for the conformational adaptations. Even though a mechanistic understanding of conformational adaptations triggered by TR antagonists is important for the development of novel therapeutics, they were not previously examined in detail as it was done here.

## Introduction

Thyroid hormone receptors (TRs) belong to the superfamiliy of nuclear receptors (NRs) and are involved in a multitude of physiological and pathological processes in humans. In particular, they mediate the action of thyroid hormones, which play a critical role in development, growth, cardiac function, and metabolism leading to severe implications in the case of dysregulation [1, 2]. The isoforms TR$\alpha$ and TR$\beta$ present different tissue distribution as well as a distinct physiological role beyond their expression patterns. This was, for example, shown by the isoform-specific regulation of gene expression [3]. For example, it is thought that cardiovascular effects of thyroid hormone are mediated by TR$\alpha$, while TR$\beta$ is responsible for actions regarding metabolism [4]. TR antagonists are therapeutically beneficial in patients suffering from hyperthyroidism and the resulting thyrotoxicosis primarily caused by Graves disease, thyroiditis, or exogenous hormone intake. Hyperthyroidism is associated with tachycardia, heart failure, as well as skeletal muscle weakness and can have life-threatening outcomes [5]. On the other hand, adverse effects of various drugs and environmental chemicals are considered to be mediated by blocking TRs resulting in hypothyroidism with cardiac symptoms such as bradycardia and complications related to the central nervous system. Even though therapeutics against hyperthyroidism are available, they suffer from considerable disadvantages such as the highly delayed onset of synthesis inhibitors due to a remaining reserve of hormones in circulation and their long half-life. Thus, direct pharmacological intervention at the receptor level instead of thyroid hormone synthesis has appealing benefits [1, 2, 5, 6, 7, 8]. Interestingly, the cardiovascular adverse effects of the analgesic celecoxib were linked to TR$\beta$ antagonism [9]. In contrast, it was proposed that TRs might constitute a target for novel antiarrhytmics due to their influence on the cardiovascular system. However, despite their therapeutic potential in rapidly treating thyrotoxicosis, TR antagonists have not yet reached clinical application [2, 7]. In this regard, the understanding of the underlying structural mechanism of TR antagonism can be critical for the development of effective compounds, especially in a structure-based design setting as it was evidenced in previous discovery efforts [5]. Since crystal structures of TR$\alpha$ or TR$\beta$ bound to antagonists have not been published and experimental methods fail to provide atomistic detail, computational methods offer an attractive approach to acquire mechanistic insight into their structural adaptations in response to antagonists [1, 2, 7, 10, 11, 12, 13]. In previous work, the unliganded ligand binding domain (LBD) of TR$\beta$ was examined with 22 ns molecular dynamics (MD) simulations and the mechanism of TR antagonism was monitored by hydrogen/deuterium exchange followed by mass spectrometry. Based on these experiments, it was illustrated that helix-12 (H12) underwent conformational changes [7]. Interestingly, a recently published model based on a combination of quantitative structure-activity relationships and machine learning evidenced accurate distinguishing of TR agonists and antagonists [10]. Another study highlighted conformational adaptations induced by mutations at the activation function-2 (AF-2), involved in coactivator binding essential for downstream signaling (Figure 1A), and similarly reported rearrangements of H12 [14]. In other NRs, multiple studies have employed microsecond MD simulations and advanced sampling techniques to

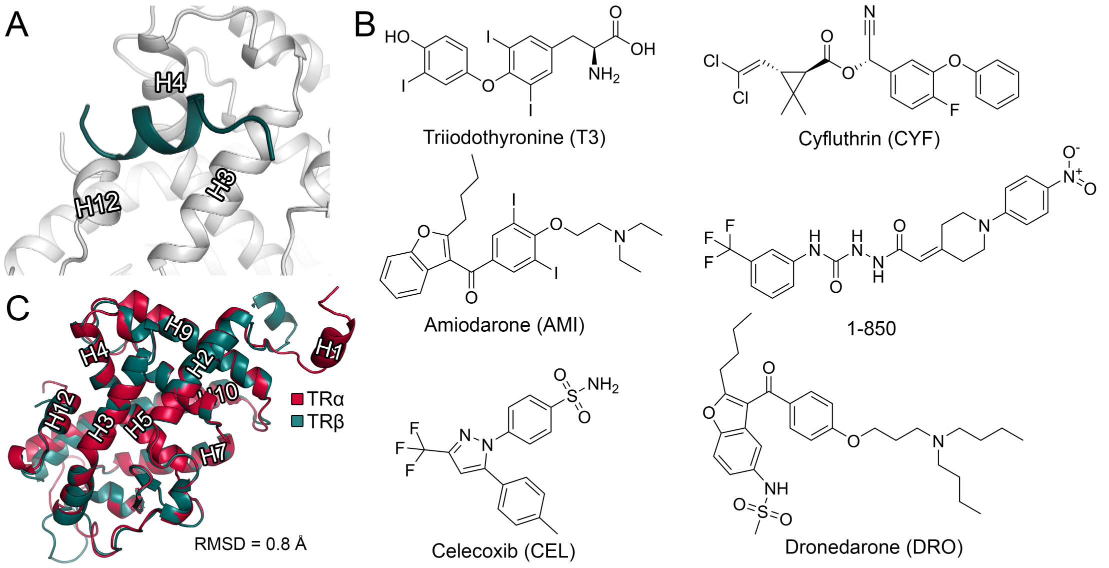


**Figure 1** Protein and ligand structures considered in this study. (A) Topology of the coactivator binding site AF-2. The cocrystallized coactivator fragment is shown in red. (B) Ligands explored in this study. (C) Superposition of TR$\alpha$ and TR$\beta$. The alpha carbon RMSD is indicated at the bottom.

study their antagonism. In the androgen receptor (AR) and estrogen receptor (ER), involved in the development and progression of malignant tumors, antagonism has been linked to structural rearrangements of H12 induced by altered interaction patterns in the ligand binding pocket (LBP). Further, alterations in other secondary structure elements such as H3 and H10 have been discussed [11, 12, 13, 15].

Here, we applied 27 $\mu$s MD simulations of TR$\alpha$ and TR$\beta$ complexed to thyriodothyronine (T3) and several antagonists to study the structural implications of antagonist binding to TRs in atomistic detail. Specifically, the experimentally verified TR antagonists evaluated in this study included the marketed drugs amiodarone, celecoxib, and dronedarone, as well as the pesticide cyfluthrin and the experimental compound 1-850 (Figure 1B and Table 1). We observed structural adaptations of multiple helices including H3, H10, and H12 including displacements, rearrangements, and loss of helicity. A mechanistic analysis based on dynamic cross-correlation combined with visual inspection revealed a functional relationship between H3 and H12 in these conformational adaptations. Further, we observed the narrowing of a major access pathway to the buried LBP as well as an increase in LBP volume with antagonists. Such a comprehensive study on the structural implications of TR antagonists was, until today, not reported.

**Table 1** Thyroid receptor antagonists examined in this study.

| **Ligand** | **Usage**[a] | **TR$\alpha$**[b] | **TR$\beta$**[c] | **Method**[d] |
|---|---|---|---|---|
| Celecoxib | Analgesic | no | yes | Luciferase assay [9] |
| Amiodarone | Antiarrhythmic | yes | yes | Radioligand assay [16] |
| Dronedarone | Antiarrhythmic | yes | no | Radioligand assay[17] |
| 1-850 | Experimental | yes | yes | Reporter gene assay [5] |
| Cyfluthrin | Pesticide | no | yes | Reporter gene assay [18] |

[a] Main practical application of the compound; [b] Antagonism at TR$\alpha$ shown in experiments; [c] Antagonism at TR$\beta$ shown in experiments; [d] Method and reference for mode of action determination.

## Results and Discussion

**The position of H12 is modified by TR antagonists.** In the signaling pathway of NRs, the association of coactivator proteins at the AF-2 site formed by helices H3, H4, H5, and H12 (Figure 1C) upon agonist binding constitutes a critical step for their transcriptional activation [19]. In order for a successful binding of coactivators to NRs, H12 must adopt a distinctive orientation [7]. As mentioned above, crystal structures of TRs bound to antagonists are unavailable as it is the case for several other NRs. In the AR, for example, it was proposed that the lack of an antagonist-bound crystal structure is caused by missing dissociation of chaperone proteins stabilizing the receptor [13]. In contrast, a structure of ER$\alpha$ bound to tamoxifen (PDB ID: 3ERT) revealed a clear displacement of H12 [20] impairing the binding of coactivators and, therefore, it is thought that many NRs behave similarly [21]. While the state of knowledge of antagonist-related conformational perturbations in TRs is limited, experiments based on hydrogen/deuterium exchange presented only minor differences regarding H12 solvent-accessibility in TR$\beta$ suggesting that it is still packed to the body of the LBD [7]. Further, a computational study on the effects of I280 mutations in TRs suggested a change in H12 orientation [14].

As shown in Table 1, there is a distinct ligand specificity of the TR isoforms regarding the interactions with the selected antagonists. Various experimental techniques have been applied to study the isotype-specific effects of the selected compounds as shown in a recent review [1]. As studies for some compounds were conducted on either TR$\alpha$ or TR$\beta$, we can not exclude their binding to the other isoform. Regardless, we limited our study on established ligand-protein complexes. Regarding ligand specificity, only one protein residue is different between TR$\alpha$ and TRB$\beta$ in the direct vicinity of the binding site. Thus, this amino acid was proposed to be the main selectivity factor [4]. In our simulations with antagonists, we could not detect a trend for increased root-mean square deviation (RMSD) of H12 based on average values (Figure 2A). In fact, the receptors responded individually to the different antagonists. The overall RMSD of the proteins demonstrated acceptable convergence of the simulations. However, the values of TR$\alpha$ indicated a slight instability, while they converged toward the end of the T3-bound and TR$\beta$ simulations except for celecoxib (Figure S1). The convergence of the simulations was also evident by the stability of the readouts discussed in the following sections. Regarding the binding modes, the same simulation systems presented some instabilities based on the ligand RMSD values, especially compared to T3 (Figure S2). However, the selected antagonists have more degrees of freedom and are structurally more complex as opposed to T3. Additionally, the antagonist-related structural changes of the receptor further allowed the ligands to reorient within the binding pocket, in which they remained in all simulations. In the five simulations with exceptionally high RMSD values, we observed a slight translation or reorganization of the flexible moiety of amiodarone, a translation of 1-850, a change of the flexible chain in dronedarone, and a substantial reorganization of the comparatively small celecoxib (Figure S3). Isoform-specific structural differences are to be expected as the sequence similarity between TR$\alpha$ and TR$\beta$ amounts to 82.3% (Figure S4) resulting in several structural differences (alpha carbon RMSD of 0.8 Å) in the direct vicinity of the binding site, the H2-H3 loop, as well as the terminal region after H12 among others (Figure S5). Since it was suggested that H12 adopts a different orientation in respect to the coactivator binding site with antagonists [7, 20, 22], we determined the distance between the centroids of H12 and a superimposed coactivator fragment (Figure 2B). In TR$\alpha$, the distance was higher for all three antagonists combined with a clear increase in fluctuations (Figure S6) suggesting a destabilization of H12 in contrast to a structural overlap with the coactivator proteins as it was observed in other receptors [11, 12, 20]. In TR$\beta$, no universal trend could be deduced regarding the increased fluctuations, but both amiodarone and 1-850 presented a similar tendency for higher distance values. Differences in H12 behavior between the isoforms might have occurred due to the structural differences at the terminus of H12 (Figure S7). In particular, the C-terminus is slightly longer in TR$\alpha$ and is lacking a terminal fixation by a charge-assisted hydrogen bond as opposed to TR$\beta$ leading to a potential for increased flexibility. A visual examination of representative structures derived from the last portion of the simulations revealed profound changes in the position of H12 for various systems (Figure 2C and Tables S1-S2). It clearly dislocated toward the coactivator binding site in one simulation with TR$\beta$ complexed to cyfluthrin occluding the coactivator binding site, as it was suggested based on hydrogen/deuterium exchange experiments with the antagonist NH3 (PubChem ID: 10027822) bound to TR$\beta$ [7]. Even though other simulations only revealed a minor displacement, H12 retained packing to the body of the LDB, in accordance with the above-mentioned experiments. Additionally, it was proposed that antagonism might be mediated by different structural mechanisms, which would explain differences between the individual compounds assessed here [5]. In conclusion, the observed conformational changes of H12 were subtle, varied between the TR isoforms as well as different ligands, and, for the most part, were in accordance with laboratory experiments.

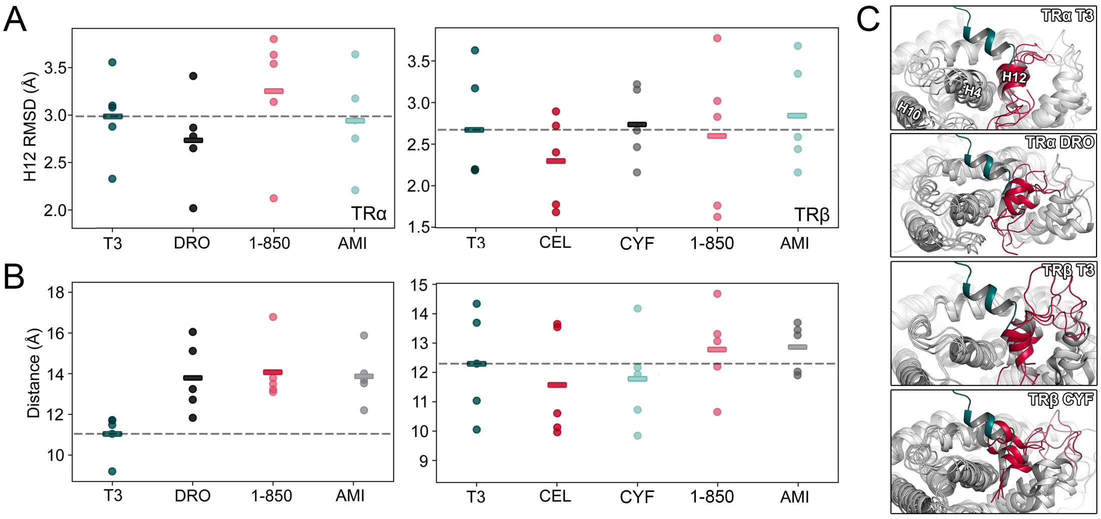


**Figure 2** Behavior of H12. (A) Average backbone RMSD values of H12 for all replica simulations. The dashed horizontal line represents the average value among the T3-bound baseline simulations, while colored short lines indicate the averages among the antagonist-bound simulations. (B) Average distance values of H12 to a superimposed coactivator fragment. The dashed horizontal line represents the average value among the T3-bound baseline simulations, while short colored lines indicate the averages among the antagonist-bound simulations. (C) Visualization of H12 (blue) and the coactivator fragment (red) in cluster structures of four different systems.

**TR antagonists impair coactivator binding by decreasing H3 helicity.** As mentioned above, H3 is equally involved in the formation of the AF-2 coactivator binding site as H12 (Figure 1A). In previous MD simulations with the AR, a distortion of H3 at its center in response to bicalutamide and hydroxyflutamide binding was described [11, 23]. While our results of the structural adaptations of H12 in response to antagonist binding suggested dissimilar behavior of TR$\alpha$ and TR$\beta$, the receptors presented similar conformational changes of H3. A visual inspection of the highest occupied cluster structures based on an RMSD matrix, presented various degrees of decreased helicity in H3, in particular around its center, in antagonist-bound complexes (Figures 3A and 3B). Intrigued by these observations, we determined the degree of helicity in all systems over the whole trajectory (Figures 3C and 3D). The analysis unveiled a certain loss of helicity in all systems exposed to antagonists suggesting that the perturbation of H3 is a common structural adaptation in both TR$\alpha$ and TR$\beta$. A mutational study with TR$\beta$ revealed the importance of H3 for coactivator recruitment [24]. Indeed, the superposition of a coactivator fragment revealed alterations of its direct interaction interface with the receptor caused by the distortion of H3 (Figure 4A), suggesting this interference to impair coactivator association, and thus, downstream signaling. In particular, the central portion of H3 undergoes hydrophobic interactions with a leucine residue of the coactivator, which are perturbed by H3 distortion (PDB ID: 1BSX). A following analysis of close contacts below 2.0 Å to a superimposed coactivator fragment revealed only few steric clashes in TR$\alpha$, while celecoxib and cyfluthrin in TR$\beta$ presented a significant increase in multiple replica simulations (Figure S8). Overall, this indicates that the coactivator protein interaction interface is not necessarily occluded, but structurally modified to prevent coactivator binding.

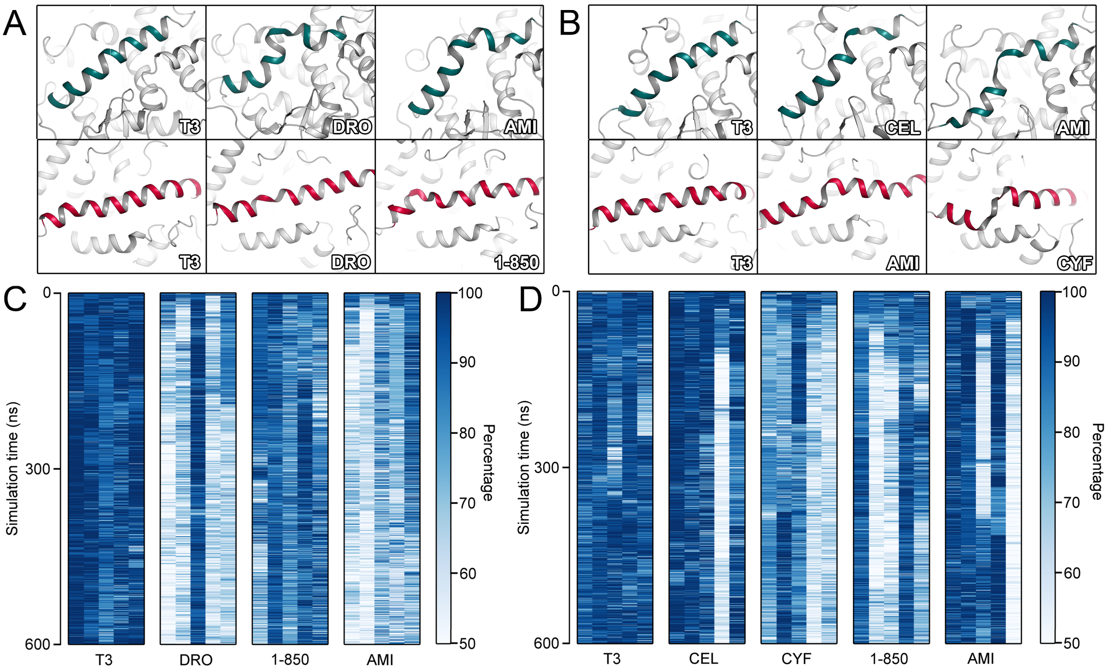


**Figure 3** Behavior of H3. (A) Structural changes of H3 (pine green) and H10 (red) in TR$\alpha$. (B) Structural changes of H3 (pine green) and H10 (red) in TR$\beta$. (C) Helicity analysis of TR$\alpha$. (D) Helicity analysis of TR$\beta$.

The visualization of several trajectories with such alterations of H3, revealed a molecular mechanism for its distortion (Figure 4B). Allosteric pathways have been associated with the inactivation of various NRs including TRs [7, 15, 23, 25, 26]. In the first step of this allosteric pathway, the antagonist interacts with residues of the H11-H12 loop and provokes a slight shift of this loop toward H3 without affecting its helicity. However, the following adaptation of H12 induced a propagating conformational change toward the center of H3 as shown in Figure S9, ultimately leading to its observed distortion coupled to a loss of helicity. Previous work on the AR concluded a similar allosteric pathway and, in one study, dynamic cross-correlation map (DCCM) analysis was used to investigate it [23, 25]. Correspondingly, we determined the DCCM for all systems (Figures 4C and S10-S16). As presented in Figure 4C, a slightly positive correlation could be determined for H3 and H12 supporting our visual observations regarding their collective motions. An overview of the observed correlations is given in Table S3. In fact, we detected the above-mentioned correlation in at least one replica simulation with all ligands and it occurred most frequently in the system with amiodarone bound to TR$\alpha$. Especially, in systems selected from visual inspection presenting major changes in H3, the correlation was particularly apparent. However, the correlation was more evident in TR$\alpha$ than TR$\beta$. Systems without such correlations presented only minor conformational changes of H3. Negative correlations around H12 indicate collective motions in the opposite direction. In a next step, we analyzed residue interaction networks to obtain the betweenness centrality (BC) measure describing the importance of a protein residue for intramolecular communication [27] (Figure S17). This analysis revealed increased BC values for residues in H3 and H12 adding more evidence for their relevance in the allosteric mechanism. Other structural regions presenting high BC values in both isoforms included the loop between H5 and the beta sheets and the H8-H9 loop in proximity to H10. Further examination of the trajectories revealed R228 in TR$\alpha$ and R282 in TR$\beta$ to adopt variable orientations depending on the nature of the bound ligand. As shown in Figure 4D, this residue took part in an ionic interaction with the carboxylate of T3, while it regularly faced towards the solvent when bound to dronedarone or celecoxib. This could be confirmed by computing the solvent accessible surface area (SASA) as well as its minimal distance to the respective ligand (Figure S18). Since this arginine residue is located at the center of H3, the lack of a negative charge of the antagonistic ligands leading to the absence of the salt bridge may contribute to the destabilization of H3. The slightly decreased SASA in the simulation with TR$\beta$ and celecoxib was caused by an interaction with the C-terminal aspartic acid residue implicated by a displacement of H12. Furthermore, visual inspection suggested different interaction patterns between hydrophobic residues located in the H11-H12 loop (L400 in TR$\alpha$ and L454 in TR$\beta$) and H3 (I226 in TR$\alpha$ and I280 in TR$\beta$). In the two above-mentioned simulations of celecoxib and dronedarone presenting decreased helicity of H3, the hydrophobic interaction energy between these residue pairs clearly correlated with the loss of helicity (Figure S19). It is important to note that not all simulations presented the same conformational changes and we specifically highlighted this structural mechanism in simulations exhibiting pronounced structural adaptations. In future studies, detailed experimental structure-activity relationships of TR agonists and antagonists might offer more elaborate insight into the initiation of the allosteric pathway and the particular ligand-protein interactions involved. In addition to H3, we detected conformational changes of H10 confirming experimental observations in TR$\beta$ [7]. The changes occurred in considerable distance to the LBP and the

AF-2 site (Figures 3A, 3B, and S20) suggesting no direct functional consequences on coactivator binding. However, H10 is part of the dimerization interface which may provide an additional functionality of antagonists to interfere with downstream signaling [28, 29]. A clear mechanism for the distortion of H10 could not be deduced, but the DCCM analysis presented a correlation between H10 and H4 to H5 in both TR$\alpha$ and TR$\beta$.

**Protein cavities and pockets are influenced by TR antagonists.** Even though the LBP of NRs is an occluded cavity within the core of the protein, crystal structures provide no obvious route for ligands to exchange with the solvent. Therefore, it is thought that molecular pathways connect this cavity to the surrounding solvent environment, as was shown for other protein systems. Due to the dynamic nature of these pathways, MD simulations offer an indispensable method to investigate them [30, 31, 32]. In a recent review, we identified the most frequently described pathway to be located among the H6-H7 loop, the H11-H12 loop, and the N-terminal section of H3 (pathway II) [30]. In TRs, steered MD simulations revealed the main pathway for ligand egress to be located among H3, the H1-H2 loop, and the adjacent beta sheets (pathway IIIa) [31]. Additional evidence for the involvement of this region in ligand egress was provided based on increased dissociation rates of T3 if certain mutations are present [28], which stands in accordance with the increased bottleneck radius we determined for pathway IIIa as opposed to pathway II. The bottleneck of a pathway describes its most narrow section, and therefore, it is thought to be the main determinant in gating the accessibility of the occluded pocket [33, 34]. While pathway II did not suffer from any drastic alterations of its bottleneck radius in response to antagonists, differences occurred for pathway IIIa (Figures 4E and S21). For multiple ligands, pathway IIIa presented decreased bottleneck radii potentially locking the antagonistic ligand in the LBP, which offers an additional explanation how antagonists efficiently block TR signaling.

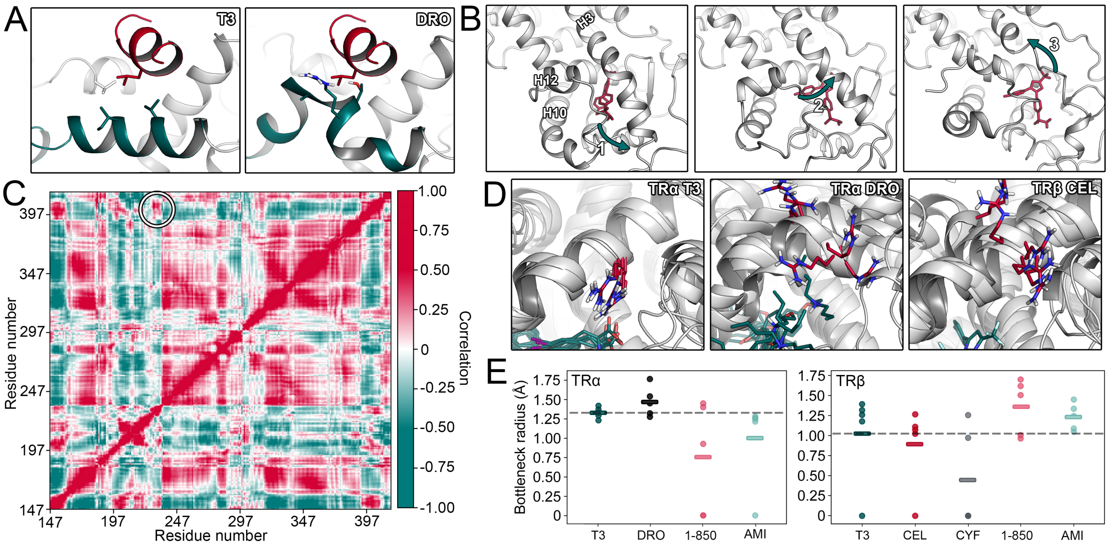


**Figure 4** Mechanistic analysis. (A) Relative position of H3 (pint green) to a superimposed coactivator fragment (red) in representative structures of TR$\alpha$ with T3 and dronedarone. (B) Visualization of the allosteric pathway involved in H3 distortion presented by the example of celecoxib bound to TR$\beta$. (C) Results of the DCCM analysis for TR$\alpha$ bound to dronedarone. The circle indicates the correlation between residues of H3 and H12. (D) Location of R228 in TR$\alpha$ and R282 in TR$\beta$ in representative structures. In the structure of TR$\beta$ and celecoxib, the interaction with the C-terminal aspartic acid residue is indicated. (E) Bottleneck radii of pathway IIIa computed in the ligand pathway analysis.

As opposed to the ligand pathways, the volume of the LBP and surrounding cavities was increased in all simulations with antagonists (Figures 5A and 5B). Similar observations have been made in the AR, for which multiple studies reported an increase in binding site volume in complexes with antagonists [11, 13]. The values predicted in our study overestimate the volume of the LBP due to the inclusion of neighboring voids and small parts of the surrounding solvent space, which did not have to be continuous as opposed to other methods [35]. Regardless, since all receptor systems are treated in the same manner, the deviation of absolute values compared to previous reports does not affect conclusions relating to the relative volumes among the complexes. An extension of the LBP volume was associated with the displacement of H12 in a crystal structure of TR$\beta$ complexed with the thyroid hormone thyroxine [36]. This suggests that the conformational changes of the H11-H12 region in presence of antagonists were responsible for the enlargement of the binding site. While the extension of the binding pocket likely has no direct functional implications on coactivator association, it seems to be a characteristic response to NR antagonists. Regardless, the increase of the active site might allosterically influence other parts of the receptor.

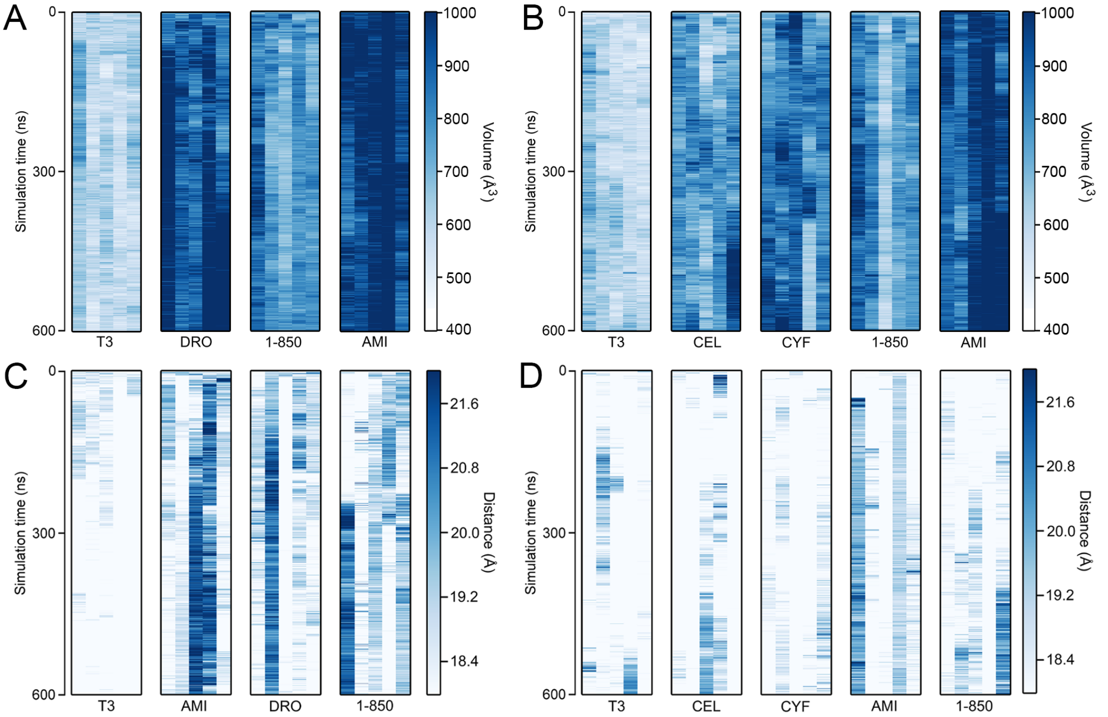


**Figure 5** Active site volume and charge clamps. (A) The active site volume of TR$\alpha$. (B) The active site volume of TR$\beta$. (C) The distance between the alpha carbons of the charge clamp residues K234 and E403 in TR$\alpha$. (D) The distance between the alpha carbons of the charge clamp residues K288 and E457 in TR$\beta$.

To get additional insight into the impaired binding of coactivators, we aimed on detecting structural alterations of the AF-2 site. A pair of charged residues on the AF-2 site referred to as charge clamps are known to be essential for coactivator association and are conserved in various NRs [13, 21, 23]. In TR$\alpha$, the charge clamps constitute of K234 and E403, while K288 and E457 are responsible in TR$\beta$ according to a structural alignment to the AR. A cocrystallized coactivator fragment is only available with TR$\beta$ (PDB ID: 1BSX) and shows a bidentate hydrogen bond from E457 to the coactivator, while K234 interacts with a neighboring protein carboxylate. Previous work focused on the AR revealed alterations of the electrostatic potential surface at the AF-2 site [13]. As this readout has been performed visually and we aimed on analyzing the relative position of the charge clamp residues over all trajectories, we determined the distance between the residues in each TR isoform. Especially in TR$\alpha$, there was a clear tendency for a larger separation in antagonist complexes compared to T3 (Figure 5C), while the changes were less pronounced in TR$\beta$ (Figure 5D). Since the lysine residue of the charge clamps is located at H3, this suggests that the conformational changes of H3 are likely involved in the increase of this distance. Such changes might significantly alter or even prevent the association of coactivator proteins.

## Materials and Methods

**Model building.** To obtain the complexes for the MD simulations, we used homology modeling for TR$\alpha$ and TR$\beta$ due to missing residues at both C-termini and the H2-H3 loop of TR$\beta$. For this purpose, we employed the Modeller (v9.21-1) toolkit [37] with two native structures of TR$\alpha$ (PDB ID: 2H79) and TR$\beta$ (PDB ID: 3GWS) bound to T3 retrieved from the Protein Data Bank [38] as starting point. Sequences in FASTA format were obtained from the UniProt database [39] according to entries B6ZGR6 and P10828. The sequence similarity was computed after alignment using the ClustalW algorithm within the UGENE (v1.32.0) toolkit [40]. The corresponding alignments were visualized with Jalview (v2.10.5) [41]. The sequence identity to the templates was 95.4% for TR$\alpha$ and 94.6% for TR$\beta$ (Figure S22). Due to the missing loop in TR$\beta$, five models were generated and the selection was made according to its DOPE score [37]. Both selected structures were treated with the Protein Preparation Wizard [42] within the Maestro Small-Molecule Drug Discovery Suite [43]. In detail, we added hydrogen atoms, determined bond orders, and predicted protonation states at pH 7.4. Next, the hydrogen bonding network was oriented using PROPKA at pH 7.4 and, finally, the system was minimized with the OPLS3e force field to a convergence threshold of 0.3 Å for protein heavy atoms.

We retrieved SMILES strings of the ligands from the PubChem database [44] and treated them with the LigPrep protocol in Maestro. Their protonation states were predicted at pH 7.4 by Epik and low-energy conformers were generated with the OPLS3e force field. In addition, favorable protonation states were determined in Marvin Sketch provided by ChemAxon [45]. In the case of T3, the predicted protonation state presented a deprotonated carboxylic acid and a protonated amine (pKa of 8.4 [46] and state occupancy of 10% at pH 7.4) function. However, while the visual inspection of the crystal structure revealed an arginine to interact with the negative charge of T3, a backbone nitrogen faced its amine (Figure S23). Therefore, in order to prevent this unfavorable interaction, we neutralized the positive charge of T3 while retaining the charge on the carboxylic acid group. For cyfluthrin containing multiple stereocenters, we modeled the form deposited in the PubChem database corresponding to (S)-cyano(4-fluoro-3-phenoxyphenyl)methyl (1R,3S)-3-(2,2-dichloroethenyl)-2,2-dimethyl-cyclopropane-1-carboxylate.

**Molecular docking.** The start conformations of the ligands within the modeled receptors were determined by rigid molecular docking with the AutoDock Vina software [47]. The centroid of the cubic search space with a side length of 22 Å was specified according to the mass center of cocrystallized T3. To ensure proper sampling of the search space, we configured an exhaustiveness of 16. Prior to the production runs, we evaluated the pose prediction accuracy of the protocol by docking the native ligand T3, as well as all available cocrystallized thyroid analogs to the selected models followed by a determination of the RMSD between the docked pose and the cocrystallized ligand. To compute the RMSD, we used the superposition panel in Maestro. Based on the sub-angstrom accuracy of the protocol for both isoforms in redocking T3 as well as the high accuracy for the thyroid analogs (Figure S24 and Table S4), we retained the previous settings for the following production runs generating complexes according to Table 1. The best pose of each antagonist was taken as input for the MD simulations, except for 1-850 in TR$\alpha$, for which no favorable orientation withing the binding site could be obtained. Thus, the binding mode of 1-850 in TR$\beta$ was superimposed on the TR$\alpha$ structure while alleviating clashes by slight adaption of the side chain dihedral angles of F215, I221, M259, and S277. The similarity between the cocrystallized ligands and T3 was computed by determining the Tanimoto coefficients based on atom connectivity fingerprints in the Similarity and Clustering panel within Maestro. As the obtained similarity coefficients underlined (Table S4), several ligands are dissimilar to T3 proving the capability of the protocol to predict the orientation of different chemotypes (Figure S24). Alternative poses obtained from AutoDock Vina were compared regarding their RMSD to the best pose as well as their relative binding free energies (Figure S25). Using the K-Means algorithm in the sklearn python toolkit, we computed clusters and determined their average binding free energy. As alternative binding modes with similar energies were either not available or within an RMSD of 2 Å to the selected pose, this analysis confirmed the best pose to be superior to the alternatives. To obtain more insight into the correctness of the binding poses, we used the induced-fit docking (IFD) protocol of Glide [48] as well as our in-house program DOLINA [49] and checked if similar binding modes could be obtained. The latter was specifically designed to cope with protein flexibility within NR binding sites. The analysis revealed a consensus of at least one of the mentioned protocols to the results obtained from AutoDock Vina providing additional confidence in the starting position for our MD simulations (Figure S26). Compounds, for which we obtained positive scores using AutoDock Vina, presented favorable negative scores in the Glide IFD protocol (Figure S27).

**MD simulations.** We selected the Desmond (v2019-1) simulation engine [50] for all MD simulations conducted in this study. The orthorhombic periodic boundary systems were solvated with TIP3P water molecules and the charge of the system was neutralized with counterions in the Desmond System Builder panel. The OPLS_2005 force field was selected to conduct the simulations with the time-step of the RESPA integrator set to 2 fs. After the default equilibration protocol, the production simulations with a duration of 600 ns were conducted in an NPT ensemble at 310 K regulated by the Nose-Hoover thermostat and atmospheric pressure maintained by the Martyna-Tobias-Klein barostat. The u-series algorithm [51] was selected by default to treat long-range interactions, while bonds to hydrogen atoms were treated with the M-SHAKE algorithm. All simulations were conducted with 5 replicas by varying the random seed for initial velocities, while atomic coordinates were recorded at an interval of 60 ps.

**Evaluation of the MD trajectories.** First off, the RMSD values of the simulations were computed in the Simulation Interaction Diagram panel in Maestro. Next, representative structures for the last 1000 frames of each trajectory were obtained using the trj_cluster.py routine that comes with Maestro and clusters the selected MD frames based on an RMSD matrix. To analyze the position of H12 in respect to the AF-2 binding site, we used an in-house python routine based on the structalign_utility.py script of Maestro to superimpose a coactivator fragment (PDB ID: 1BSX) to the respective MD frame. The distance between H12 and the coactivator was defined by their centroids based on the atoms presented in Table S5. Similarly, close contacts at a threshold of 2.0 Å between the protein heavy atoms and a superimposed coactivator protein were determined for all simulations. To estimate the helicity of H3 and H10, we exported 1000 frames of the complete trajectories and processed them with the STRIDE (v29.01.96) [52] program to assign a secondary structure to all residues. Using the output files, the percentage of helicity (including 3-10 helices) of H3 and H10 was determined based on the residues given in Table S6. For the correlation and network analyses, we used MD-TASK (v1.0.1) [27]. The trajectories were treated with the trj_extract_subsystem.py script and were centered on the protein with the trj_align.py routine included in Maestro. To ensure reliability, the trajectories were converted to the DCD file format using VMD [53]. The DCCM analysis was performed for every second frame of the first 300 ns of the trajectories, while every fifth frame was supplied to the network analysis to determine the BC. The BC values were averaged using the avg_network.py routine in MD-TASK. The pathways to the active site were analyzed with CAVER (v3.0) [33] by exporting 500 frames at a timestep of 1.2 ns centered on the protein. Ligand pathways were clustered at a threshold of 4.5 and visually assigned according to our previous work [30]. The starting point for the tunnel computation was identified based on the heavy atom centroid of T3 in the TR$\alpha$ crystal structure. To estimate the volume of the active site cavity, we selected the POVME (v2.0) [54] tool and processed 1000 trajectory frames of each simulation. The inclusion sphere was selected to have a radius of 12 Å (Figure S23) and a grid spacing of 0.5 Å. Similar to the tunnel computation, the centroid of the inclusion sphere to calculate the binding site volume was defined as the centroid of a cocrystallized ligand. The distance between the charge clamp residues was computed by an in-house python routine processing MD frames at a time step of 600 ps. The SASA was computed with the POPS algorithm [55] at default settings, while the distance between the arginine in H3 and the ligand was determined by an in-house python routine taking all ligand atoms and the terminal carbon of arginine into consideration. The hydrophobic energy between the leucine and isoleucine residues was determined with a previously published in-house routine [56] computing the energy according to the VSGB 2.0 energy model [57]. For visualization, we used PyMol (v2.1.1) [58] and ChemDraw (v16.0.1.4) [59].

## Conclusion

Even though it is known that impaired binding of coactivator proteins is the main principle behind TR antagonism, its exact molecular basis is still poorly understood. Here we applied microsecond MD simulations of protein-ligand complexes followed by their detailed examination to elucidate the conformational changes and molecular mechanisms involved in TR antagonism. While the conformational adaptations of H12, often referred to as the main driver of impairing coactivator association, were subtle confirming previous experiments, H3 was distorted in complexes with multiple antagonists. Since H3 is equally involved in forming the coactivator binding site AF-2, this suggested alterations in H3 to be primarily responsible for antagonism. Aiming on deducing the intramolecular mechanism for H3 distortion, we visualized MD trajectories and, in a next step, analyzed the DCCM of the protein residues. The results pointed to an allosteric pathway from the antagonist over H12 to H3. Further, we observed the narrowing of a major ligand access pathway to the buried binding pocket of TRs and an increase of the distance between charged coactivator recognition residues offering additional insight into the function of antagonists. Even though a mechanistic understanding of these conformational adaptations triggered by TR antagonists is important for the development of efficient novel therapeutics against hyperthyroidism, they were not previously examined in atomistic detail.

## References

[1] Louise S. Mackenzie. *Thyroid Hormone Receptor Antagonists: From Environmental Pollution to Novel Small Molecules*, volume 106. Elsevier Inc., 1 edition, 2018.

[2] Johan Malm, Mathias Farnegardh, Gary Grover, and Paul Ladenson. Thyroid Hormone Antagonists: Potential Medical Applications and Structure Activity Relationships. *Current Medicinal Chemistry*, 16(25):3258–3266, 2009.

[3] Tânia M. Ortiga-Carvalho, Aniket R. Sidhaye, and Fredric E. Wondisford. Thyroid hormone receptors and resistance to thyroid hormone disorders. *Nature Reviews Endocrinology*, 10(10):582–591, 2014.

[4] Girish Raparti, Suyog Jain, Karuna Ramteke, Mangala Murthy, Ravi Ghanghas, Sunita Ramanand, and Jaiprakash Ramanand. Selective thyroid hormone receptor modulators. *Indian Journal of Endocrinology and Metabolism*, 17(2):211, 2013.

[5] Matthieu Schapira, Bruce M. Raaka, Sharmistha Das, Li Fan, Maxim Totrov, Zhiguo Zhou, Stephen R. Wilson, Ruben Abagyan, and Herbert H. Samuels. Discovery of diverse thyroid hormone receptor antagonists by high-throughput docking. *Proceedings of the National Academy of Sciences of the United States of America*, 100(12):7354–7359, 2003.

[6] Michelle Leemans, Stephan Couderq, Barbara Demeneix, and Jean-Baptiste Fini. Pesticides With Potential Thyroid Hormone-Disrupting Effects: A Review of Recent Data. *Frontiers in endocrinology*, 10:743, 12 2019.

[7] A C M Figueira, D M Saidemberg, P C T Souza, L Martinez, T S Scanlan, J D Baxter, M S Skaf, M S Palma, P Webb, and I Polikarpov. Analysis of Agonist and Antagonist Effects on Thyroid Hormone Receptor Conformation by Hydrogen/Deuterium Exchange. *Molecular Endocrinology*, 25(1):15–31, 1 2011.

[8] Danielle Devereaux and Semhar Z Tewelde. Hyperthyroidism and Thyrotoxicosis. *Emergency Medicine Clinics of North America*, 32(2):277–292, 2014.

[9] Mire Zloh, Noelia Perez-Diaz, Leslie Tang, Pryank Patel, and Louise S. Mackenzie. Evidence that diclofenac and celecoxib are thyroid hormone receptor beta antagonists. *Life Sciences*, 146:66–72, 2016.

[10] Fangfang Wang and Jinyi Xing. Classification of thyroid hormone receptor agonists and antagonists using statistical learning approaches. *Molecular diversity*, 23(1):85–92, 2 2019.

[11] D. J. Osguthorpe and A. T. Hagler. Mechanism of androgen receptor antagonism by bicalutamide in the treatment of prostate cancer. *Biochemistry*, 50(19):4105–4113, 2011.

[12] Mojie Duan, Na Liu, Wenfang Zhou, Dan Li, Minghui Yang, and Tingjun Hou. Structural Diversity of Ligand-Binding Androgen Receptors Revealed by Microsecond Long Molecular Dynamics Simulations and Enhanced Sampling. *Journal of Chemical Theory and Computation*, 12(9):4611–4619, 2016.

[13] Sugunadevi Sakkiah, Rebecca Kusko, Bohu Pan, Wenjing Guo, Weigong Ge, Weida Tong, and Huixiao Hong. Structural changes due to antagonist binding in ligand binding pocket of androgen receptor elucidated through molecular dynamics simulations. *Frontiers in Pharmacology*, 9(MAY):1–13, 2018.

[14] Paulo C T Souza, Gustavo B Barra, Lara F R Velasco, Isabel C J Ribeiro, Luiz A Simeoni, Marie Togashi, Paul Webb, Francisco A R Neves, Munir S Skaf, Leandro Martínez, and Igor Polikarpov. Helix 12 Dynamics and Thyroid Hormone Receptor Activity: Experimental and Molecular Dynamics Studies of Ile280 Mutants. *Journal of Molecular Biology*, 412(5):882–893, 2011.

[15] Abbas Khan, Ashfaq-Ur-Rehman, Muhammad Junaid, Cheng Dong Li, Shoaib Saleem, Fahad Humayun, Shazia Shamas, Syed Shujait Ali, Zainib Babar, and Dong Qing Wei. Dynamics Insights Into the Gain of Flexibility by Helix-12 in ESR1 as a Mechanism of Resistance to Drugs in Breast Cancer Cell Lines. *Frontiers in Molecular Biosciences*, 6 (January):1–14, 2020.

[16] Wilmar M Wiersinga. Pharmacological Effects of Amiodarone and Dronedarone on Cardiac Thyroid Hormone Receptors BT - Thyroid and Heart Failure: From Pathophysiology to Clinics. pages 89–95. Springer Milan, Milano, 2009. ISBN 978-88-470-1143-4.

[17] H. C. Van Beeren, W. M.C. Jong, E. Kaptein, T. J. Visser, O. Bakker, and W. M. Wiersinga. Dronerarone acts as a selective inhibitor of 3,5,3'-triiodothyronine binding to thyroid hormone receptor-$\alpha$1: In vitro and in vivo evidence. *Endocrinology*, 144(2):552–558, 2003.

[18] Guizhen Du, Ouxi Shen, Hong Sun, Juan Fei, Chuncheng Lu, Ling Song, Yankai Xia, Shoulin Wang, and Xinru Wang. Assessing hormone receptor activities of pyrethroid insecticides and their metabolites in reporter gene assays. *Toxicological Sciences*, 116(1): 58–66, 2010.

[19] Zheng Sun and Yong Xu. Nuclear Receptor Coactivators (NCOAs) and Corepressors (NCORs) in the Brain. *Endocrinology*, 161(8):1–12, 2020.

[20] A K Shiau, D Barstad, P M Loria, L Cheng, P J Kushner, D A Agard, and G L Greene. The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen. *Cell*, 95(7):927–937, 12 1998.

[21] N. R.Carina Alves, Adali Pecci, and Lautaro D. Alvarez. Structural Insights into the Ligand Binding Domain of the Glucocorticoid Receptor: A Molecular Dynamics Study. *Journal of Chemical Information and Modeling*, 60(2):794–804, 2020.

[22] Ashley C.W. Pike, A. Marek Brzozowski, Julia Walton, Roderick E. Hubbard, Ann Gerd Thorsell, Yi Lin Li, Jan Ake Gustafsson, and Mats Carlquist. Structural insights into the mode of action of a pure antiestrogen. *Structure*, 9(2):145–153, 2001.

[23] Ye Jin, Mojie Duan, Xuwen Wang, Xiaotian Kong, Wenfang Zhou, Huiyong Sun, Hui Liu, Dan Li, Huidong Yu, Youyong Li, and Tingjun Hou. Communication between the Ligand-Binding Pocket and the Activation Function-2 Domain of Androgen Receptor Revealed by Molecular Dynamics Simulations. *Journal of Chemical Information and Modeling*, 59 (2):842–857, 2019.

[24] T N Collingwood, R Wagner, C H Matthews, R J Clifton-Bligh, M Gurnell, O Rajanayagam, M Agostini, R J Fletterick, P Beck-Peccoz, W Reinhardt, G Binder, M B Ranke, A Hermus, R D Hesch, J Lazarus, P Newrick, V Parfitt, P Raggatt, F de Zegher, and V K Chatterjee. A role for helix 3 of the TRbeta ligand-binding domain in coactivator recruitment identified by characterization of a third cluster of mutations in resistance to thyroid hormone. *The EMBO journal*, 17(16):4760–4770, 8 1998.

[25] Na Liu, Wenfang Zhou, Yue Guo, Junmei Wang, Weitao Fu, Huiyong Sun, Dan Li, Mojie Duan, and Tingjun Hou. Molecular Dynamics Simulations Revealed the Regulation of Ligands to the Interactions between Androgen Receptor and its Coactivator. *Journal of Chemical Information and Modeling*, 58:1652–1661, 2018.

[26] Shoshana J Wodak, Emanuele Paci, Nikolay V Dokholyan, Igor N Berezovsky, Amnon Horovitz, Jing Li, Vincent J Hilser, Ivet Bahar, John Karanicolas, Gerhard Stock, Peter Hamm, Roland H Stote, Jerome Eberhardt, Yassmine Chebaro, Annick Dejaegere, Marco Cecchini, Jean-Pierre Changeux, Peter G Bolhuis, Jocelyne Vreede, Pietro Faccioli, Simone Orioli, Riccardo Ravasio, Le Yan, Carolina Brito, Matthieu Wyart, Paraskevi Gkeka, Ivan Rivalta, Giulia Palermo, J Andrew McCammon, Joanna Panecka-Hofman, Rebecca C Wade, Antonella Di Pizio, Masha Y Niv, Ruth Nussinov, Chung-Jung Tsai, Hyunbum Jang, Dzmitry Padhorny, Dima Kozakov, and Tom McLeish. Allostery in Its Many Disguises: From Theory to Applications. *Structure*, 27(4):566–578, 2019.

[27] David K Brown, David L Penkler, Olivier Sheik Amamuddy, Caroline Ross, Ali Rana Atilgan, Canan Atilgan, and Oezlem Tastan Bishop. MD-TASK: a software suite for analyzing molecular dynamics trajectories. *Bioinformatics (Oxford, England)*, 33(17):2768–2771, 9 2017.

[28] Suzana T Cunha Lima and Edson D Rodrigues. The oligomeric state of thyroid receptor regulates hormone binding kinetics. *Journal of Endocrinology*, 210(1):125–134, 2011.

[29] Richard L Wagner, James W Apriletti, Mary E McGrath, Brian L West, John D Baxter, and Robert J Fletterick. A structural role for hormone in the thyroid hormone receptor. *Nature*, 378(6558):690–697, 1995.

[30] André Fischer, Martin Smiesko, and Martin Smieško. Ligand Pathways in Nuclear Receptors. *Journal of Chemical Information and Modeling*, 59(7):3100–3109, 2019.

[31] Leandro Martínez, Paul Webb, Igor Polikarpov, and Munir S Skaf. Molecular dynamics simulations of ligand dissociation from thyroid hormone receptors: Evidence of the likeliest escape pathway and its implications for the design of novel ligands. *Journal of Medicinal Chemistry*, 49(1):23–26, 2006.

[32] Weihua Li, Jing Fu, Feixiong Cheng, Mingyue Zheng, Jian Zhang, Guixia Liu, and Yun Tang. Unbinding pathways of GW4064 from human farnesoid X receptor as revealed by molecular dynamics simulations. *Journal of Chemical Information and Modeling*, 52(11): 3043–3052, 2012.

[33] Eva Chovancova, Antonin Pavelka, Petr Benes, Ondrej Strnad, Jan Brezovsky, Barbora Kozlikova, Artur Gora, Vilem Sustr, Martin Klvana, Petr Medek, Lada Biedermannova, Jiri Sochor, and Jiri Damborsky. CAVER 3.0: A Tool for the Analysis of Transport Pathways in Dynamic Protein Structures. *PLoS Computational Biology*, 8(10):23–30, 2012.

[34] Artur Gora, Jan Brezovsky, and Jiri Damborsky. Gates of enzymes. *Chemical Reviews*, 113(8):5871–5923, 2013.

[35] Thomas A Halgren. Identifying and Characterizing Binding Sites and Assessing Druggability. *Journal of Chemical Information and Modeling*, 49(2):377–389, 2 2009.

[36] Iván Lazcano, Gabriela Hernández-Puga, Juan Pablo Robles, and Aurea Orozco. Alternative ligands for thyroid hormone receptors. *Molecular and cellular endocrinology*, 493: 110448, 8 2019.

[37] Benjamin Webb and Andrej Sali. Comparative Protein Structure Modeling Using MODELLER. *Current Protocols in Bioinformatics*, 54(1):1–5, 6 2016.

[38] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, T N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 1 2000.

[39] The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1):D506–D515, 1 2019.

[40] Konstantin Okonechnikov, Olga Golosova, Mikhail Fursov, Alexey Varlamov, Yuri Vaskin, Ivan Efremov, O. G. German Grehov, Denis Kandrov, Kirill Rasputin, Maxim Syabro, and Timur Tleukenov. Unipro UGENE: A unified bioinformatics toolkit. *Bioinformatics*, 28 (8):1166–1167, 2012.

[41] Andrew M. Waterhouse, James B. Procter, David M.A. Martin, Michèle Clamp, and Geoffrey J. Barton. Jalview Version 2-A multiple sequence alignment editor and analysis workbench. *Bioinformatics*, 25(9):1189–1191, 2009.

[42] G. Madhavi Sastry, Matvey Adzhigirey, Tyler Day, Ramakrishna Annabhimoju, and Woody Sherman. Protein and ligand preparation: Parameters, protocols, and influence on virtual screening enrichments. *Journal of Computer-Aided Molecular Design*, 27(3): 221–234, 2013.

[43] Schrödinger LCC. Maestro Small-Molecule Drug Discovery Suite 2019-3. 2019.

[44] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan E Bolton. PubChem 2019 update: improved access to chemical data. *Nucleic acids research*, 47(D1):D1102–D1109, 1 2019.

[45] ChemAxon. Marvin (v.20.4.0), 2020.

[46] Marc W Harrold and Robin M Zavod. Basic Concepts in Medicinal Chemistry. *Drug Development and Industrial Pharmacy*, 40(7):988, 7 2014.

[47] Oleg Trott and Arthur J Olson. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. *Journal of computational chemistry*, 31(2):455–461, 1 2010.

[48] Thomas A. Halgren, Robert B. Murphy, Richard A. Friesner, Hege S. Beard, Leah L. Frye, W. Thomas Pollard, and Jay L. Banks. Glide: A New Approach for Rapid, Accurate

Docking and Scoring. 2. Enrichment Factors in Database Screening. *Journal of Medicinal Chemistry*, 47(7):1750–1759, 2004.

[49] Martin Smieško. DOLINA – Docking Based on a Local Induced-Fit Algorithm: Application toward Small-Molecule Binding to Nuclear Receptors. *Journal of Chemical Information and Modeling*, 53(6):1415–1423, 6 2013.

[50] Kevin Bowers, Edmond Chow, Huafeng Xu, Ron Dror, Michael Eastwood, Brent Gregersen, John Klepeis, Istvan Kolossvary, Mark Moraes, Federico Sacerdoti, John Salmon, Yibing Shan, and David Shaw. Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters. *ACM/IEEE SC 2006 Conference (SC'06)*, (November):43, 2006.

[51] David E. Shaw, J. P. Grossman, Joseph A. Bank, Brannon Batson, J. Adam Butts, Jack C. Chao, Martin M. Deneroff, Ron O. Dror, Amos Even, Christopher H. Fenton, Anthony Forte, Joseph Gagliardo, Gennette Gill, Brian Greskamp, C. Richard Ho, Douglas J. Ierardi, Lev Iserovich, Jeffrey S. Kuskin, Richard H. Larson, Timothy Layman, Li Siang Lee, Adam K. Lerer, Chester Li, Daniel Killebrew, Kenneth M. Mackenzie, Shark Yeuk Hai Mok, Mark A. Moraes, Rolf Mueller, Lawrence J. Nociolo, Jon L. Peticolas, Terry Quan, Daniel Ramot, John K. Salmon, Daniele P. Scarpazza, U. Ben Schafer, Naseer Siddique, Christopher W. Snyder, Jochen Spengler, Ping Tak Peter Tang, Michael Theobald, Horia Toma, Brian Towles, Benjamin Vitale, Stanley C. Wang, and Cliff Young. Anton 2: Raising the Bar for Performance and Programmability in a Special-Purpose Molecular Dynamics Supercomputer. *International Conference for High Performance Computing, Networking, Storage and Analysis, SC*, 2015-Janua(January):41–53, 2014.

[52] D Frishman and P Argos. Knowledge-based protein secondary structure assignment. *Proteins*, 23(4):566–579, 12 1995.

[53] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD: Visual molecular dynamics. *Journal of Molecular Graphics*, 14(1):33–38, 1996.

[54] Jacob D. Durrant, Lane Votapka, Jesper Sørensen, and Rommie E. Amaro. POVME 2.0: An enhanced tool for determining pocket shape and volume characteristics. *Journal of Chemical Theory and Computation*, 10(11):5047–5056, 2014.

[55] Luigi Cavallo, Jens Kleinjung, and Franca Fraternali. POPS: A fast algorithm for solvent accessible surface areas at atomic and residue level. *Nucleic acids research*, 31(13):3364–3366, 7 2003.

[56] André Fischer and Martin Smieško. Spontaneous Ligand Access Events to Membrane-Bound Cytochrome P450 2D6 Sampled at Atomic Resolution. *Scientific Reports*, 9(1): 16411, 2019.

[57] Jianing Li, Robert Abel, Kai Zhu, Yixiang Cao, Suwen Zhao, and Richard A. Friesner. The VSGB 2.0 model: A next generation energy model for high resolution protein structure modeling. *Proteins.*, 79(10):2794–2812, 2011.

[58] Schrodinger LLC. The PyMOL Molecular Graphics System, Version 2.1.1. 2018.

[59] PerkinElmer. ChemDraw 16.0.1.4. 2017.

## 7.1 Supporting Information

# Supporting Results and Discussion

### The position of H12 is modified by TR antagonists

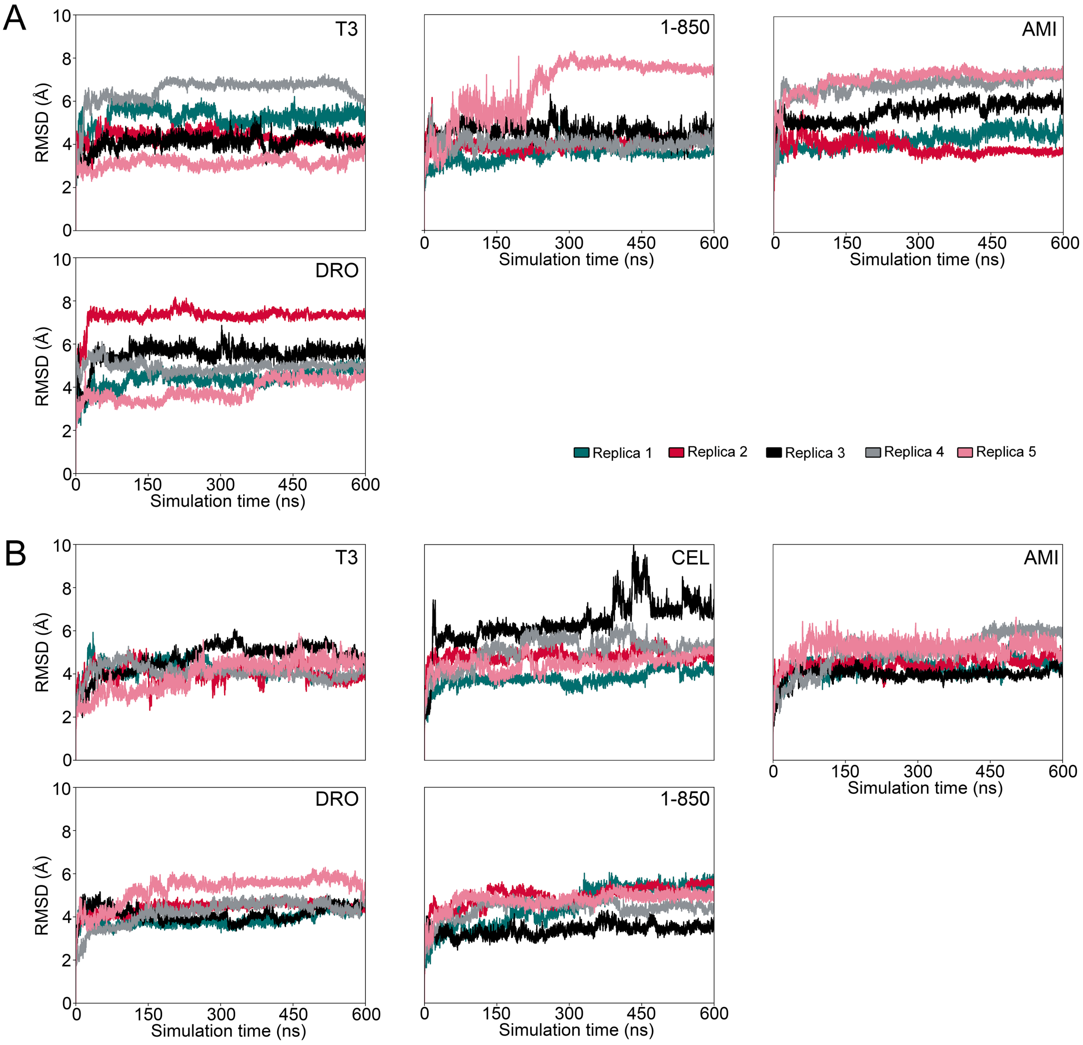


**Figure S 1** RMSD of (A) TR$\alpha$ and (B) TR$\beta$ systems.

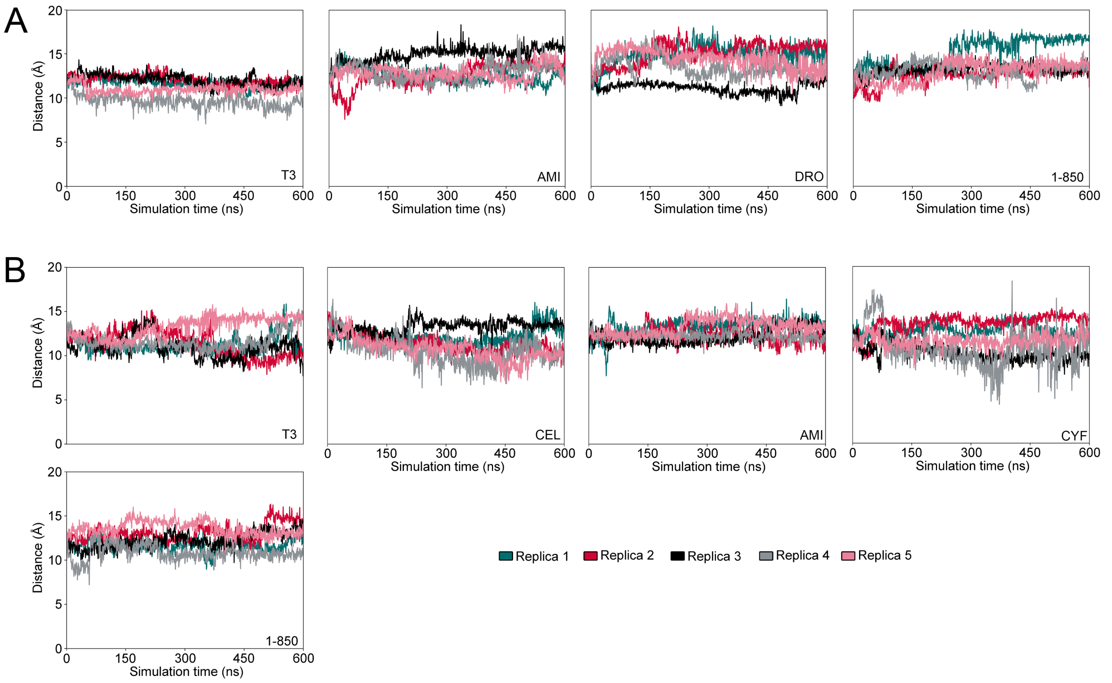


**Figure S 2** Distance of H12 to superimposed coactivator fragment of (A) TR$\alpha$ and (B) TR$\beta$ systems.

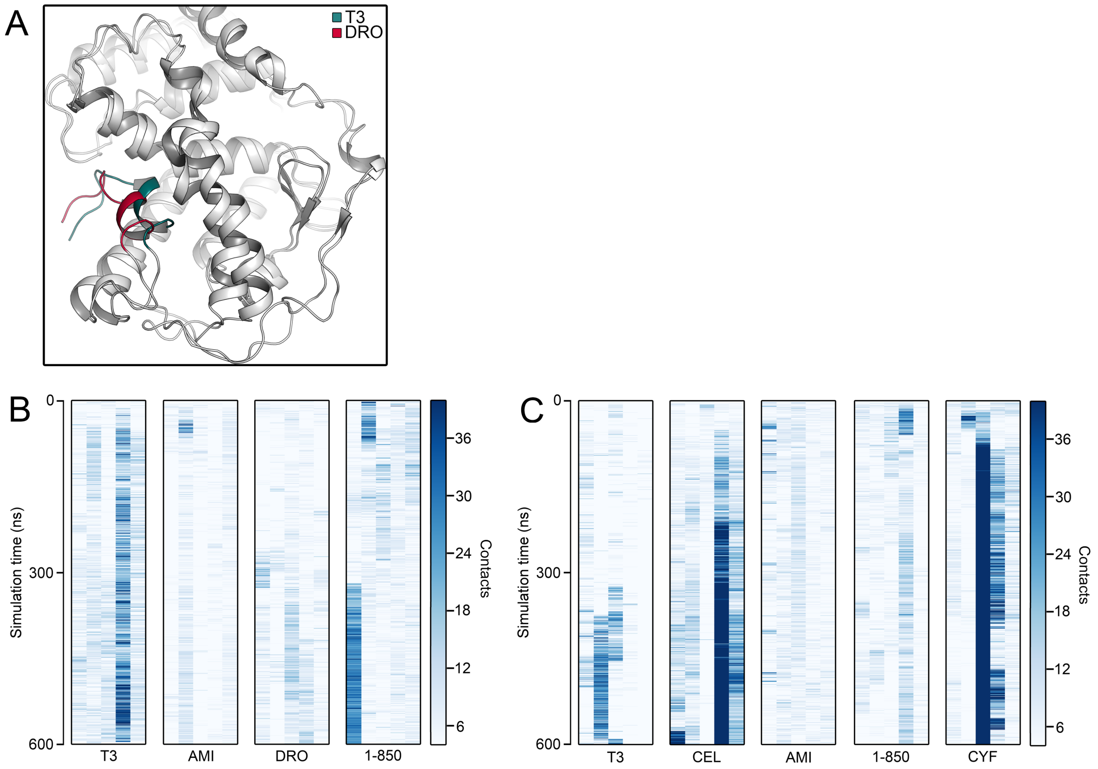


**Figure S 3** Interference with coactivator binding. (A) Superposition of H12 in TR$\alpha$ with T3 and dronedarone. H12 is presented in two different colors to illustrate its displacement. (B) Close contacts between the protein and a superimposed coactivator fragment in TR$\alpha$. (C) Close contacts between the protein and a superimposed coactivator fragment in TR$\beta$.

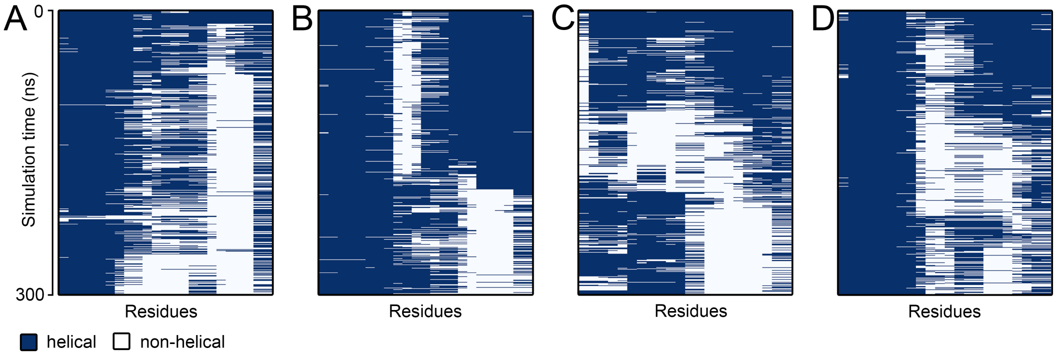


**Figure S 4** The propagation of H3 distortion. Per-residue helicity of H3 of (A) and (B) TR$\alpha$ with dronedarone, (C) TR$\beta$ with celecoxib, and (D) TR$\beta$ with 1-850. The helicity was determined for the first half of the trajectory and is shown from N-terminus (left) to C-terminus (right) of H3.

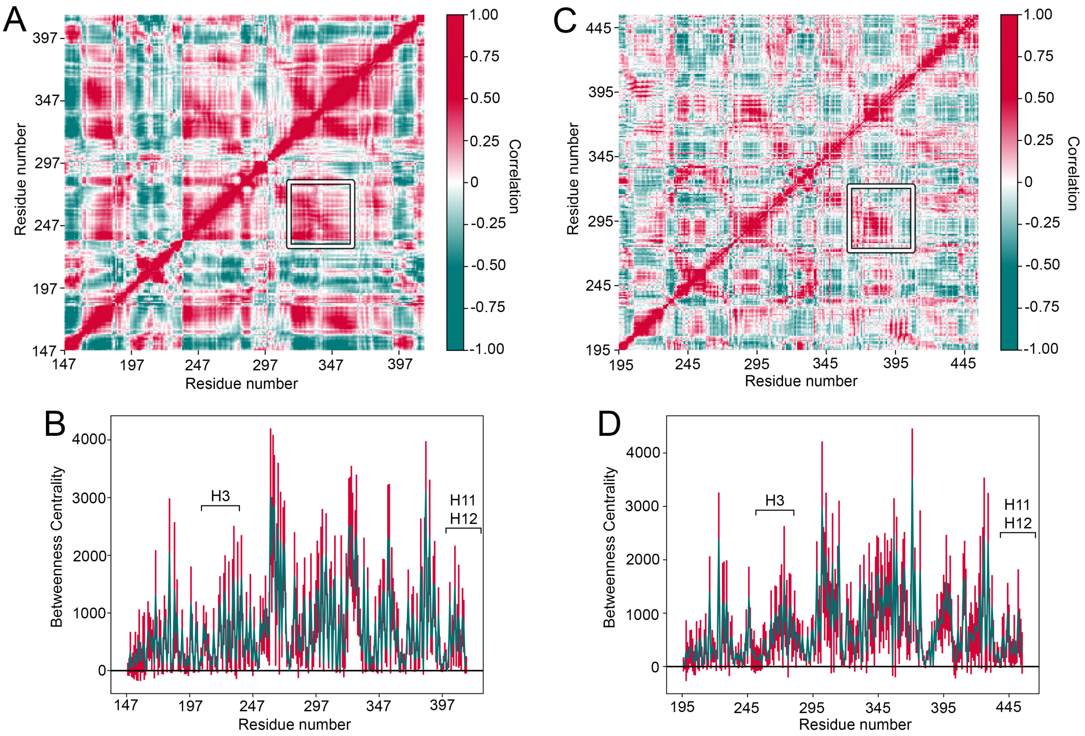


**Figure S 5** The allosteric mechanism of H3 distortion. (A) DCCM analysis of TR$\alpha$ complexed with dronedarone. The correlation between H10 and H4-H5 is indicated by a rectangle. (B) BC analysis of TR$\alpha$ complexed with dronedarone. The sections of H3 and H11/H12 are indicated. (C) DCCM analysis of TR$\beta$ complexed with celecoxib. The correlation between H10 and H4-H5 is indicated by a rectangle. (D) BC analysis of TR$\beta$ complexed with celecoxib. The sections of H3 and H11/H12 are indicated.

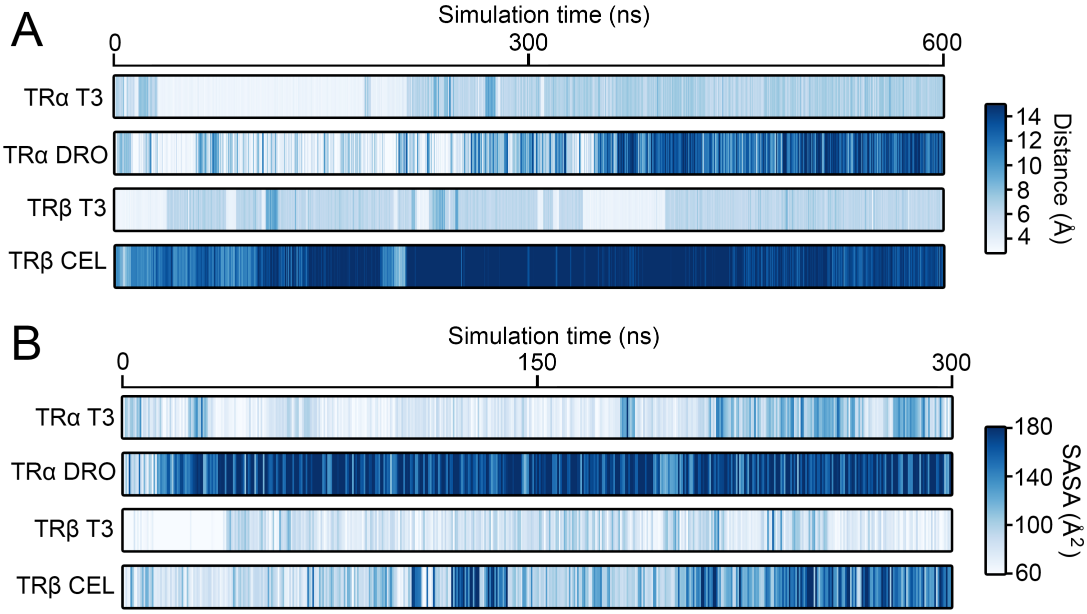


**Figure S 6** Metrics of the arginine in H3. (A) Shortest distance between R228 in TR$\alpha$ or R282 in TR$\beta$ and the respective ligand over the whole trajectory. (B) The SASA of R228 in TR$\alpha$ and R282 in TR$\beta$ over the first half of the respective trajectory.

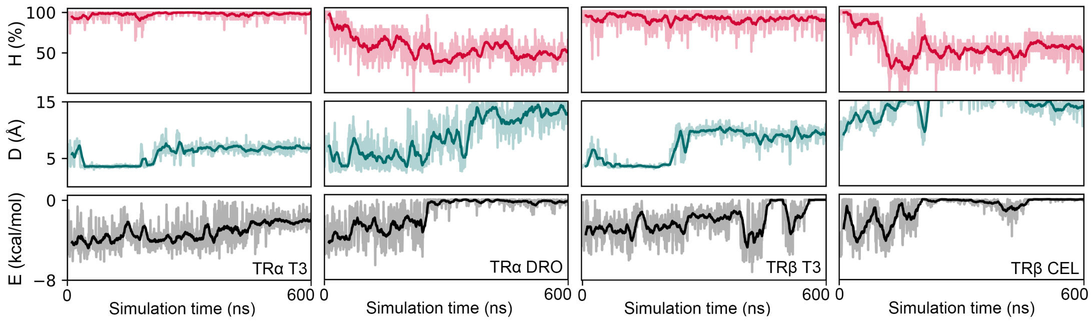


**Figure S 7** H3 helicity, shortest distance from R228 (TR$\alpha$) or R282 (TR$\beta$) to the ligand, and hydrophobic energy between I226-L400 (TR$\alpha$) or I280-L454 (TR$\beta$). The floating average was computed with a window of 20.

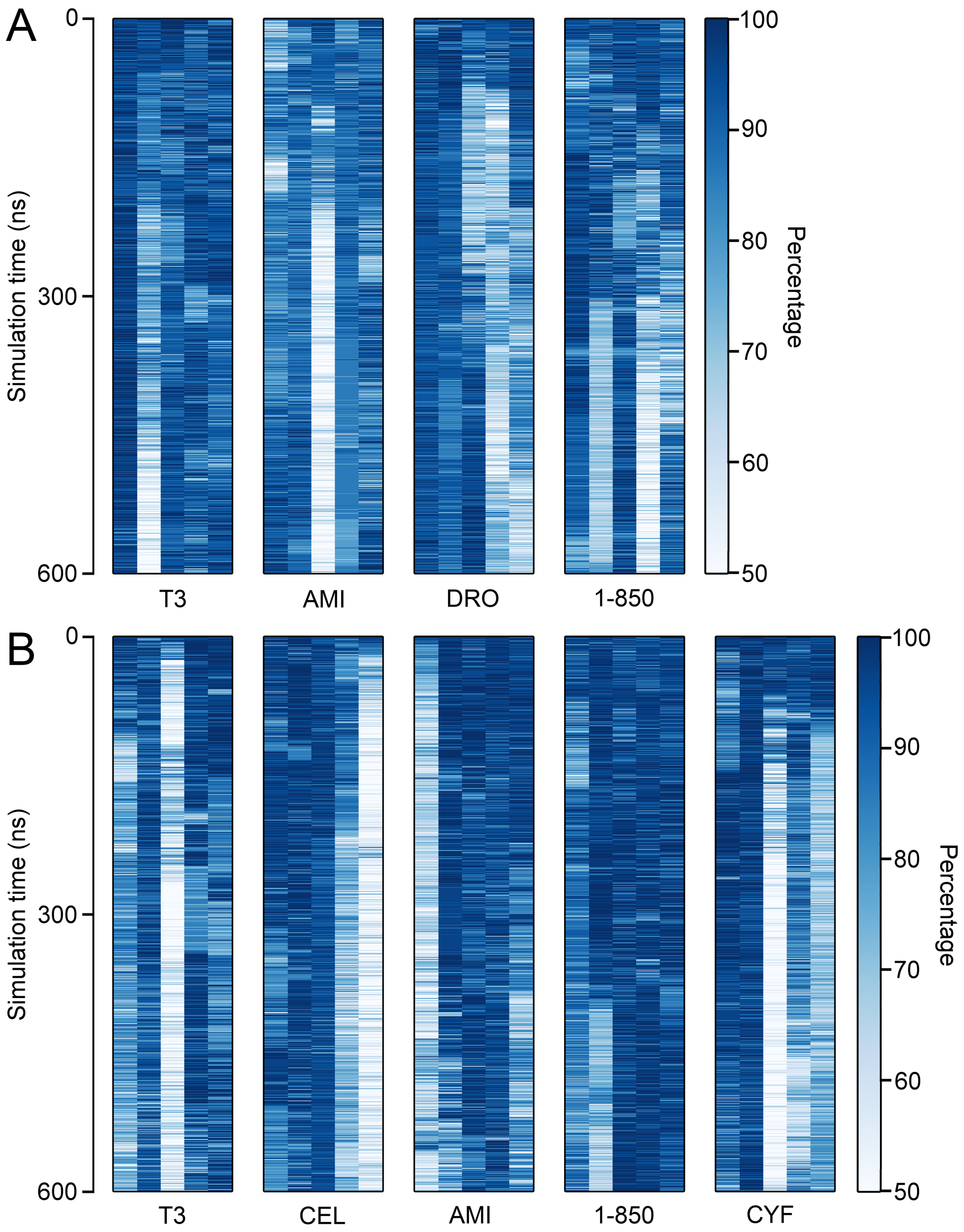


**Figure S 8** Helicity analysis of H10 in (A) TR$\alpha$ and (B) TR$\beta$.

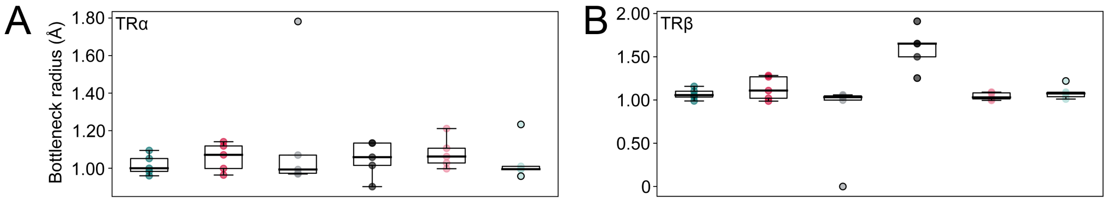


**Figure S 9** Pathway II bottleneck radius analysis of (A) TR$\alpha$ and (B) TR$\beta$.

# Supporting Materials and Methods

## Model building

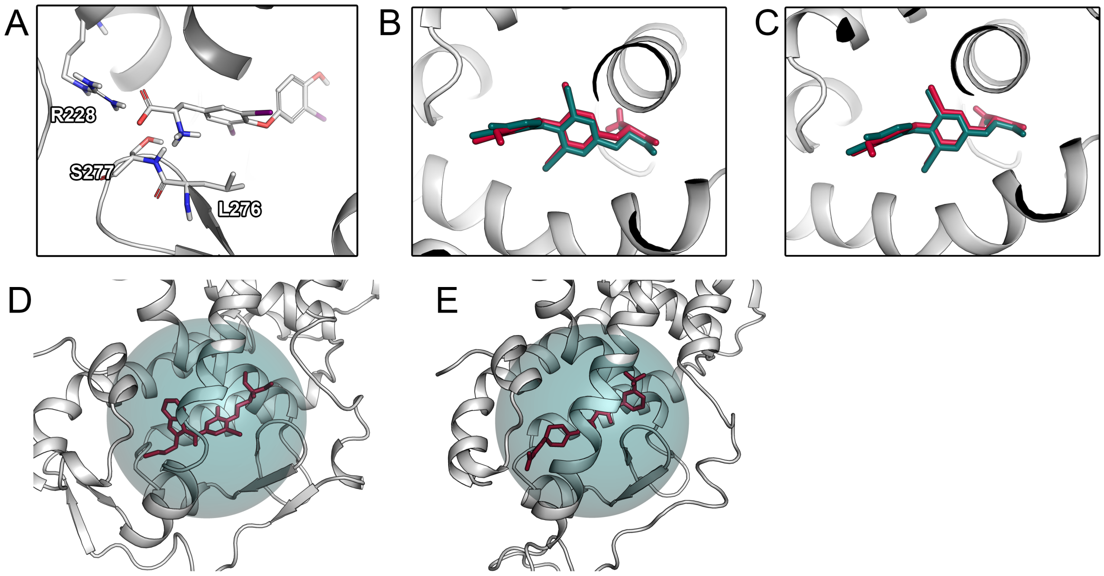


**Figure S 10** (A) Missing interaction partner of the protonated amine of T3. (B) Redocking of T3 in TR$\alpha$. (C) Redocking of T3 in TR$\beta$. (D) Inclusion sphere of TR$\alpha$ and amiodarone for active site calculation. (E) Inclusion sphere of TR$\beta$ and 1-850 for active site calculation.

## Molecular docking

**Table S 1** RMSD values obtained from redocking T3.

| Receptor | Reference PDB | RMSD (Å) |
|---|---|---|
| TR$\alpha$ | 1NAV | 1.94 |
| | 2H79[a] | 0.59 |
| | 3HZF | 1.04 |
| TR$\beta$ | 1N46 | 1.50 |
| | 1NAX | 0.67 |
| | 1Q4X | 2.02 |
| | 1R6G | 2.47 |
| | 2J4A | 0.90 |
| | 3GWS[a] | 0.59 |
| | 3IMY | 0.97 |
| | 6KKB | 1.64 |

[a] Structures with T3 used in this study.

## Evaluation of the MD trajectories

**Table S 2** Residues considered to be part of H12.

| **Receptor** | **H12 residues** |
|---|---|
| TR$\alpha$ | P399, L400, F401, L402, E403, V404 |
| TR$\beta$ | P453, L454, F455, L456, E457, V458 |

**Table S 3** Residues considered to be part of H3 and H10.

| **Receptor** | **H3 residues** | **H10 residues** |
|---|---|---|
| TR$\alpha$ | L212-K234 | W364-E391 |
| TR$\beta$ | E267-K288 | W418-V444 |

## CHAPTER 8

# Allosteric Binding Sites On Nuclear Receptors: Focus On Drug Efficacy and Selectivity

Similar to CYPs, NRs have allosteric binding sites that influence their signaling. This chapter is focused on the systematic assessment of the efficacy and selectivity of compounds designed to allosterically modulate the function of eight NRs responsible for the action of steroid and thyroid hormones. Compounds binding to the AF-2 or BF-3 allosteric sites have been designed to treat hormone-dependent cancers. In this work, several aspects of molecular recognition were addressed including ligand-protein interactions and solvation.

---

**Author contributions:** Conceptualization, A.F.; formal analysis, A.F.; writing and original draft preparation, A.F.; writing, review and editing, A.F., M.S.; visualization, A.F.; supervision, M.S.

---

*Based on the published research article:*

Fischer, A.; Smieško, M. Allosteric Binding Sites on Nuclear Receptors: Focus on Drug Efficacy and Selectivity. *Int. J. Mol. Sci.* 2020, 21, 6–8.

## Abstract

Nuclear receptors (NRs) are highly relevant drug targets in major indications such as oncologic, metabolic, reproductive and immunologic diseases. However, currently marketed drugs designed towards the orthosteric binding site of NRs often suffer from resistance mechanisms and poor selectivity. The identification of two superficial allosteric sites activation function-2 (AF-2) and binding function-3 (BF-3) as novel drug targets sparked the development of inhibitors, while selectivity concerns due to a high conservation degree remained. To determine important pharmacophores and hydration sites among AF-2 and BF-3 of eight hormonal NRs, we systematically analyzed over 10 $\mu$s of molecular dynamics simulations including simulations in explicit water and solvent mixtures. In addition, a library of over 300 allosteric inhibitors was evaluated by molecular docking. Based on our results, we suggest the BF-3 site to offer a higher potential for drug selectivity as opposed to the AF-2 site that is more conserved among the selected receptors. Detected similarities among the AF-2 sites of various NRs urge for a broader selectivity assessment in future studies. In combination with the supporting materials, this work provides a foundation to improve both selectivity and potency of allosteric inhibitors in a rational manner and increase the therapeutic applicability of this promising compound class.

## Introduction

Nuclear receptors (NRs) are ligand-inducible transcription factors that are attractive drug targets due to their involvement in a multitude of physiological and pathological processes. Currently marketed drugs designed to interact with the buried ligand binding pocket (LBP) of the respective receptor are used in major indications such as oncologic, metabolic, reproductive, and immunologic diseases [1, 2]. However, the success of these therapeutics is often limited by poor selectivity and resistance mechanisms that, in the worst case, reverse the antagonistic effect of a drug and promote disease [3, 4, 5]. Additionally, undesirable effects are promoted by the fact that both inhibitors and natural substrates share the same binding pocket. In recent years, two allosteric sites on the surface of several NRs, called activation function-2 (AF-2) and binding function-3 (BF-3), have been identified and considered as alternative sites for drug binding (Figure 1A). The AF-2 site corresponds to a protein-protein interaction surface for the binding of coactivator proteins essential for downstream signaling which renders it an attractive target for potential inhibitors. While the BF-3 site has been initially shown to allosterically regulate binding of coactivators to the AF-2 site [2, 6, 7, 8], it has been suggested as interaction surface for the engagement with chaperones that associate NRs [2, 9, 10]. In recent years, several hundreds of compounds have been identified to modulate NR activity through either of these allosteric sites at various receptors [11, 12, 13, 14, 15, 16, 17]. Selectivity testing in the mentioned projects was, if conducted, in most cases limited to a single other NR [18, 19, 20]. Since especially steroidal NRs such as androgen receptor (AR), estrogen receptors (ER), glucocorticoid receptor (GR), progesterone receptor (PR), and minearlocorticoid receptor (MR) share a common domain architecture as well as a similar fold regarding their ligand binding domain (LBD) , the selectivity concern for allosteric NR inhibitors remains (Figure 1B). For example, it has been reported that the AF-2 and BF-3 sites of AR and GR have a sequence identity of approximately 50% [9, 21]. Further on, drug-like mimetics of coactivator peptides at the AF-2 site have the potential to disrupt protein-protein interactions for multiple NRs and ultimately promote off-target toxicity [19, 22].

While several structures with co-crystallized ligands have been determined for the AR, targeting superficial binding sites such as AF-2 and BF-3 of other receptors remains a challenge due to their comparably large size, shallowness, and high flexibility [23, 24]. Knowledge regarding binding hotspots and distinct pharmacophores among structurally similar NRs is crucial for the design of effective and selective inhibitory compounds [24, 25, 26, 27]. In this regard, cosolvent molecular dynamics (MD) simulations are a suitable computational tool to determine hotspots and assess their druggability, as well as to obtain detailed information on potentially useful pharmacophores to improve drug potency and selectivity for the site of interest. In this simulation protocol, which was inspired by crystallographic observations of small fragments binding to protein surfaces, organic solvent molecules mimicking drug fragments are added to the aqueous phase to monitor and quantify their interaction with a protein. Compared to similar methods for binding site detection, this protocol depends less on the input structure, allows for conformational changes of the protein, and is generally more reliable due to the intrinsic treatment of protein flexibility and explicit solvation [24, 25, 28, 29]. Even though cosolvent simulations have been previously applied to study the allosteric sites of AR, ER$\alpha$, and ER$\beta$ the main objective of these studies was to proof the applicability of the simulation protocol. In two studies that considered the AR, researchers were able to detect both allosteric sites in the top hotspots and, based on an assessment regarding the maximally achievable binding affinity, the AF-2 site was deemed more druggable [24, 30, 31]. Another known drawback of solvent-exposed binding sites is the influence of water molecules on the recognition, efficacy and selectivity of ligands due to their potential displacement or mediation of ligand-protein interactions. Whether a water molecule can be favourably displaced depends on its environment in the respective binding site [32, 33, 34, 35]. The desolvation free energy of a water molecule can be estimated based on MD simulations followed by the quantitative assessment of the trajectory snapshots and can ultimately guide the design of novel compounds or improve scoring in virtual screening projects [34, 35, 36].

Here, we applied cosolvent MD simulations, hydration site prediction, and molecular docking to a set of eight NRs including AR, ER$\alpha$, ER$\beta$, GR, MR, PR, and the thyroid receptors $\alpha$ and $\beta$ (TR$\alpha$ and TR$\beta$). We assessed a large share of compounds reported to modulate NR signaling through either AF-2 or BF-3 of the respective receptor.

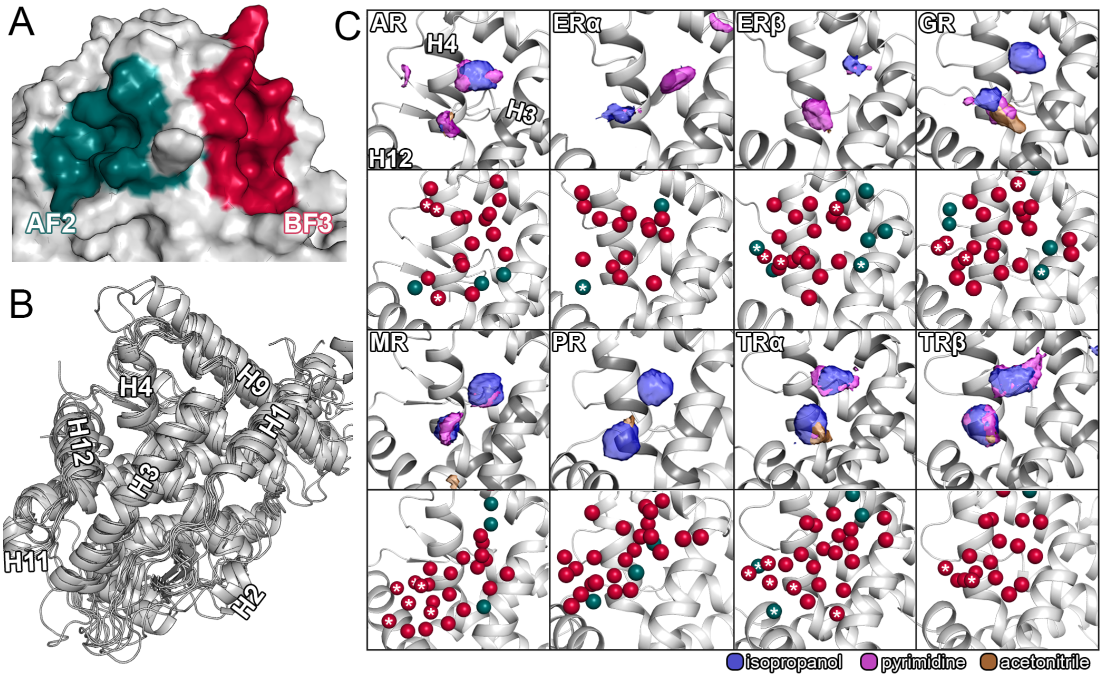


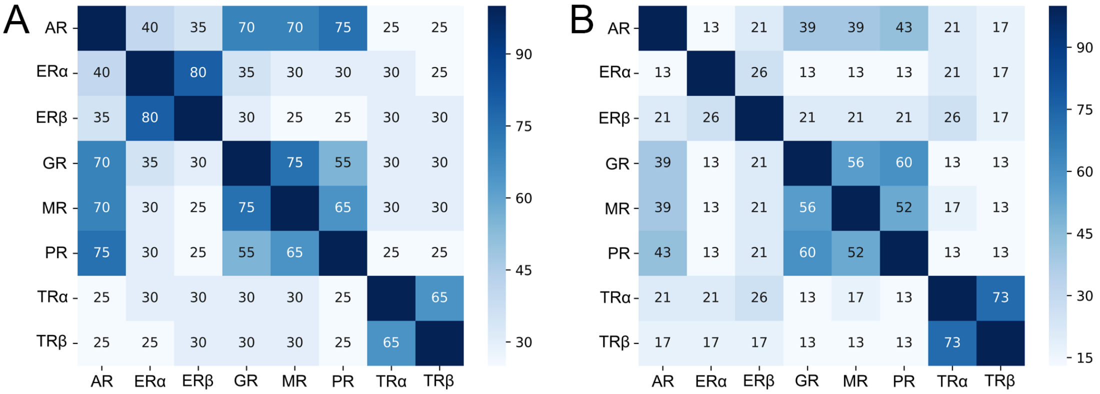
**Figure 1** Structural overview, results from cosolvent simulations, and hydration site prediction for the AF-2 site. (A) AF-2 and BF-3 sites of the androgen receptor (PDB ID: 3L3X). (B) Structural alignment of AR, ER$\alpha$, ER$\beta$, GR, MR, PR, TR$\alpha$, and TR$\alpha$. Secondary structure elements were assigned according to Tan and colleagues [37]. (C) For each receptor, the results of cosolvent simulations (upper part) and hydration site prediction (lower part) from WATSite for the AF-2 site are given. The color scheme for the cosolvent densities is given below the figure. The densities are shown at an isovalue of 12. Water molecules, that were found to be conserved based on the crystal structure analysis were colored in pine green and water molecules with a negative enthalpy ($\Delta$H $<$ -1.0 kcal/mol) were indicated with asterisks.

In contrast to previous works, our objective was the systematic determination of the main pharmacophores and positions of structural waters in order to compare them within our selection of human hormonal NRs to ultimately navigate the design of potent and selective inhibitors for each particular receptor.

## Results and Discussion

**Sequence Similarity Among Hormonal NRs.** The structures of LBD of AR, ER$\alpha$, ER$\beta$, GR, MR, PR, TR$\alpha$, and TR$\beta$ feature a similar conserved fold (Figure 1B). We conducted a sequence-based analysis of residues in the 5 Å range around a cocrystallized ligand and compared those residue regarding to their similarity among all receptors (Figure 2). While the analysis revealed similarities between the isoforms of the ERs and TRs that were expected, receptors with high identity in both sites included

AR, GR, MR, and PR. We observed values up to 75% between MR and GR as well as PR and AR in the AF-2, which raises serious concerns regarding off-target binding for ligands targeting either of these receptors. In slight contrast to previously reported conservation degree in the literature [2, 9], our comparably high percentages could be explained by the different definitions of conservation and binding site residues. Compared to the other receptors, the TRs offer a good potential for selective binding to either site, but especially for BF-3. Overall, the results suggest the BF-3 site to offer a higher potential for the design of selective inhibitors due to the generally lower values in similarity among the receptors compared to the AF-2. The conservation of residues from a three-dimensional perspective can be assessed in Figure S1 and S2.



A

| | AR | ERα | ERβ | GR | MR | PR | TRα | TRβ |
|---|---|---|---|---|---|---|---|---|
| AR | | 40 | 35 | 70 | 70 | 75 | 25 | 25 |
| ERα | 40 | | 80 | 35 | 30 | 30 | 30 | 25 |
| ERβ | 35 | 80 | | 30 | 25 | 25 | 30 | 30 |
| GR | 70 | 35 | 30 | | 75 | 55 | 30 | 30 |
| MR | 70 | 30 | 25 | 75 | | 65 | 30 | 30 |
| PR | 75 | 30 | 25 | 55 | 65 | | 25 | 25 |
| TRα | 25 | 30 | 30 | 30 | 30 | 25 | | 65 |
| TRβ | 25 | 25 | 30 | 30 | 30 | 25 | 65 | |

B

| | AR | ERα | ERβ | GR | MR | PR | TRα | TRβ |
|---|---|---|---|---|---|---|---|---|
| AR | | 13 | 21 | 39 | 39 | 43 | 21 | 17 |
| ERα | 13 | | 26 | 13 | 13 | 13 | 21 | 17 |
| ERβ | 21 | 26 | | 21 | 21 | 21 | 26 | 17 |
| GR | 39 | 13 | 21 | | 56 | 60 | 13 | 13 |
| MR | 39 | 13 | 21 | 56 | | 52 | 17 | 13 |
| PR | 43 | 13 | 21 | 60 | 52 | | 13 | 13 |
| TRα | 21 | 21 | 26 | 13 | 17 | 13 | | 73 |
| TRβ | 17 | 17 | 17 | 13 | 13 | 13 | 73 | |

**Figure 2** Sequence identity analysis of residues in (A) the AF-2 and (B) the BF-3 sites. The identity is given as percentage of the maximally achievable score based on the considered residues.

**Distinct Pharmacophores of the Allosteric Sites.** To this date, the AR is the most intensively studied NR regarding the development of allosteric inhibitors, due to its involvement in the genesis and progression of prostate cancer, which is one of the leading causes for cancer-related death in men. [2, 18]. Even though constitutively active splice variants of the AR lacking the LBD regularly arise in late stages of the disease, the AR LBD remains a drug target of high interest, especially in early stages of pharmacological treatment [2]. Likewise, efforts were put into the design of inhibitors against ER$\alpha$ since the majority of breast cancer cases depend on this receptor [14]. Unfortunately, currently available therapeutics often suffer from resistance mechanisms, in some cases only caused by as little as a single amino acid mutation in the LBP [4], which likely contributed to the number of works that applied cosolvent simulations to the allosteric sites of the AR and the ERs. In these studies, the AF-2 site was detected in both AR and ERs, while densities at the BF-3 site were only reviewed for the AR [24, 30, 31]. The aforementioned work inspired us to systematically apply this simulation protocol to eight NRs which are known to suffer from poor drug selectivity [2, 14, 17]. Based on the evaluation of our simulations, we were able to identify probe molecules binding to the AF-2 and BF-3 sites of all receptors, with the exception of the BF-3 site in ER$\alpha$ (Figures 1C and 3A). In accordance with the sequence analysis, the similarity among the AF-2 sites regarding the probe densities of all receptors along with the diversity of the individual BF-3 sites was one of the most apparent outcomes of our simulations. The results do not only reflect the preference of multiple NRs for similar coactivator sequences, a known concern for receptor selectivity [38], but also support the fact that a higher degree of selectivity could be achieved when targeting the BF-3 site over both the orthosteric pocket and the AF-2 site due to its uniqueness among the receptors. Clearly, the selectivity concerns regarding inhibitors interacting with the AF-2 site were justified because especially simulations of GR, MR, PR, and the the TRs presented a highly similar pattern of probe densities. However, despite the comparably high sequence similarity of GR, MR, and PR regarding the AF-2 site, the GR resulted in a notably higher density of acetonitrile which points towards a higher degree of amiphaticity that is favored there. Interestingly, the ERs not only displayed distinct differences to the other receptors, but also between themselves based on two isolated densities of isopropanol and pyrimidine that were interchanged between the two isoforms. Even though it is possible that compounds assume reversed binding modes in either receptor, such distinct differences offer potential to improve isoform selectivity, especially if structure-based design is employed. Less obviously, the density map of the ER$\alpha$ revealed a smaller third hotspot in the vicinity of V368 unique to this receptor suggesting this to be the reason for the selectivity differences regarding AR and ER$\alpha$ observed among particularly decorated inhibitors with a common pyrimidine core [10]. In consensus with that, the AR displaced an isolated density of pyrimidine towards H4 and a generally more pronounced aromatic density. The higher aromaticity of the AR AF-2 site compared to the ERs likely reflects its preference for phenylalanine or tryptophan residues as opposed to leucines in coactivator fragments [39]. Notably, simultaneous binding to multiple NRs can be desired in certain therapeutic scenarios as for the inhibition of the AR in breast cancer treatment, since it might adopt roles of ER$\alpha$ when the function of this receptor is absent due to pharmacological inhibition [40]. Therefore, compounds binding to both AR and ER$\alpha$ such as several ones identified by Gunther *et al.* might be beneficial depending on the therapeutic indication [2, 41]. While the literature suggests differences in the hydrophobic relief of the AF-2 to be responsible for TR isoform selectivity [16], our results only presented minor differences in any of the probe densities between TR$\alpha$ and TR$\beta$ apart from a slightly oblonged density for pyrimidine in the TR$\beta$.

As mentioned before, the BF-3 site of the studied NRs displayed a higher degree of heterogeneity regarding the cosolvent densities (Figure 3A). While AR, GR, and to some degree MR and PR showed a somehwat comparable pattern of probe densities, both ERs and TRs presented a high degree of diversity among them despite being most closely related based on their sequence. Most obviously, the BF-3 site of ER$\alpha$ was barely mapped by the probe molecules at the selected isovalue (twelve times the density in bulk solvent) and only presented a slight density of pyrimidine indicating a region for an aromatic moiety. The same region in ER$\beta$ was mapped by isopropanol suggesting the placement of an amphipathic as opposed to an aromatic functional group in this isoform, to achieve selective ligand-protein interactions. The lack of probe density at the ER$\alpha$ BF-3 site points towards poor druggability of the site [42], which is indirectly supported by experimental results since inhibitors directed against the AR did not inhibit ER$\alpha$ [18]. The density maps of the BF-3 site of both TRs substantially differ from the ones of other receptors, which reduces the odds for the cross-binding of compounds harboring the proposed density-based pharmacophores. Even though the TR$\beta$ shared densities for isopropanol and acetonitrile with the TR$\alpha$, the latter displays additional densities in distal regions of the site and a pronounced density of acetonitrile in the center. Therefore, an amphipathic group, potentially containing a nitrogen atom, would offer potential to increase compound selectivity between the two TRs. Moreover, the density maps revealed a high identity between AR and GR, especially regarding the regions mapped by isopropanol. In contrast, the MR and PR presented distinct differences despite the comparable degree of conservation among all four receptors.

A comparison of the AF-2 and BF-3 displayed density patterns that are shared among the two sites depending on the receptor. For example, the densities for the ER$\alpha$ AF-2 and PR BF-3 site showed a similar arrangement consisting of an amphipathic group coupled to an aromatic moiety. Furthermore, the BF-3 sites of both AR and GR showed a distinct resemblance to the AF-2 sites of most other receptors. The consideration of compounds designed for the AF-2 site to simultaneously interact with the BF-3 site and vice versa is not only supported by our results, but is also based on crystallographic data. Most interestingly, a crystal structure of the AR (PDB ID: 2YLP) [43] revealed two ligands concurrently bound to both allosteric sites. Therefore, a complete selectivity assessment should consider binding to the other allosteric site as well. In addition to the allosteric sites, we measured comparably high probe densities in several orthosteric sites, regions that were suggested to be involved in the access to the buried binding pocket of NRs, and other zones of the receptors [44]. For detailed review and to assist the development of novel compounds [45], we supply the complete density maps for every receptor in our supporting material. The root mean square deviation (RMSD) of the cosolvent simulations was assessed (Tables S1-S8) and presented deviations ranging from 0.81 to 2.57 Å confirming good conformational stability of the protein backbone throughout the simulations.

**Conformational Changes of the Allosteric Sites.** Even though it was suggested that the association of allosteric inhibitors is dependent on the presence of an agonist in the orthosteric site [46], we did not observe significant differences regarding the cosolvent densities between our apo and holo simulations (Figure S3 and S4). Potentially, a protocol with prolonged individual simulations or the application of biasing potentials might induce more pronounced changes, since conformational adaptations affecting the surface of the receptor have to occur over a long distance and naturally require substantial simulation efforts [29]. For example, association of inhibitors to the LBP has been shown to structurally modulate the AF-2 and its capability to interact with coactivator proteins, mainly by conformational change of helix-12 (H12) [10]. Combination therapy with multiple drugs is regularly applied in cancer pharmacotherapy [47, 48] and therefore potential synergistic effects of allosteric and orthosteric inhibitors will have to be considered in future studies. Likewise, the simultaneous treatment with AF-2 and

BF-3 inhibitors might produce mixed results, since binding of inhibitors to the BF-3 site is known to reduce the affinity of coactivator peptides in an allosteric mechanism and might affect a potential drug-drug synergy. In general, the AF-2 site of NRs is known to be capable of significant conformational changes as the example of the AR nicely underlines, since this receptor accepts a diverse set of coactivator fragments and has to structurally adapt in order to do so [39, 11]. Our examination of available AR crystal structures (Figure S5) revealed certain residues in both AF-2 and BF-3 capable of structural adaptation to ligand molecules. In particular, the residues K720, R726, and M734 in the AF-2 displayed various rotamers depending on the interaction partner. In the BF-3 site, Q670, F826, N727, E829, K836, and R840 appeared flexible, suggesting this site to exhibits a higher degree of flexibility.

To quantify the conformational change induced by different probe molecules in either allosteric site, we compared representative structures of our simulations in pure water to ones obtained from cosolvent simulations (Figure S6). In accordance with the above-mentioned flexibility in crystal structures, this analysis uncovered a higher degree of structural adaptation of the BF-3 as opposed to the AF-2. In this context, it is worth noting that several residues of the BF-3 site are located in close vicinity to the N-terminus, which would likely be more rigid due to its direct linkage to the DNA-binding domain that was not considered here. Probe molecules in cosolvent simulations have been shown to induce cryptic binding pockets, which can significantly contribute to drug selectivity and therefore knowledge on such pharmacophores can be instrumental for rational drug design. It is known that individual cosolvent molecules can cause distinct conformational adaptations of the respective protein [29, 49]. Indeed, our RMSD based analysis presented the highest degree of structural change with isopropanol as probe molecule suggesting the associated pharmacophores as promising to exploit the intrinsic flexibility of both allosteric pockets. Interestingly, we observed a different extent of adaptation from receptor to receptor with the ER$\alpha$ showing the most distinct changes. In the literature, residues with a high degree of flexibility have been reported for the AR and include K720, M734, N727, F826, E829, and F837 [50, 43]. Although our analysis detected several of these residues to be involved in a comparably large structural adaptation, we observed conformational changes of additional residues that have not been reported before. Most notably, the residue L685 in the GR (corresponding to F826 in the AR) located in the BF-3 site displayed a particularly high RMSD. A large share of allosteric NR inhibitors have been developed based on the interplay of computational screening and experimental characterization. Importantly, our results could support future virtual screening studies by recommending flexible residues in the binding site for molecular docking calculations. One limitation we experienced during this analysis was the truncation of termini in several crystal structures preventing a quantification of the RMSD for these regions and adjacent structural elements. The backbone RMSD analysis (Figure S7) of all triplicates simulations in pure water revealed excellent (AR, ER$\alpha$a, GR, PR, TR$\beta$) to sufficient (ER$\beta$, MR, TR$\alpha$) convergence.

**Displacement of Water Molecules from the Allosteric Sites.** Besides their often limited selectivity, the major drawback for the clinical application of allosteric inhibitors is their comparably low potency which is a typical characteristic of superficial protein-protein interaction inhibitors [51, 52]. Regarding the ER$\alpha$ it was proposed that the modest efficacy of small molecules designed to bind to the AF-2 surface is also associated with the lower amount of water molecules that are displaced by the inhibitor in contrast to the physiologic interaction partner [13]. As mentioned in the introduction, the importance of considering water molecules is well known in structure-based drug design because, in almost any case, waters are displaced during ligand association. Displacing a tightly bound water molecule generally results in a favorable gain of entropy and is a commonly used strategy exploited by medicinal chemists to improve compound efficacy and selectivity [32, 33, 53]. However, if such a displacement is favorable depends on enthalpic and entropic contributions that are determined by the environment of the water molecule. Ultimately, the desolvation free energy of a water molecule can be estimated based on these contributions and give valuable input for ligand design [35]. Especially for solvent-exposed binding sites, as they often are when protein-protein interactions are considered, contributions of water molecules can gain even higher importance [13].

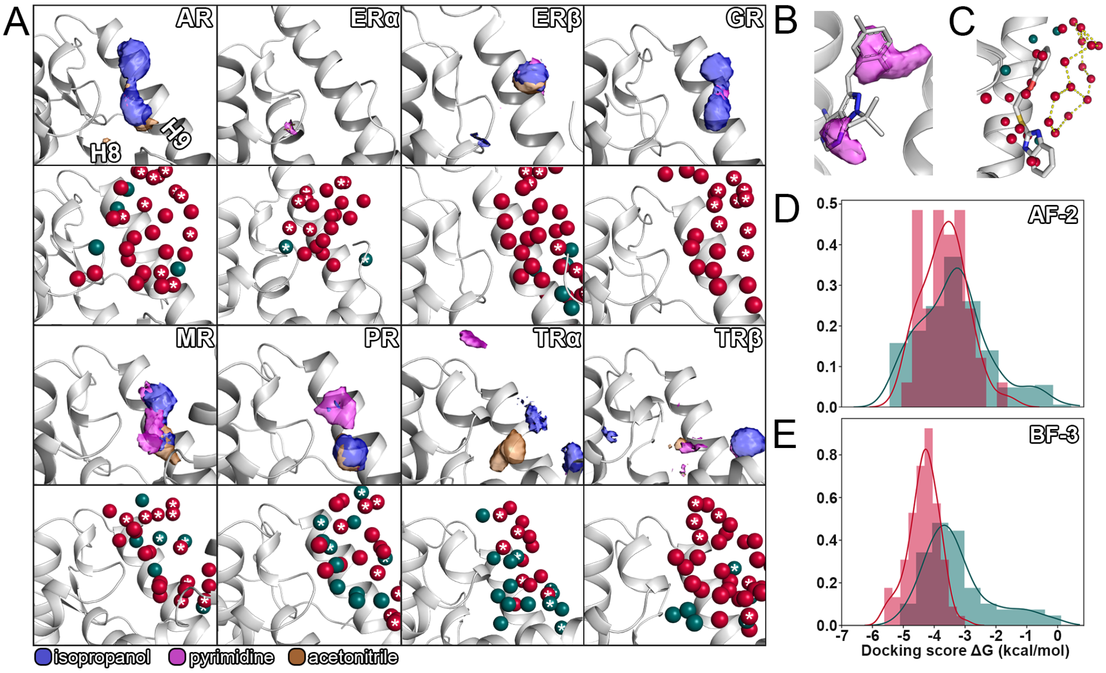


**Figure 3** Results from cosolvent simulations, hydration site prediction for the BF-3 site, and molecular docking.(A) For each receptor, the results of cosolvent simulations (upper part) and hydration site prediction (lower part) from WATsite for the AF-2 site are given. The color scheme for the cosolvent densities is given below the figure. The densities are shown at an isovalue of 12. Water molecules, that were found to be conserved based on a crystal structure analysis were colored in pine green and water molecules with a negative enthalpy ($\Delta H < -1.0$ kcal/mol) were indicated with asterisks. (B) Density of pyrimidine at the AF-2 overlapping with cocrystallized ligand (PDB ID: 2PIP) (C) Cluster of water molecules at the BF-3 of the AR. A cocrystallized ligand molecule is shown as comparison (PDB ID: 4HLW). Polar contacts were visualized in PyMol. (D) Distribution of docking scores of AR AF-2 inhibitors. Confirmed actives are shown in red, while the remaining compounds of the library colored pine green. (E) Distribution of docking scores of AR BF-3 inhibitors.

We evaluated the hydration sites for the selected set of NRs by two different approaches. The WATsite program allows to set up a restrained MD simulation in explicit water, which is post-processed using machine learning techniques to estimate the thermodynamic contributions of a water molecule to be displaced. Following this analysis, we retrieved available crystal structures for every receptor and determined conserved hydration sites among them (Figure S8) to ultimately combine both predictions and find a consensus. Although a large share of the water positions we determined were unique to each NR, there were sites shared by multiple receptors (Figure 1C, 3A, and S8). For example, we identified a conserved water molecule at the AF-2 site of both isoforms of the ER that could be favorably displaced based on its desolvation enthalpy (Figure S9C).

Another interesting example was a water molecule with a negative enthalpic contribution and conservation within crystal structures close to H12 that was predicted to occur in only one isoform of the TR (Figure 1C). Therefore, displacing this water molecule with a TR$\beta$ AF-2 inhibitor might increase the selectivity towards TR$\alpha$. In a similar fashion, displacing a conserved water molecule that occurs in AR, ER$\beta$, GR, and MR from a favorable environment might decrease binding of a compound to any of these receptors (Figure S9A). Remarkably, no water molecules with a negative enthalpic contribution were identified for the AF-2 site of the PR. At the BF-3, we observed a higher diversity of hydration sites compared to the AF-2 site following the previous trends regarding our conservation analysis and the cosolvent densities. However, we noticed a reoccurring network of water molecules in vicinity of the H9 N-terminus that, depending on the compound, might form a favorable first-shell hydration layer based on enthalpic contributions (Figure 3C) [36]. The simulations showed small backbone fluctuations values, which was to be expected due to the applied restraints on the protein atoms (Figure S10). Analogous to our cosolvent simulations, structure files resulting from both procedures used for hydration prediction together with a complete list of enthalpic and entropic contributions for each water molecule (Table S9-S16) are provided in the supporting material. These contributions can be used to estimate the gain in free energy of a particular ligand molecule by considering the water molecules it displaces in the bound state.

**Selectivity of Allosteric Inhibitors Explored by Molecular Docking.** The design of numerous allosteric inhibitors considered in this study was itself assisted by computational chemistry methods such as virtual screening [14, 18, 43, 54]. Molecular docking is an accepted technique with high throughput to explore off-target binding of potential drug compounds, preferably in an early stage of their development [55, 56]. Here, we retrieved more than 300 confirmed allosteric inhibitors from the literature and cross-docked them to investigate their potential to interact with the AF-2 and BF-3 of other NRs in our selection. Independent of the site towards the inhibitors were designed, they were docked to both allosteric sites, because most studies experimentally excluded a LBP-based mechanism, but did not distinguish between them [4, 14, 15]. As a common practice, we assessed the accuracy in pose prediction by redocking cocrystallized ligands to the respective site (Figure S11-14). The determined heavy-atom RMSD values between the poses reached from 1.25 to 7.46 Å for the standard precision (SP) docking protocol and from 0.84 to 8.22 Å for the extra precision (XP) protocol in Glide (Table S17) [57, 58]. A visual inspection of the obtained poses revealed a reversed orientation for multiple ligands at the BF-3 site, which might be explained by the symmetry of several compounds regarding their aromatic moieties as well as the increased conformational freedom at such solvent-exposed binding sites. In addition, the probe densities from our cosolvent simulations (Figure 1C and 3A) justify a reversed orientation of the main pharmacophores in certain cases. Another complication, potentially causing inaccuracies in pose prediction, is the presence of crystal mates in close vicinity to the cocrystallized ligand [59], which we determined in various crystal structures as exemplarily shown in Figure S15. Despite a more sophisticated scoring function and the considerably higher computational cost of the XP protocol, it did not offer any obvious improvement regarding pose prediction in the selected cases. Additionally, the Glide SP protocol was successfully used for the determination of inhibitors towards the AF-2 and BF-3 sites of the AR [50, 54, 13, 12] and we therefore selected it to evaluate our library of ligands. To further validate the performance of the chosen docking protocol, we generated a decoy ligand set of and determined the area under the curve of the Receiver Operator Characteristic (ROC AUC) for each compound group. Based on a poor score in this metric (Table S18), compounds designed towards the AF-2 site of TR$\alpha$ and TR$\beta$ were excluded from further investigation. An analysis of the score distribution was conducted for the remaining compounds series with more than ten entries, meaning ligands designed towards the AF-2 sites of AR and ER$\alpha$, as well as the BF-3 of the AR were docked into their target site to compare their scores to the ones of all other compounds in our ligand set. Between the allosteric sites of the AR, a distinct difference regarding the score distribution could be observed (Figure 3D and E). While the distribution of docking scores of the AF-2 compound series showed a high overlap, the curves displayed an astounding degree of separation for the BF-3 site. The largest share of compounds designed to interact with the BF-3 site of the AR showed an average improvement in binding free energy of approximately 1.0 kcal/mol with a high number of the remaining compounds scored below -3.0 kcal/mol. Again, these results suggest targeting the BF-3 site to offer a higher degree of selectivity compared to the AF-2 site, especially since the distribution of the ER$\alpha$ AF-2 compound series presented a similar pattern as the AR AF-2 series (Figure S16). The score distribution obtained from the XP docking protocol (Figure S17) confirmed the results obtained from the SP protocol.

Even though the majority of studies neglected off-target binding of their compounds, one study extensively evaluated their inhibitors against other NRs overlapping with our selection [14]. Although their ER$\alpha$ ligand showed reasonable selectivity against AR and PR, two receptors presenting a high degree of similarity throughout our work, it was shown to interact with the GR to a reasonable extent. Inspired by these results, we reviewed the docking poses of the compound and discovered a halogen bonding interaction [33] between a lysine residue, which is part of a so-called charge clamp, and the chlorine atom of the inhibitor shared by both GR and ER$\alpha$ (Figure S18). The charge clamps, flanking the hydrophobic subpockets of the AF-2 sites in various NRs, were often proposed as a selectivity factor [10]. The described interaction did not appear in neither AR or PR and we therefore suggest this specific interaction as a determinant for the selectivity of this compound.

## Materials and Methods

**Sequence Alignment and Analysis.** After a sequence alignment in the UGENE v1.32.0 [60] suite using the ClustalW algorithm [61] (Figure S19), we determined residues of either site based on a spherical zone around a co-crystallized ligand (PDB ID: 2YLP) that can bind both AF-2 and BF-3 in the AR. We then used an in-house python routine to determine the conservation of the selected residues as follows: identical residues were valued at 1.0, while the same residue group was scored at 0.5 (Table S19). The scores were summed up for 20 AF-2 residues and 23 BF-3 residues respectively to ultimately calculate a percentage value for the conservation.

**Ligand Preparation.** All ligands were retrieved from various publications that evaluated their compounds and provided evidence for binding to either of the allosteric sites (Table S20). Compounds were included if a reasonable biological activity ($IC_{50}$ or $K_i$ below 100 $\mu M$) was measured. Three-dimensional conformers were generated in the LigPrep panel [62] within the Maestro Small-Molecule Drug Discovery Suite 2019-3

[63] using the OPLS3e force field. The protonation states of the ligands were predicted using Epik [64] at physiological pH (pH = 7.4). The highest scored ligand conformations were selected, potential tautomerization was accounted for, and in the case of unspecified chiral centers both stereoisomers were considered.

*Protein Preparation.* The protein structures used in this study were retrieved from the Protein Data Bank (Table S21) and prepared using the Protein Preparation Wizard [65] within Maestro. In the case of missing loops, they were added based on complete template structures for ER$\alpha$, ER$\beta$, MR, and TR$\beta$. The amino acid sequence of the protein structures was compared to the sequence reported in the UniProt database [66] and, in the case of engineered amino acids, the sequence was manually corrected to represent the wild-type receptor sequence. For MD simulations, we aimed on selecting structures with physiologic ligands bound to the LBP and a resolution below 2.5 Å. In the case of less than five missing amino acids at the C-terminus to complete the sequence, these residues were manually added to the structure. While the N-terminus was modeled with an acetamide cap, since it would be further linked to the DNA-binding domain, the C-terminus was modeled as free carboxylic acid group. Ions and organic solvents were removed, before hydrogen atoms were added to the structures, the protonation state predicted at pH 7.4, and the hydrogen bonding network was oriented. As a last step, the structures were refined by the means of a restrained minimization using the OPLS3e force field with a RMSD convergence threshold of 0.30 Å.

**MD Simulations and Evaluation.** The simulations in pure water were conducted using the Desmond simulation engine (v.2019-3) [67]. Using the System Builder, the prepared protein structures were solvated with SPC water molecules in cubic periodic boundary system with a buffer of 10 Å to the next protein atom. Ions were added to neutralize the systems, before they were relaxed for 100 ps using the MD-based Desmond Minimization protocol. The simulations were conducted using the OPLS_2005 force field in an NPT ensemble combined with the Martyna-Tobias-Klein barostat with a relaxation time of 2.0 ps at 300 K and the Nose-Hoover thermostat with a relaxation time of 1.0 ps. The u-series algorithm was used to treat long range interactions with a cut-off of 9 Å for short range interactions [68]. By default, the M-SHAKE algorithm was applied to constrain bonds to hydrogen atoms. We left the time step for the RESPA integrator at 2.0 fs and files with atomic coordinates were saved at an interval of 4.8 ps. After the default relaxation protocol (Table S22), the simulations were carried out in triplicates for a duration of 40 ns per receptor at a temperature of 300 K and, to ensure a unique course of the individual trajectories, we generated random seeds for the initial velocities. The backbone RMSD of the pure water MD simulations were determined in the Simulation Interaction Diagram panel within Maestro.

For the cosolvent MD simulations, we used the Mixed Solvent MD workflow that comes with the Desmond simulation engine [67]. As probe molecules, isopropanol, acetonitrile, and pyrimidine at a concentration of 5% (by volume) were selected since these solvents are water-miscible, offer a low potential for aggregation, and therefore do not require the application of repulsive forces [27, 69]. In addition to the recommended simulation protocol with apo structures, we ran simulations with the cocrystallized ligand remaining in the orthosteric binding pocket. For the ER$\beta$ and GR, the water buffer parameter was increased from the default value of 12.0 to 15.0 as described in the provided documentation. The default relaxation protocol for this workflow (Table S23) was conducted, before the 5 ns production simulations were run at a temperature of 300 K in an NPT ensemble using the OPLS_2005 force field. For each probe molecule, ten simulations were conducted resulting in a cumulative simulation time of 1.2 $\mu$s per receptor. The remaining specifications were left on default. The backbone RMSD of the cosolvent MD simulations were determined based on the output frame of the protocol using an in-house python routine.

To quantify the conformational change of the AF-2 and BF-3 sites induced by the probe molecules, we determined the heavy-atom RMSD between representative structures of the cosolvent simulations and the pure water simulations of the individual residues located in the allosteric sites. First, we determined the representative structure of the cosolvent simulations by inputting the last frame of each simulation for each probe into the MaxCluster algorithm and selecting the structure with the highest rank according to the 3Djury score [70]. Similarly, we chose the last 30 frames of each pure water simulation of each receptor as input for MaxCluster to determine representative structures. Before determining the heavy atom RMSD using an in-house python routine, we superimposed the obtained structures.

The simulations to determine the hydration sites were performed using the WATsite 3.0 protocol [34, 35] that comes as a PyMol plugin [71]. The prepared structures of the eight receptors were used as input for the simulations and, since WATsite requires information about the location binding site, an AR ligand molecule known to bind both AF-2 and BF-3 was derived from a crystal structure (PDB ID: 2YLP) and superimposed to be located in the allosteric sites of the respective receptor. For each binding site, a separate simulation with an equilibration phase of 2 ns and a production stage of 20 ns at 298.15 K was run totaling to 352 ns of simulation time. We took the default timestep of 2.0 fs and frames with atomic coordinates collected every 2.0 ps. Long range interactions were treated with the Particle Mesh Ewald method, non-bonded interactions were cut off at 10 Å, and heavy atoms were restrained with a spring constant of 2.5 kcal/mol/Å$^2$. In the post-processing stage, we selected the DBSCAN clustering algorithm to determine the hydration sites and their occupancy. The backbone RMSD of the simulations was assessed using an in-house python routine.

**Crystal Structure Analysis.** For each receptor, all crystal structures with a resolution below 2.5 Å were retrieved from the Protein Data Bank and superimposed. Next, only the water molecules were kept in the structures and merged into a single PDB file that was used as input for an in-house python routine that determined the cluster centroids along with their occupancy using the the DBSCAN algorithm with an epsilon value of 0.9 and $n$ set to 2.0, similar to other protocols [72]. Clusters fulfilling the selected minimal occupancy criterion, depending on the number of input structures (Table S24), were considered as conserved and further compared to the prediction from WATsite using a distance threshold of 1.4 Å to establish a consensus between the two approaches.

**Molecular Docking.** We used the Glide protocol [57, 58] to dock the prepared ligands into the AF-2 and BF-3 sites of the selected panel of NRs. In the Recepor Grid Generation panel within Maestro, we defined the cubic grid box to be located at either site with an inner box size of 10 Å and an outer box size of 22.4 Å. In order to define the binding site, we superimposed an AR ligand molecule on each receptor in the Protein Structure Alignment panel [63]. All actives were grouped according to their target site and receptor before they were docked, using both SP and XP protocols of Glide, to all sites in the set. Also, we redocked known cocrystallized ligands and calculated the RMSD to the native pose in the Superposition panel in Maestro. In order to assess the reliability of the SP docking protocol, we further generated a decoy dataset for each compound group using the DUD-E webserver based on SMILES codes [73]. The ROC AUC metric, which characterizes if a randomly chosen known active molecule will rank higher than a randomly chosen decoy, was measured in the Enrichment Calculator within Maestro as described in detail in our previous work [63, 74]. Crystal mates were visually inspected in Maestro.

## Conclusions

Several allosteric inhibitors for the AR proved their efficacy in blocking AR signaling through experiments with cells or xenograft in vivo studies [2, 8, 54]. Besides limitations in potency, the clinical application of this interesting compound class is hampered by off-target binding due to high sequence identity among hormonal NRs. From this viewpoint, the BF-3 site displayed advantages over the AF-2 site as a drug target in our analysis focused on sequence identity, pharmacophores, and hydration sites. Further, we recommend intensive selectivity testing to a wider array of NRs, especially when inhibitors targeting the AF-2 site based on our results. Differences in probe densities reported in this study might be exploited to rationally design novel compounds and give insight into important structure-activity relationships. In our supplementary material, we provide the complete density maps obtained from the cosolvent simulations that can, for example, be incorporated in a pharmacophore-based screening campaign. Importantly, certain therapeutic scenarios might benefit by the concurrent binding to multiple NRs as we discussed regarding ER$\alpha$ inhibitors. In addition, future studies will have to consider potential synergistic effects of the simultaneous administration of orthosteric and allosteric inhibitors as well as combinations of inhibitors targeting the AF-2 and BF-3 sites concurrently. In our hydration site analysis, we identified water molecules that were conserved among multiple receptors including a reoccurring network of water molecules that formed an enthalpically favorable first-shell hydration layer around inhibitors at the BF-3 site. By the means of the provided supplementary files, the gain in desolvation free energy for a particular ligand can be estimated by accounting for the displaced waters in the bound state. By docking a large set of allosteric inhibitors, we demonstrated a modest accuracy of the applied protocol and suggest the inclusion of water molecules and protein flexibility into future predictions. Additionally, we suggest residues that could be considered in flexible docking calculations based on a quantification of the per-residue conformational adaptation in the presence of different cosolvent molecules. In conclusion, this work provides a foundation to refine both selectivity and potency of allosteric inhibitors in a rational manner. Improving these properties will likely increase the therapeutic applicability of this interesting compound class.

## References

[1] Michal Pawlak, Philippe Lefebvre, and Bart Staels. General molecular biology and architecture of nuclear receptors. *Current Topics in Medicinal Chemistry*, 12(6):486–504, 2012.

[2] Guillermo Martinez-Ariza and Christopher Hulme. Recent advances in allosteric androgen receptor inhibitors for the potential treatment of castration-resistant prostate cancer. *Pharmaceutical patent analyst*, 4(5):387–402, 2015.

[3] Wenqing Gao, Casey E. Bohl, and James T. Dalton. Chemistry and structural biology of androgen receptor. *Chemical Reviews*, 105(9):3352–3370, 2005.

[4] Fuqiang Ban, Eric Leblanc, Huifang Li, Ravi S N Munuganti, Kate Frewin, Paul S Rennie, and Artem Cherkasov. Discovery of 1 H-indole-2-carboxamides as novel inhibitors of the androgen receptor binding function 3 (BF3). *Journal of Medicinal Chemistry*, 57(15): 6867–6872, 2014.

[5] Jillian R Gunther, Terry W Moore, Margaret L Collins, and John A Katzenellenbogen. Amphipathic benzenes are designed inhibitors of the estrogen receptor alpha/steroid receptor coactivator interaction., 5 2008.

[6] E Estebanez-Perpina, L A Arnold, P Nguyen, E D Rodrigues, E Mar, R Bateman, P Pallai, K M Shokat, J D Baxter, R K Guy, P Webb, and R J Fletterick. A surface on the androgen receptor that allosterically regulates coactivator binding. *Proceedings of the National Academy of Sciences*, 104(41):16074–16079, 2007.

[7] Laura Caboni, Gemma K Kinsella, Fernando Blanco, Darren Fayne, William N Jagoe, Miriam Carr, D Clive Williams, Mary J Meegan, and David G Lloyd. ”True” antiandrogens-selective non-ligand-binding pocket disruptors of androgen receptor-coactivator interactions: Novel tools for prostate cancer. *Journal of Medicinal Chemistry*, 55(4):1635–1644, 2012.

[8] Ravi S.N. Munuganti, Mohamed D.H. Hassona, Eric Leblanc, Kate Frewin, Kriti Singh, Dennis Ma, Fuqiang Ban, Michael Hsing, Hans Adomat, Nada Lallous, Christophe Andre, Jon Paul Selvam Jonadass, Amina Zoubeidi, Robert N. Young, Emma Tomlinson Guns,

Paul S. Rennie, and Artem Cherkasov. Identification of a Potent Antiandrogen that Targets the BF3 Site of the Androgen Receptor and Inhibits Enzalutamide-Resistant Prostate Cancer. *Chemistry and Biology*, 21(11):1476–1485, 11 2014.

[9] Víctor Buzón, Laia R. Carbó, Sara B. Estruch, Robert J. Fletterick, and Eva Estébanez-Perpiñ. A conserved surface on the ligand binding domain of nuclear receptors for allosteric control. *Molecular and Cellular Endocrinology*, 348(2):394–402, 2012.

[10] Eric Biron and François Bédard. Recent progress in the development of protein-protein interaction inhibitors targeting androgen receptor-coactivator binding in prostate cancer, 7 2016.

[11] Alice L Rodriguez, Anobel Tamrazi, Margaret L Collins, and John A Katzenellenbogen. Design, Synthesis, and in Vitro Biological Evaluation of Small Molecule Inhibitors of Estrogen Receptor $\alpha$ Coactivator Binding. *Journal of Medicinal Chemistry*, 47(3):600–611, 1 2004.

[12] Kriti Singh, Ravi Shashi Nayana Munuganti, Eric Leblanc, Yu Lun Lin, Euphemia Leung, Nada Lallous, Miriam Butler, Artem Cherkasov, and Paul S Rennie. In silico discovery and validation of potent small-molecule inhibitors targeting the activation function 2 site of human oestrogen receptor alpha. *Breast cancer research : BCR*, 17:27, 2 2015.

[13] Aiming Sun, Terry W Moore, Jillian R Gunther, Mi-Sun Kim, Eric Rhoden, Yuhong Du, Haian Fu, James P Snyder, and John A Katzenellenbogen. Discovering small-molecule estrogen receptor $\alpha$/coactivator binding inhibitors: high-throughput screening, ligand development, and models for enhanced potency. *ChemMedChem*, 6(4):654–666, 4 2011.

[14] Kriti Singh, Ravi S.N. Munuganti, Nada Lallous, Kush Dalal, Ji Soo Yoon, Aishwariya Sharma, Takeshi Yamazaki, Artem Cherkasov, and Paul S. Rennie. Benzothiophenone derivatives targeting mutant forms of estrogen receptor-$\alpha$ in hormone-resistant breast cancers. *International Journal of Molecular Sciences*, 19(2), 2018.

[15] Leggy A. Arnold, Eva Estebanez-Perpina, Marie Togashi, Natalia Jouravel, Anang Shelat, Andrea C. McReynolds, Ellena Mar, Phuong Nguyen, John D. Baxter, Robert J. Fletterick, Paul Webb, and R. Kiplin Guy. Discovery of small molecule inhibitors of the interaction of the thyroid hormone receptor with transcriptional coregulators. *Journal of Biological Chemistry*, 280(52):43048–43055, 2005.

[16] Yeon Hwang Jong, Leggy A. Arnold, Fangyi Zhu, Aaron Kosinski, Thomas J. Mangano, Vincent Setola, Bryan L. Roth, and R. Kiplin Guy. Improvement of pharmacological properties of irreversible thyroid receptor coactivator binding inhibitors. *Journal of Medicinal Chemistry*, 52(13):3892–3901, 2009.

[17] Jong Yeon Hwang, Wenwei Huang, Leggy A. Arnold, Ruili Huang, Ramy R. Attia, Michele Connelly, Jennifer Wichterman, Fangyi Zhu, Indre Augustinaite, Christopher P.

Austin, James Inglese, Ronald L. Johnson, and R. Kiplin Guy. Methylsulfonylnitrobenzoates, a new class of irreversible inhibitors of the interaction of the thyroid hormone receptor and its obligate coactivators that functionally antagonizes thyroid hormone. *Journal of Biological Chemistry*, 286(14):11895–11908, 2011.

[18] Ravi Shashi Nayana Munuganti, Eric Leblanc, Peter Axerio-Cilies, Christophe Labriere, Kate Frewin, Kriti Singh, Mohamed D H Hassona, Nathan A Lack, Huifang Li, Fuqiang Ban, Emma Tomlinson Guns, Robert Young, Paul S Rennie, and Artem Cherkasov. Targeting the binding function 3 (BF3) site of the androgen receptor through virtual screening. 2. Development of 2-((2-phenoxyethyl) thio)-1H-benzimidazole derivatives. *Journal of Medicinal Chemistry*, 56(3):1136–1148, 2013.

[19] Preethi Ravindranathan, Tae Kyung Lee, Lin Yang, Margaret M. Centenera, Lisa Butler, Wayne D. Tilley, Jer Tsong Hsieh, Jung Mo Ahn, and Ganesh V. Raj. Peptidomimetic targeting of critical androgen receptor-coregulator interactions in prostate cancer. *Nature Communications*, 4(May), 2013.

[20] Alexander A Parent, Jillian R Gunther, and John A Katzenellenbogen. Blocking estrogen signaling after the hormone: pyrimidine-core inhibitors of estrogen receptor-coactivator binding. *Journal of medicinal chemistry*, 51(20):6512–6530, 10 2008.

[21] Nada Lallous, Kush Dalal, Artem Cherkasov, and Paul S. Rennie. Targeting alternative sites on the androgen receptor to treat Castration-Resistant Prostate Cancer. *International Journal of Molecular Sciences*, 14(6):12496–12519, 2013.

[22] Subhamoy Dasgupta, David M. Lonard, and Bert W. O'Malley. Nuclear Receptor Coactivators: Master Regulators of Human Health and Disease. *Annual Review of Medicine*, 65 (1):279–292, 2014.

[23] Ratna Rajesh Thangudu, Stephen H Bryant, Anna R Panchenko, and Thomas Madej. Modulating protein-protein interactions with small molecules: the importance of binding hotspots. *Journal of molecular biology*, 415(2):443–453, 1 2012.

[24] Yaw Sing Tan, David R. Spring, Chris Abell, and Chandra S. Verma. The Application of Ligand-Mapping Molecular Dynamics Simulations to the Rational Design of Peptidic Modulators of Protein-Protein Interactions. *Journal of Chemical Theory and Computation*, 11(7):3199–3210, 2015.

[25] Wenbo Yu, Sirish Kaushik Lakkaraju, E. Prabhu Raman, Lei Fang, and Alexander D. Mackerell. Pharmacophore modeling using site-identification by ligand competitive saturation (SILCS) with multiple probe molecules. *Journal of Chemical Information and Modeling*, 55(2):407–420, 2015.

[26] Sirish Kaushik Lakkaraju, Wenbo Yu, E Prabhu Raman, Alena V Hershfeld, Lei Fang, Deepak A Deshpande, and Alexander D MacKerell. Mapping Functional Group Free

Energy Patterns at Protein Occluded Sites: Nuclear Receptors and G-Protein Coupled Receptors. *Journal of Chemical Information and Modeling*, 55(3):700–708, 3 2015.

[27] Chao-Yie Yang and Shaomeng Wang. Hydrophobic Binding Hot Spots of Bcl-xL Protein-Protein Interfaces by Cosolvent Molecular Dynamics Simulation. *ACS medicinal chemistry letters*, 2(4):280–284, 4 2011.

[28] Shota Uehara and Shigenori Tanaka. Cosolvent-Based Molecular Dynamics for Ensemble Docking: Practical Method for Generating Druggable Protein Conformations. *Journal of chemical information and modeling*, 57(4):742–756, 4 2017.

[29] Amr H. Mahmoud, Ying Yang, and Markus A. Lill. Improving Atom-Type Diversity and Sampling in Cosolvent Simulations Using $\lambda$-Dynamics. *Journal of Chemical Theory and Computation*, 15(5):3272–3287, 2019.

[30] Phani Ghanakota and Heather A. Carlson. Moving beyond Active-Site Detection: MixMD Applied to Allosteric Systems. *Journal of Physical Chemistry B*, 120(33):8685–8695, 2016.

[31] Jesus Seco, F. Javier Luque, and Xavier Barril. Binding site detection and druggability index from first principles. *Journal of Medicinal Chemistry*, 52(8):2363–2371, 2009.

[32] Sarah E. Graham, Richard D. Smith, and Heather A. Carlson. Predicting Displaceable Water Sites Using Mixed-Solvent Molecular Dynamics. *Journal of Chemical Information and Modeling*, 58(2):305–314, 2018.

[33] Caterina Bissantz, Bernd Kuhn, and Martin Stahl. A Medicinal Chemist's Guide to Molecular Interactions. *J. Med. Chem.*, 53(16):6241–6241, 2010.

[34] Ying Yang, Bingjie Hu, and Markus A Lill. WATsite2.0 with PyMOL Plugin: Hydration Site Prediction and Visualization BT - Protein Function Prediction: Methods and Protocols. pages 123–134. Springer New York, New York, NY, 2017. ISBN 978-1-4939-7015-5.

[35] Bingjie Hu and Markus A. Lill. WATsite: Hydration site prediction program with PyMOL interface. *Journal of Computational Chemistry*, 35(16):1255–1260, 2014.

[36] Markus Lill, Ying Yang, Amr Mahmoud, Matthew Masters, and Analytics ChemRxiv Preprint. Elucidating the Multiple Roles of Hydration in Protein-Ligand Binding via Layerwise Relevance Propagation and Big Data Analytics: Elucidating the Multiple Roles of Hydration in Protein-Ligand Binding via Layerwise Relevance Propagation and Big Data. pages 1–17, 2019.

[37] Mh Eileen Tan, Jun Li, H. Eric Xu, Karsten Melcher, and Eu Leong Yong. Androgen receptor: Structure, role in prostate cancer and drug discovery. *Acta Pharmacologica Sinica*, 36(1):3–23, 2015.

[38] Nerea Gallastegui and Eva Estébanez-Perpiñá. Thinking Outside the Box: Alternative Binding Sites in the Ligand Binding Domain of Nuclear Receptors BT - Nuclear Receptors: From Structure to the Clinic. pages 179–203. Springer International Publishing, Cham, 2015. ISBN 978-3-319-18729-7.

[39] Yangguang Liu, Meng Wu, Tianqi Wang, Yongli Xie, Xiangling Cui, Liujun He, Yang He, Xiaoyu Li, Mingliang Liu, Laixing Hu, Shan Cen, and Jinming Zhou. Structural Based Screening of Antiandrogen Targeting Activation Function-2 Binding Site , 2018.

[40] Rachel Bleach and Marie McIlroy. The divergent function of androgen receptor in breast cancer; analysis of steroid mediators and tumor intracrinology. *Frontiers in Endocrinology*, 9(OCT):1–19, 2018.

[41] Jillian R. Gunther, Alexander A. Parent, and John A. Katzenellenbogen. Alternative inhibition of androgen receptor signaling: Peptidomimetic pyrimidines as direct androgen receptor/coactivator disruptors. *ACS Chemical Biology*, 4(6):435–440, 2009.

[42] Phani Ghanakota, Debarati DasGupta, and Heather A Carlson. Free Energies and Entropies of Binding Sites Identified by MixMD Cosolvent Simulations. *Journal of Chemical Information and Modeling*, 59(5):2035–2045, 5 2019.

[43] Nathan A. Lack, Peter Axerio-Cilies, Peyman Tavassoli, Frank Q. Han, Ka Hong Chan, Clementine Feau, Eric LeBlanc, Emma Tomlinson Guns, R. Kiplin Guy, Paul S. Rennie, and Artem Cherkasov. Targeting the binding function 3 (BF3) site of the human androgen receptor through virtual screening. *Journal of Medicinal Chemistry*, 54(24):8563–8573, 2011.

[44] André Fischer, Martin Smiesko, and Martin Smieško. Ligand Pathways in Nuclear Receptors. *Journal of Chemical Information and Modeling*, 59(7):3100–3109, 2019.

[45] Phani Ghanakota and Heather A. Carlson. Driving Structure-Based Drug Discovery through Cosolvent Molecular Dynamics. *Journal of Medicinal Chemistry*, 59(23):10383–10399, 2016.

[46] Dalei Shao, Thomas J. Berrodin, Eric Manas, Diane Hauze, Robert Powers, Ashok Bapat, Daniel Gonder, Richard C. Winneker, and Donald E. Frail. Identification of novel estrogen receptor $\alpha$ antagonists. *Journal of Steroid Biochemistry and Molecular Biology*, 88(4-5): 351–360, 2004.

[47] Zenita Adhireksan, Giulia Palermo, Tina Riedel, Zhujun Ma, Reyhan Muhammad, Ursula Rothlisberger, Paul J. Dyson, and Curt A. Davey. Allosteric cross-talk in chromatin can mediate drug-drug synergy. *Nature Communications*, 8:1–11, 2017.

[48] Reza Bayat Mokhtari, Tina S. Homayouni, Narges Baluch, Evgeniya Morgatskaya, Sushil Kumar, Bikul Das, and Herman Yeger. Combination therapy in combating cancer. *Oncotarget*, 8(23):38022–38043, 2015.

[49] S. Roy Kimura, Hai Peng Hu, Anatoly M. Ruvinsky, Woody Sherman, and Angelo D. Favia. Deciphering Cryptic Binding Sites on Proteins by Mixed-Solvent Molecular Dynamics. *Journal of Chemical Information and Modeling*, 57(6):1388–1401, 2017.

[50] Peter Axerio-Cilies, Nathan A Lack, M Ravi Shashi Nayana, Ka Hong Chan, Anthony Yeung, Eric Leblanc, Emma S.Tomlinson Guns, Paul S Rennie, and Artem Cherkasov. Inhibitors of androgen receptor activation function-2 (AF2) site identified through virtual screening. *Journal of Medicinal Chemistry*, 54(18):6197–6205, 2011.

[51] Hai-Bing Zhou, Margaret L Collins, Jillian R Gunther, John S Comninos, and John A Katzenellenbogen. Bicyclo[2.2.2]octanes: close structural mimics of the nuclear receptor-binding motif of steroid receptor coactivators. *Bioorganic and medicinal chemistry letters*, 17(15):4118–4122, 8 2007.

[52] Timothy R. Geistlinger and R. Kiplin Guy. Novel selective inhibitors of the interaction of individual nuclear hormone receptors with a mutually shared steroid receptor coactivator 2. *Journal of the American Chemical Society*, 125(23):6852–6853, 2003.

[53] Joel Wahl and Martin Smieško. Thermodynamic Insight into the Effects of Water Displacement and Rearrangement upon Ligand Modifications using Molecular Dynamics Simulations. *ChemMedChem*, 13(13):1325–1335, 2018.

[54] Nada Lallous, Eric Leblanc, Ravi S N Munuganti, Mohamed D H Hassona, Nader Al Nakouzi, Shannon Awrey, Helene Morin, Mani Roshan-Moniri, Kriti Singh, Sam Lawn, Takeshi Yamazaki, Hans H Adomat, Christophe Andre, Mads Daugaard, Robert N Young, Emma S Tomlinson Guns, Paul S Rennie, and Artem Cherkasov. Targeting Binding Function-3 of the Androgen Receptor Blocks Its Co-Chaperone Interactions, Nuclear Translocation, and Activation. *Molecular cancer therapeutics*, 15(12):2936–2945, 12 2016.

[55] Angelo Vedani, Max Dobler, Zhenquan Hu, and Martin Smieško. OpenVirtualToxLab-A platform for generating and exchanging in silico toxicity data. *Toxicology Letters*, 232(2): 519–532, 2015.

[56] Lei Xie, Li Xie, and Philip E Bourne. Structure-based systems biology for analyzing off-target binding. *Current opinion in structural biology*, 21(2):189–199, 4 2011.

[57] Thomas A. Halgren, Robert B. Murphy, Richard A. Friesner, Hege S. Beard, Leah L. Frye, W. Thomas Pollard, and Jay L. Banks. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. *Journal of Medicinal Chemistry*, 47(7):1750–1759, 2004.

[58] Richard A. Friesner, Robert B. Murphy, Matthew P. Repasky, Leah L. Frye, Jeremy R. Greenwood, Thomas A. Halgren, Paul C. Sanschagrin, and Daniel T. Mainz. Extra precision glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *Journal of Medicinal Chemistry*, 49(21):6177–6196, 2006.

[59] Matthew P Repasky, Robert B Murphy, Jay L Banks, Jeremy R Greenwood, Ivan Tubert-Brohman, Sathesh Bhat, and Richard A Friesner. Docking performance of the glide program as evaluated on the Astex and DUD datasets: a complete set of glide SP results and selected results for a new scoring function integrating WaterMap and glide. *Journal of Computer-Aided Molecular Design*, 26(6):787–799, 2012.

[60] Konstantin Okonechnikov, Olga Golosova, Mikhail Fursov, Alexey Varlamov, Yuri Vaskin, Ivan Efremov, O. G. German Grehov, Denis Kandrov, Kirill Rasputin, Maxim Syabro, and Timur Tleukenov. Unipro UGENE: A unified bioinformatics toolkit. *Bioinformatics*, 28 (8):1166–1167, 2012.

[61] Julie D. Thompson, Desmond G. Higgins, and Toby J. Gibson. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22(22): 4673–4680, 1994.

[62] Schrödinger LCC. LigPrep 2019-3. 2019.

[63] Schrödinger LCC. Maestro Small-Molecule Drug Discovery Suite 2019-3. 2019.

[64] Jeremy R. Greenwood, David Calkins, Arron P. Sullivan, and John C. Shelley. Towards the comprehensive, rapid, and accurate prediction of the favorable tautomeric states of drug-like molecules in aqueous solution. *Journal of Computer-Aided Molecular Design*, 24(6-7):591–604, 2010.

[65] G. Madhavi Sastry, Matvey Adzhigirey, Tyler Day, Ramakrishna Annabhimoju, and Woody Sherman. Protein and ligand preparation: Parameters, protocols, and influence on virtual screening enrichments. *Journal of Computer-Aided Molecular Design*, 27(3): 221–234, 2013.

[66] Alex Bateman, Maria Jesus Martin, Claire O'Donovan, Michele Magrane, Emanuele Alpi, Ricardo Antunes, Benoit Bely, Mark Bingley, Carlos Bonilla, Ramona Britto, Borisas Bursteinas, Hema Bye-AJee, Andrew Cowley, Alan Da Silva, Maurizio De Giorgi, Tunca Dogan, Francesco Fazzini, Leyla Garcia Castro, Luis Figueira, Penelope Garmiri, George Georghiou, Daniel Gonzalez, Emma Hatton-Ellis, Weizhong Li, Wudong Liu, Rodrigo Lopez, Jie Luo, Yvonne Lussi, Alistair MacDougall, Andrew Nightingale, Barbara Palka, Klemens Pichler, Diego Poggioli, Sangya Pundir, Luis Pureza, Guoying Qi, Steven Rosanoff, Rabie Saidi, Tony Sawford, Aleksandra Shypitsyna, Elena Speretta, Edward Turner, Nidhi Tyagi, Vladimir Volynkin, Tony Wardell, Kate Warner, Xavier Watkins, Rossana Zaru, Hermann Zellner, Ioannis Xenarios, Lydie Bougueleret, Alan Bridge, Sylvain Poux, Nicole Redaschi, Lucila Aimo, Ghislaine ArgoudPuy, Andrea Auchincloss, Kristian Axelsen, Parit Bansal, Delphine Baratin, Marie Claude Blatter, Brigitte Boeckmann, Jerven Bolleman, Emmanuel Boutet, Lionel Breuza, Cristina Casal-Casas, Edouard De Castro, Elisabeth Coudert, Beatrice Cuche, Mikael Doche, Dolnide Dornevil, Severine Duvaud, Anne Estreicher, Livia Famiglietti, Marc Feuermann, Elisabeth Gasteiger,

Sebastien Gehant, Vivienne Gerritsen, Arnaud Gos, Nadine Gruaz-Gumowski, Ursula Hinz, Chantal Hulo, Florence Jungo, Guillaume Keller, Vicente Lara, Philippe Lemercier, Damien Lieberherr, Thierry Lombardot, Xavier Martin, Patrick Masson, Anne Morgat, Teresa Neto, Nevila Nouspikel, Salvo Paesano, Ivo Pedruzzi, Sandrine Pilbout, Monica Pozzato, Manuela Pruess, Catherine Rivoire, Bernd Roechert, Michel Schneider, Christian Sigrist, Karin Sonesson, Sylvie Staehli, Andre Stutz, Shyamala Sundaram, Michael Tognolli, Laure Verbregue, Anne Lise Veuthey, Cathy H. Wu, Cecilia N. Arighi, Leslie Arminski, Chuming Chen, Yongxing Chen, John S. Garavelli, Hongzhan Huang, Kati Laiho, Peter McGarvey, Darren A. Natale, Karen Ross, C. R. Vinayaka, Qinghua Wang, Yuqi Wang, Lai Su Yeh, and Jian Zhang. UniProt: The universal protein knowledgebase. *Nucleic Acids Research*, 45(D1):D158–D169, 2017.

[67] Kevin Bowers, Edmond Chow, Huafeng Xu, Ron Dror, Michael Eastwood, Brent Gregersen, John Klepeis, Istvan Kolossvary, Mark Moraes, Federico Sacerdoti, John Salmon, Yibing Shan, and David Shaw. Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters. *ACM/IEEE SC 2006 Conference (SC'06)*, (November):43, 2006.

[68] David E. Shaw, J. P. Grossman, Joseph A. Bank, Brannon Batson, J. Adam Butts, Jack C. Chao, Martin M. Deneroff, Ron O. Dror, Amos Even, Christopher H. Fenton, Anthony Forte, Joseph Gagliardo, Gennette Gill, Brian Greskamp, C. Richard Ho, Douglas J. Ierardi, Lev Iserovich, Jeffrey S. Kuskin, Richard H. Larson, Timothy Layman, Li Siang Lee, Adam K. Lerer, Chester Li, Daniel Killebrew, Kenneth M. Mackenzie, Shark Yeuk Hai Mok, Mark A. Moraes, Rolf Mueller, Lawrence J. Nociolo, Jon L. Peticolas, Terry Quan, Daniel Ramot, John K. Salmon, Daniele P. Scarpazza, U. Ben Schafer, Naseer Siddique, Christopher W. Snyder, Jochen Spengler, Ping Tak Peter Tang, Michael Theobald, Horia Toma, Brian Towles, Benjamin Vitale, Stanley C. Wang, and Cliff Young. Anton 2: Raising the Bar for Performance and Programmability in a Special-Purpose Molecular Dynamics Supercomputer. *International Conference for High Performance Computing, Networking, Storage and Analysis, SC*, 2015-Janua(January):41–53, 2014.

[69] Lucas A. Defelipe, Juan Pablo Arcon, Carlos P. Modenutti, Marcelo A. Marti, Adrián G. Turjanski, and Xavier Barril. Solvents to fragments to drugs: MD applications in drug design. *Molecules*, 23(12):1–14, 2018.

[70] Angel R Ortiz, Charlie E M Strauss, and Osvaldo Olmea. MAMMOTH (Matching molecular models obtained from theory): An automated method for model comparison. *Protein Science*, 11(11):2606–2621, 11 2002.

[71] Schrodinger LLC and Schrodinger LLC. The PyMOL Molecular Graphics Development Component, Version 1.8, 2015.

[72] Marko Jukič, Janez Konc, Stanislav Gobec, and Dušanka Janežič. Identification of Con-

served Water Sites in Protein Structures for Drug Design. *Journal of Chemical Information and Modeling*, 57(12):3094–3103, 2017.

[73] Michael M Mysinger, Michael Carchia, John. J Irwin, and Brian K Shoichet. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *Journal of Medicinal Chemistry*, 55(14):6582–6594, 7 2012.

[74] Joel Wahl and Martin Smieško. Endocrine disruption at the androgen receptor: Employing molecular dynamics and docking for improved virtual screening and toxicity prediction. *International Journal of Molecular Sciences*, 19(6), 2018.

## 8.1 Supporting Information

## Supporting Results and Discussion

### Sequence Similarity Among Hormonal NRs

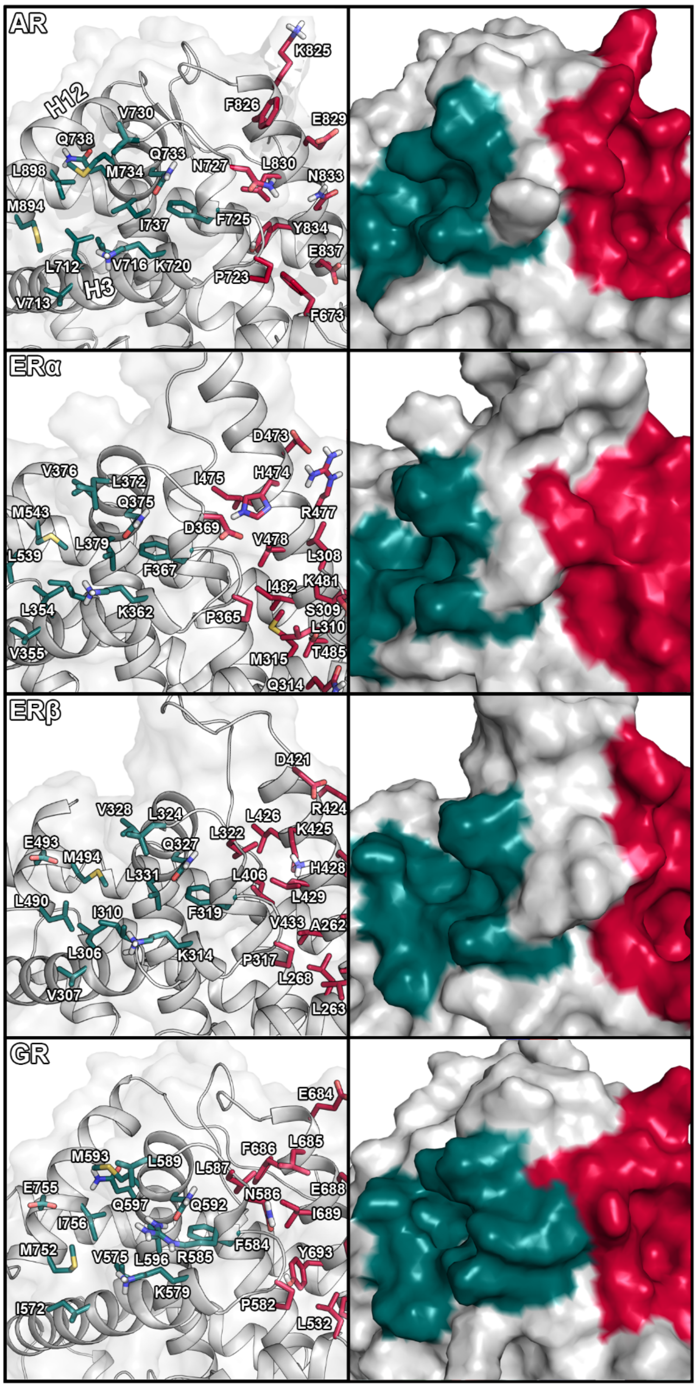


**Figure S 1** Residues and surface representation of AF-2 and BF-3 sites for AR, ER$\alpha$, ER$\beta$, and GR. The AF-2 is shown in pine green, while the BF-3 site was colored red. The surface was colored according to the type of residue (blue, positive charge; red, negative charge; green, non-polar; yellow, cysteine; purple, glycine; light blue, histidine).

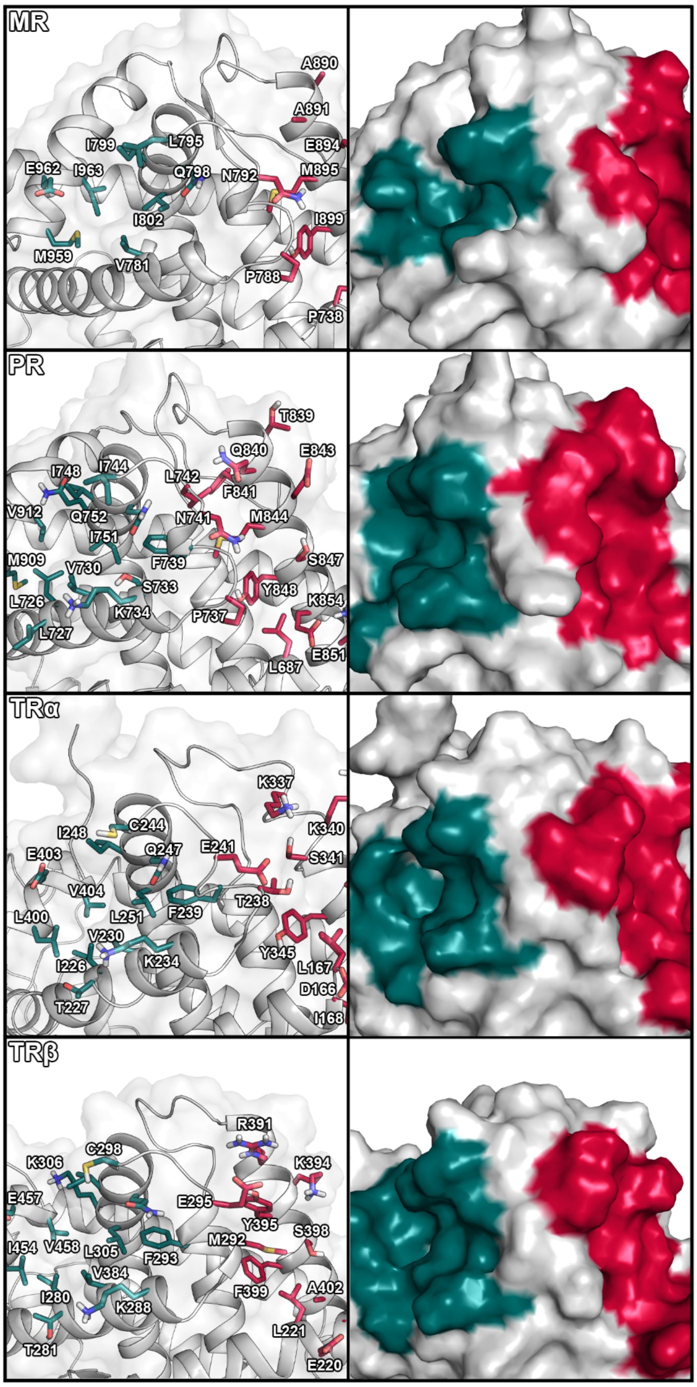


**Figure S 2** Residues and surface representation of AF-2 and BF-3 sites for MR, PR, TR$\alpha$, and TR$\beta$. The AF-2 is shown in pine green, while the BF-3 site was colored red. The surface was colored according to the type of residue (blue, positive charge; red, negative charge; green, non-polar; yellow, cysteine; purple, glycine; light blue, histidine).

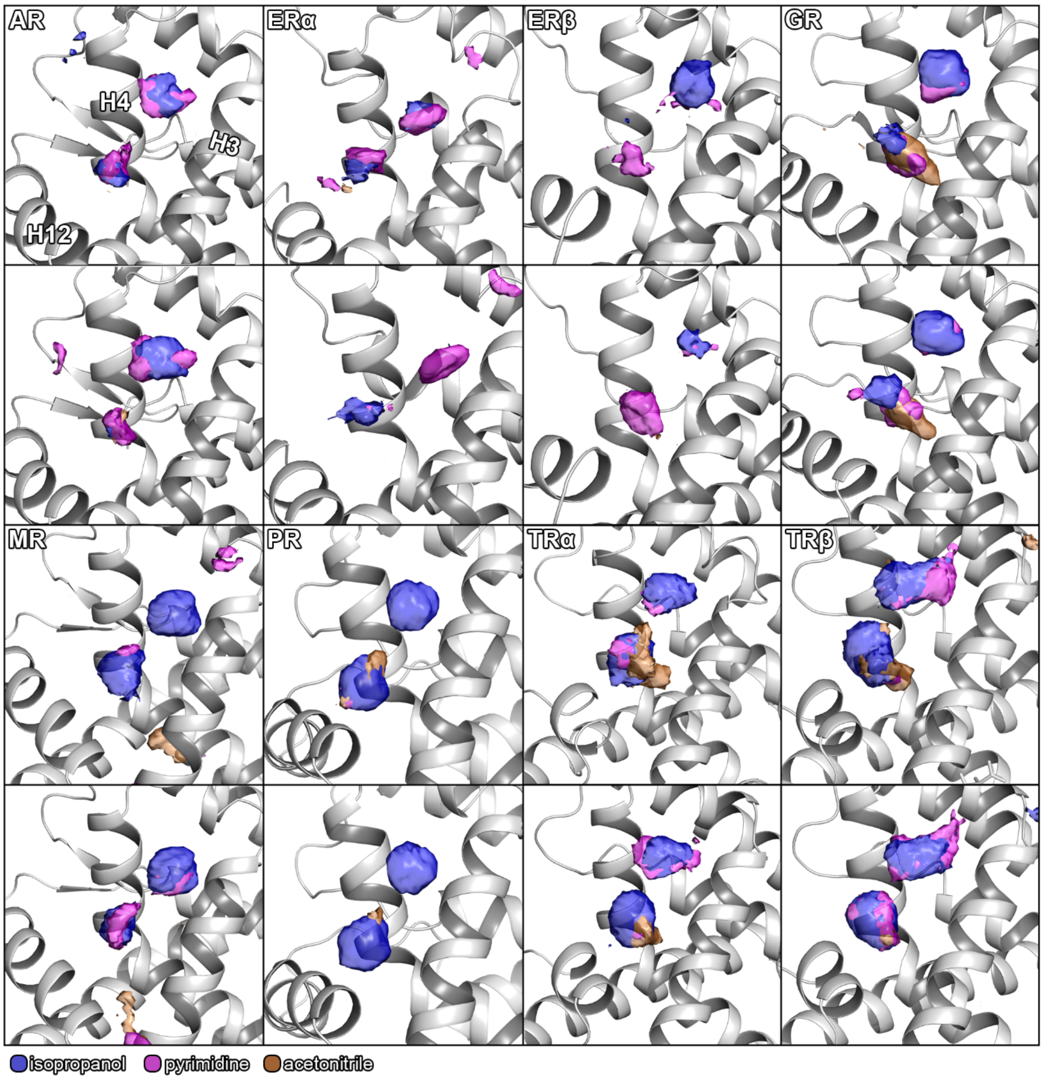


**Figure S 3** Comparison between cosolvent densities between apo and holo protein for the AF-2 site. For each receptor, a comparison of the probe densities between holo (upper part) and apo (lower part) structure is shown. The densities are shown at an isovalue of 12. A legend to interpret the colors is given below the figure. The viewpoint was held consistent.

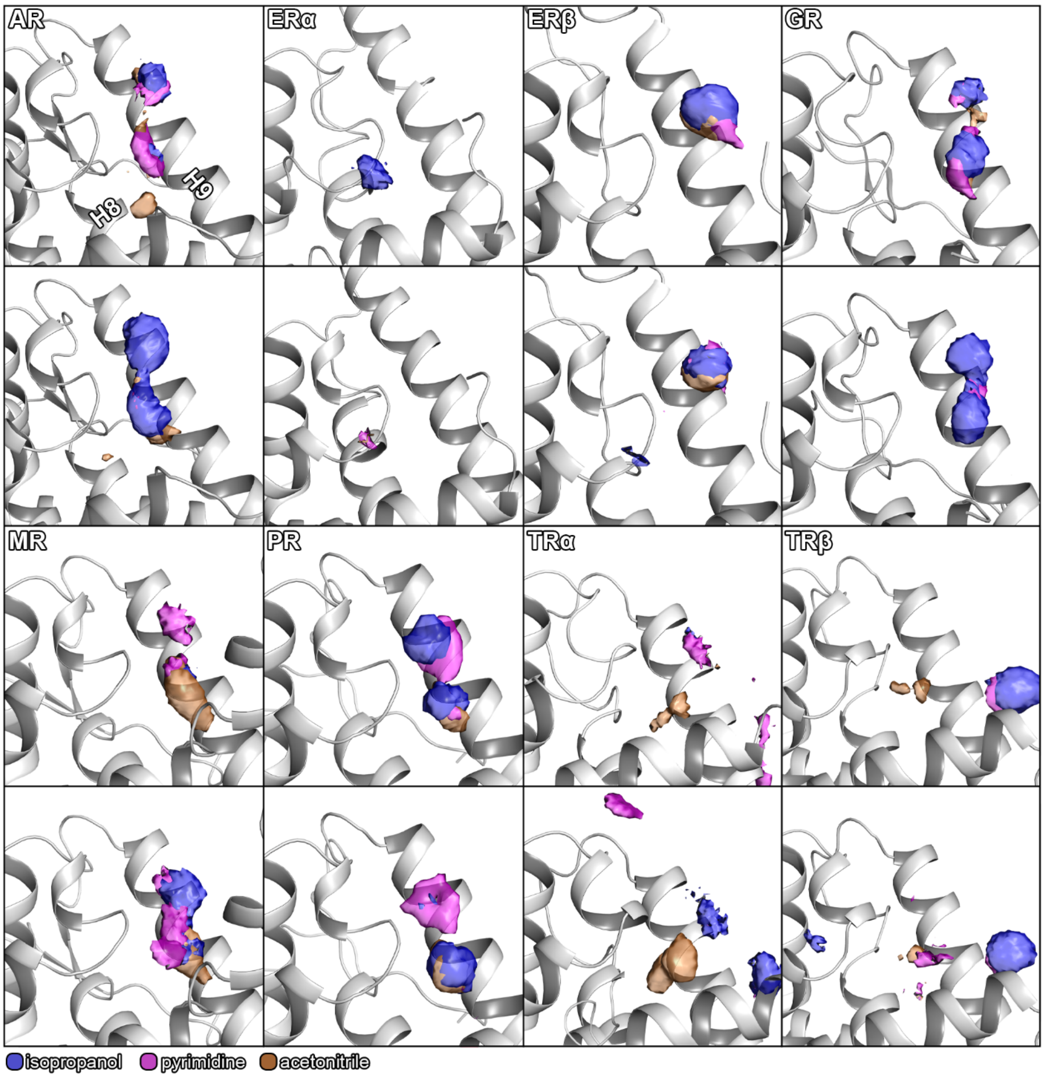


**Figure S 4** Comparison between cosolvent densities between apo and holo protein for the BF-3 site. For each receptor, a comparison of the probe densities between holo (upper part) and apo (lower part) structure is shown. The densities are shown at an isovalue of 12. A legend to interpret the colors is given in below the figure. The viewpoint was held consistent.

**Table S 1** Backbone RMSD of AR cosolvent MD simulations.

| *Replica* | *Acetonitrile apo* | *Isopropanol apo* | *Pyrimidine apo* | *Acetonitrile holo* | *Isopropanol holo* | *Pyrimidine holo* |
|---|---|---|---|---|---|---|
| 1 | 1.30 | 1.20 | 1.59 | 1.14 | 1.27 | 1.10 |
| 2 | 1.24 | 1.33 | 1.46 | 1.13 | 1.07 | 1.24 |
| 3 | 1.51 | 1.37 | 1.66 | 1.08 | 1.29 | 1.23 |
| 4 | 1.44 | 1.34 | 1.44 | 1.16 | 1.26 | 1.36 |
| 5 | 1.33 | 1.50 | 1.67 | 1.06 | 1.40 | 1.22 |
| 6 | 1.47 | 1.12 | 1.56 | 1.15 | 1.22 | 1.27 |
| 7 | 1.55 | 1.39 | 1.80 | 1.08 | 1.07 | 1.10 |
| 8 | 1.45 | 1.14 | 1.07 | 1.23 | 1.25 | 1.19 |
| 9 | 1.40 | 1.42 | 1.37 | 0.97 | 1.31 | 1.13 |
| 10 | 1.46 | 1.33 | 1.36 | 1.49 | 1.25 | 1.29 |

The backbone RMSD (Å) was determined between the input structure of the simulations and the last frame of the respective replica.

**Table S 2** Backbone RMSD of ER$\alpha$ cosolvent MD simulations.

| *Replica* | *Acetonitrile apo* | *Isopropanol apo* | *Pyrimidine apo* | *Acetonitrile holo* | *Isopropanol holo* | *Pyrimidine holo* |
|---|---|---|---|---|---|---|
| 1 | 2.19 | 1.61 | 1.88 | 1.36 | 1.48 | 1.48 |
| 2 | 1.60 | 1.59 | 1.29 | 1.24 | 1.35 | 1.67 |
| 3 | 1.90 | 1.54 | 1.74 | 1.81 | 1.27 | 1.59 |
| 4 | 1.55 | 1.74 | 1.68 | 1.59 | 1.28 | 1.48 |
| 5 | 1.75 | 1.87 | 2.01 | 1.39 | 1.55 | 1.45 |
| 6 | 1.88 | 1.36 | 1.99 | 1.45 | 1.22 | 1.28 |
| 7 | 1.45 | 1.79 | 1.56 | 1.27 | 1.81 | 1.48 |
| 8 | 1.74 | 1.54 | 1.43 | 1.44 | 1.65 | 1.40 |
| 9 | 1.44 | 1.67 | 1.87 | 1.64 | 1.22 | 1.42 |
| 10 | 1.88 | 1.51 | 1.89 | 1.40 | 1.19 | 1.74 |

The RMSD (Å) was determined between the input structure of the simulations and the last frame of the respective replica.

**Table S 3** Backbone RMSD of GR cosolvent MD simulations.

| *Replica* | *Acetonitrile apo* | *Isopropanol apo* | *Pyrimidine apo* | *Acetonitrile holo* | *Isopropanol holo* | *Pyrimidine holo* |
|---|---|---|---|---|---|---|
| 1 | 1.30 | 1.43 | 1.14 | 1.30 | 1.40 | 1.29 |
| 2 | 1.33 | 1.36 | 1.14 | 1.04 | 1.28 | 1.35 |
| 3 | 1.29 | 1.38 | 1.07 | 1.17 | 1.42 | 1.36 |
| 4 | 1.27 | 1.16 | 1.10 | 1.22 | 1.18 | 1.14 |
| 5 | 1.38 | 1.29 | 1.39 | 1.15 | 1.26 | 1.20 |
| 6 | 1.26 | 1.42 | 1.21 | 1.23 | 1.16 | 1.19 |
| 7 | 1.24 | 1.14 | 1.24 | 1.15 | 1.24 | 1.07 |
| 8 | 1.24 | 1.51 | 1.39 | 1.23 | 1.08 | 1.36 |
| 9 | 1.27 | 1.41 | 1.35 | 1.25 | 1.23 | 1.25 |
| 10 | 1.36 | 1.22 | 1.14 | 1.33 | 1.08 | 1.29 |

The RMSD (Å) was determined between the input structure of the simulations and the last frame of the respective replica.

**Table S 4** Backbone RMSD of MR cosolvent MD simulations.

| *Replica* | *Acetonitrile apo* | *Isopropanol apo* | *Pyrimidine apo* | *Acetonitrile holo* | *Isopropanol holo* | *Pyrimidine holo* |
|---|---|---|---|---|---|---|
| 1 | 1.31 | 1.44 | 1.86 | 1.61 | 1.60 | 1.46 |
| 2 | 1.56 | 1.30 | 1.47 | 1.69 | 1.30 | 1.78 |
| 3 | 1.68 | 1.76 | 1.73 | 1.41 | 1.51 | 1.50 |
| 4 | 1.78 | 1.68 | 1.88 | 1.38 | 1.50 | 1.37 |
| 5 | 1.67 | 1.48 | 1.74 | 1.48 | 1.59 | 1.49 |
| 6 | 1.61 | 1.69 | 1.53 | 1.67 | 1.58 | 1.33 |
| 7 | 1.41 | 1.40 | 1.84 | 1.35 | 1.48 | 1.78 |
| 8 | 1.39 | 1.54 | 1.59 | 1.49 | 1.45 | 1.40 |
| 9 | 1.42 | 1.43 | 1.66 | 1.72 | 1.59 | 1.56 |
| 10 | 1.25 | 1.87 | 1.70 | 1.68 | 1.28 | 1.36 |

The RMSD (Å) was determined between the input structure of the simulations and the last frame of the respective replica.

**Table S 5** Backbone RMSD of PR cosolvent MD simulations.

| *Replica* | *Acetonitrile apo* | *Isopropanol apo* | *Pyrimidine apo* | *Acetonitrile holo* | *Isopropanol holo* | *Pyrimidine holo* |
|---|---|---|---|---|---|---|
| 1 | 1.29 | 1.10 | 1.18 | 1.10 | 0.87 | 1.19 |
| 2 | 1.12 | 1.05 | 1.04 | 0.98 | 0.81 | 1.28 |
| 3 | 0.98 | 1.26 | 1.06 | 1.07 | 1.04 | 0.91 |
| 4 | 1.03 | 0.97 | 1.21 | 1.00 | 1.06 | 0.90 |
| 5 | 1.07 | 0.89 | 1.17 | 1.17 | 1.02 | 1.09 |
| 6 | 1.04 | 1.41 | 1.02 | 1.12 | 1.29 | 1.03 |
| 7 | 1.14 | 1.14 | 1.23 | 1.05 | 1.40 | 1.30 |
| 8 | 1.07 | 1.23 | 1.06 | 1.05 | 1.05 | 1.03 |
| 9 | 1.18 | 0.91 | 1.26 | 0.98 | 1.02 | 1.27 |
| 10 | 1.16 | 1.13 | 1.14 | 1.05 | 1.07 | 1.12 |

The RMSD (Å) was determined between the input structure of the simulations and the last frame of the respective replica.

**Table S 6** Backbone RMSD of TR$\alpha$ cosolvent MD simulations.

| *Replica* | *Acetonitrile apo* | *Isopropanol apo* | *Pyrimidine apo* | *Acetonitrile holo* | *Isopropanol holo* | *Pyrimidine holo* |
|---|---|---|---|---|---|---|
| 1 | 2.15 | 1.82 | 1.96 | 1.84 | 1.67 | 2.41 |
| 2 | 1.85 | 1.45 | 1.75 | 1.97 | 2.02 | 1.24 |
| 3 | 1.86 | 1.30 | 1.64 | 2.03 | 2.28 | 2.01 |
| 4 | 1.64 | 2.12 | 1.90 | 1.60 | 1.32 | 1.51 |
| 5 | 1.94 | 2.57 | 1.56 | 1.69 | 2.35 | 2.23 |
| 6 | 2.37 | 1.91 | 1.86 | 2.44 | 1.69 | 2.11 |
| 7 | 2.03 | 1.76 | 2.25 | 1.97 | 1.61 | 1.82 |
| 8 | 1.74 | 2.01 | 2.21 | 1.82 | 1.91 | 1.72 |
| 9 | 1.71 | 1.73 | 1.93 | 1.91 | 2.32 | 1.82 |
| 10 | 2.03 | 1.85 | 1.63 | 1.88 | 1.92 | 2.08 |

The RMSD (Å) was determined between the input structure of the simulations and the last frame of the respective replica.

**Table S 7** Backbone RMSD of TR$\beta$ cosolvent MD simulations.

| *Replica* | *Acetonitrile apo* | *Isopropanol apo* | *Pyrimidine apo* | *Acetonitrile holo* | *Isopropanol holo* | *Pyrimidine holo* |
|---|---|---|---|---|---|---|
| 1 | 1.51 | 1.49 | 1.38 | 1.51 | 1.85 | 1.95 |
| 2 | 1.42 | 1.77 | 1.75 | 1.55 | 1.40 | 1.64 |
| 3 | 1.56 | 1.72 | 1.51 | 1.54 | 1.55 | 1.69 |
| 4 | 1.64 | 1.24 | 1.32 | 1.51 | 1.51 | 1.68 |
| 5 | 1.87 | 1.61 | 1.73 | 1.55 | 1.69 | 1.91 |
| 6 | 1.58 | 1.90 | 1.50 | 1.71 | 1.70 | 1.61 |
| 7 | 1.48 | 1.65 | 1.32 | 1.55 | 1.31 | 1.74 |
| 8 | 1.63 | 1.62 | 1.43 | 1.56 | 1.65 | 1.59 |
| 9 | 1.48 | 1.40 | 1.73 | 1.65 | 1.66 | 1.60 |
| 10 | 1.63 | 1.74 | 1.73 | 1.59 | 1.51 | 1.31 |

The RMSD (Å) was determined between the input structure of the simulations and the last frame of the respective replica.

## Conformational Change

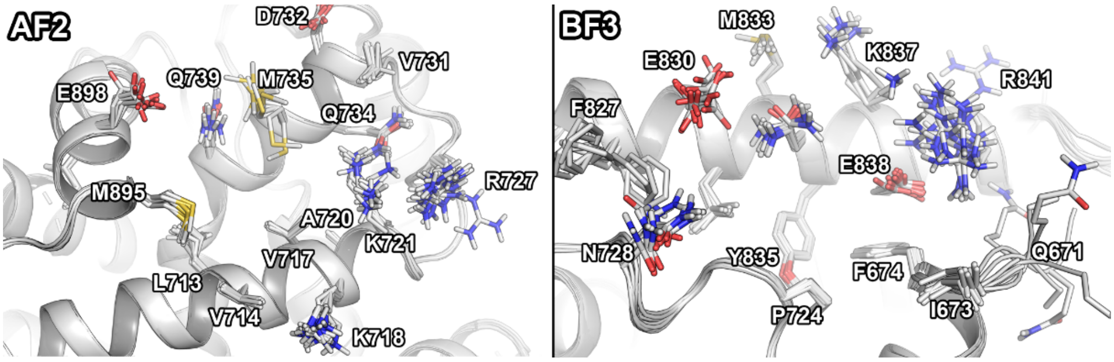


**Figure S 5** Conformational change at AF-2 and BF-3 crystal structures. Superposition of holo crystal structures of the allosteric site (PDB IDs: 2PIP, 2PIV, 2YHD, 2YLO, 2YLP, 2PIT, 2PIU, 2PIO, 2PKL, 2YLQ, 2PIW, 4HLW).

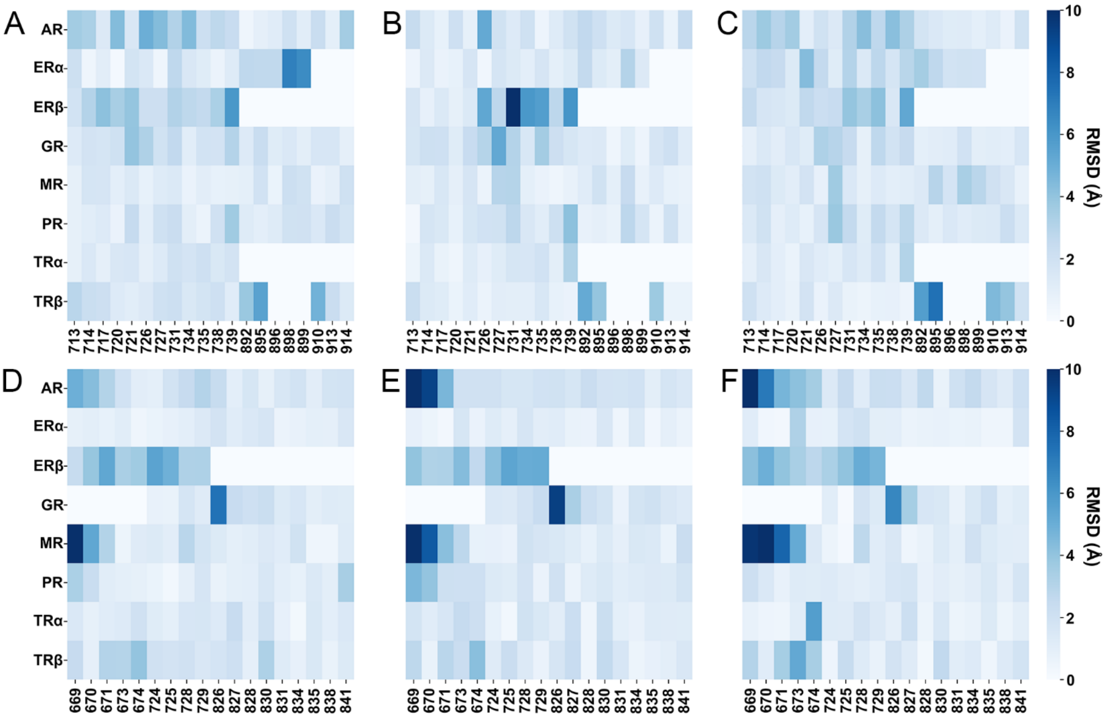


**Figure S 6** Conformational change at AF-2 and BF-3 determined by RMSD. The RMSD between the representative structures of cosolvent and pure water simulations is shown for (A) AF-2 site in acetonitrile, (B) AF-2 in isopropanol, (B) AF-2 in pyrimidine, (D) BF-3 in acetonitrile, (E) BF-3 in isopropanol, and (F) BF-3 in pyrimidine.

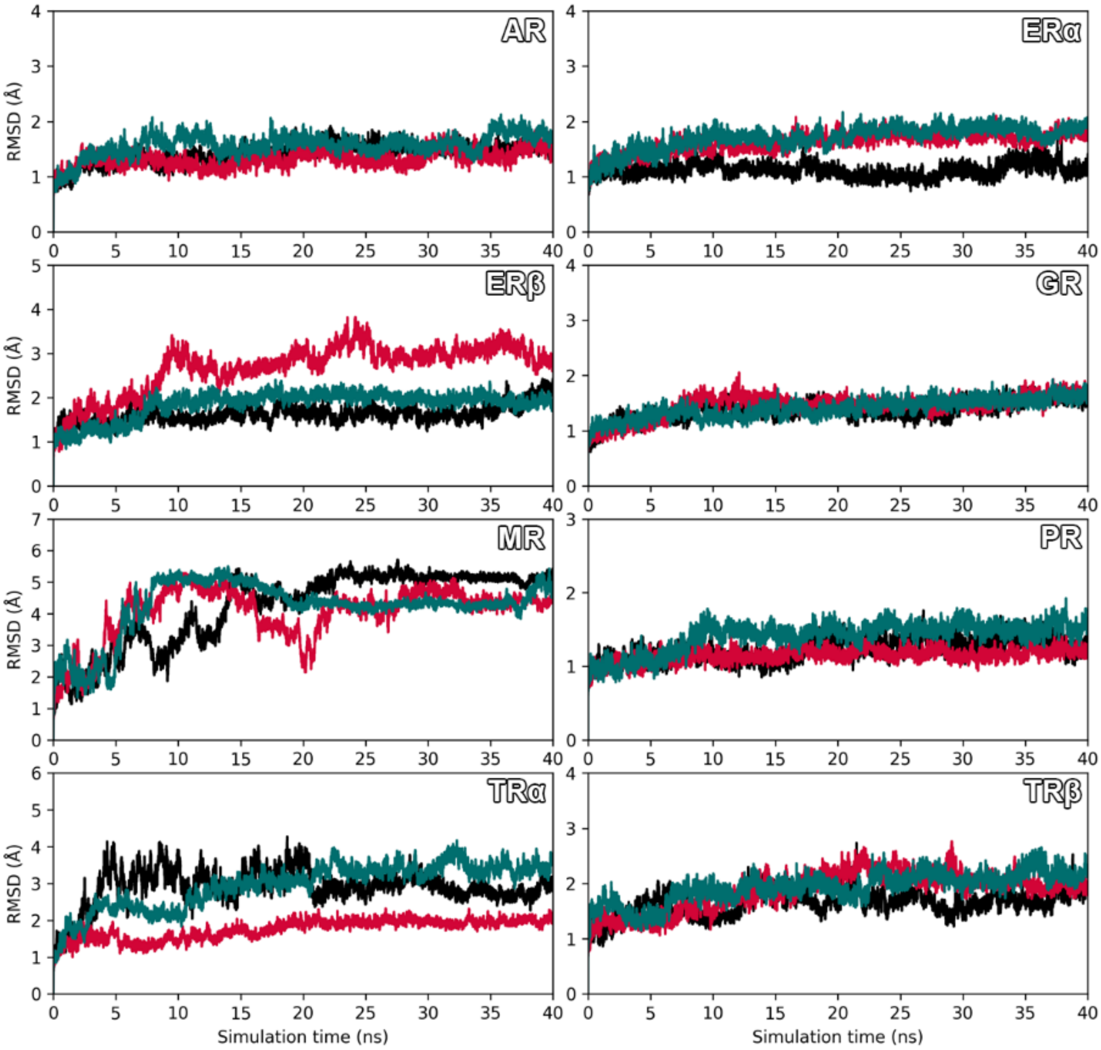


**Figure S 7** RMSD of simulations in pure water. The backbone RMSD of simulations in pure water (performed in triplicates) is presented for each receptor.

## Hydration Sites of the Allosteric Sites

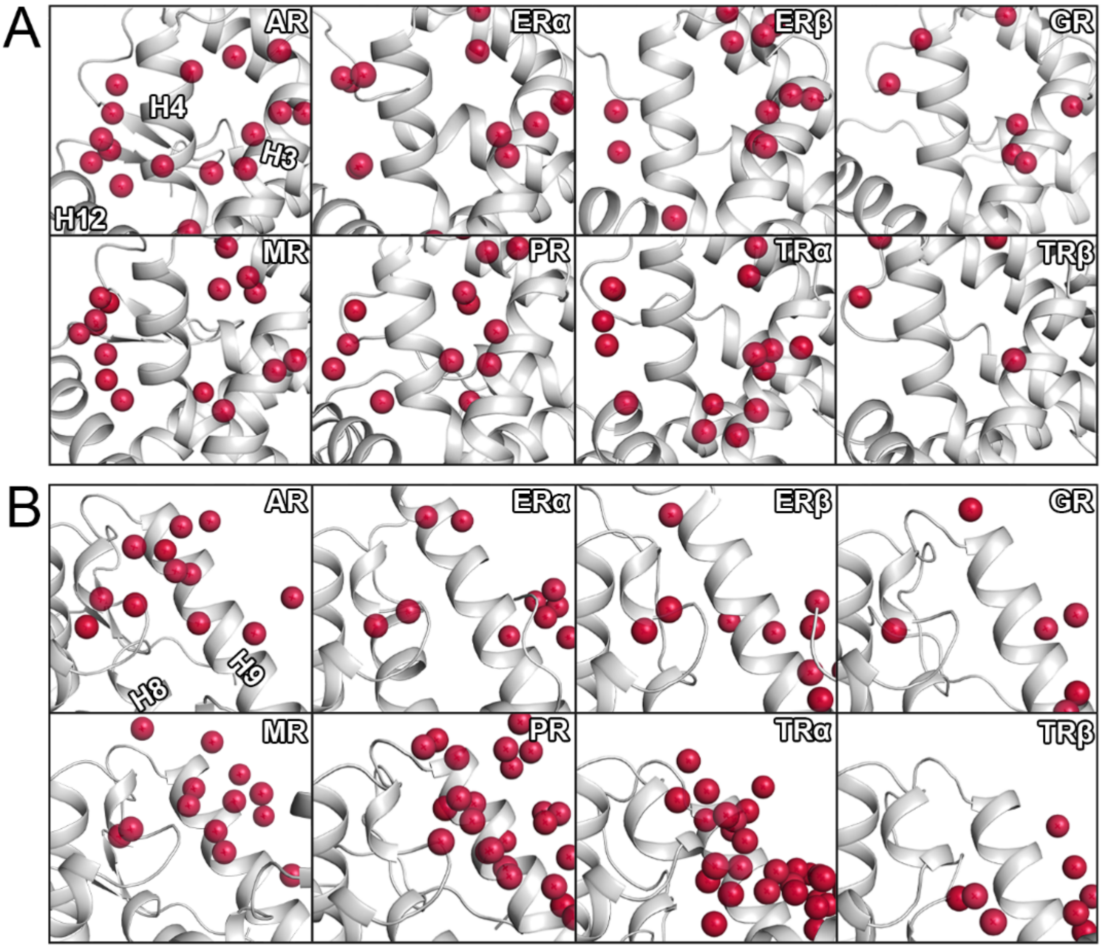


**Figure S 8** Hydration sites determined from crystal structure analysis. The hydration sites determined to be conserved in the hydration site analysis based on crystal structures. While (A) highlights the AF-2 site the (B) panel presents the BF-3 site.

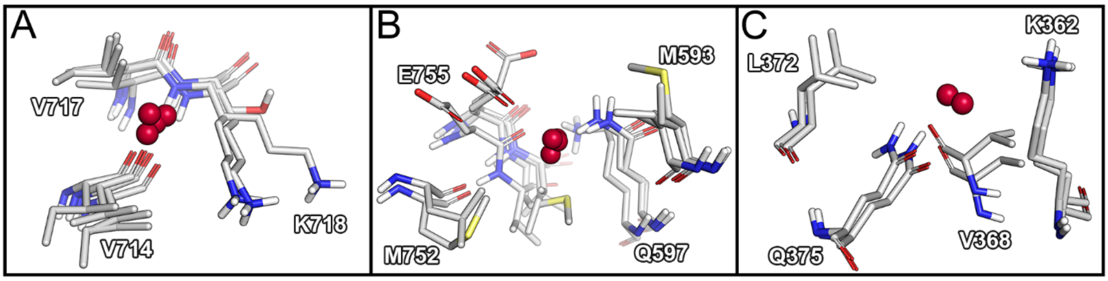


**Figure S 9** Hydration sites determined from crystal structure analysis. (A) Hydration site conserved among AR, ER$\beta$, GR and MR. (B) Hydration site conserved among ER$\beta$, GR, PR, TR$\beta$. (C) Hydration site conserved among ERs. The nomenclature for the shown residues was selected based on (A) AR, (B) GR, and (C) ER$\alpha$.

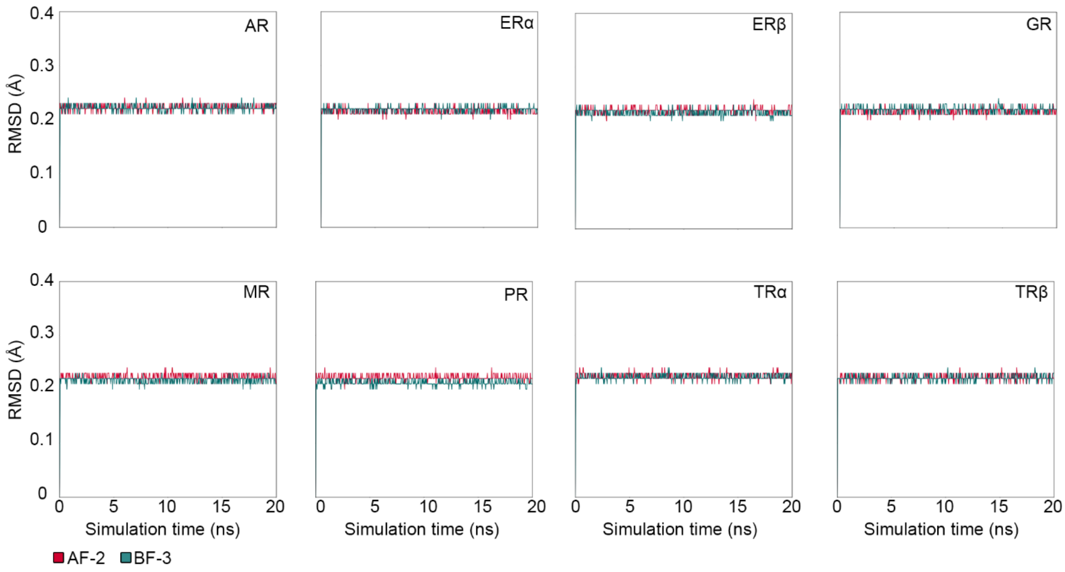


**Figure S 10** RMSD analysis of WATsite simulations. The backbone RMSD of WATsite simulations is presented for each receptor. Since a separate simulation was performed for each site, different colors were used to indicate the respective simulation.

### Tables S8-S16

Please refer to the original article for this table as it contains a large amount of raw data.

### Molecular Docking

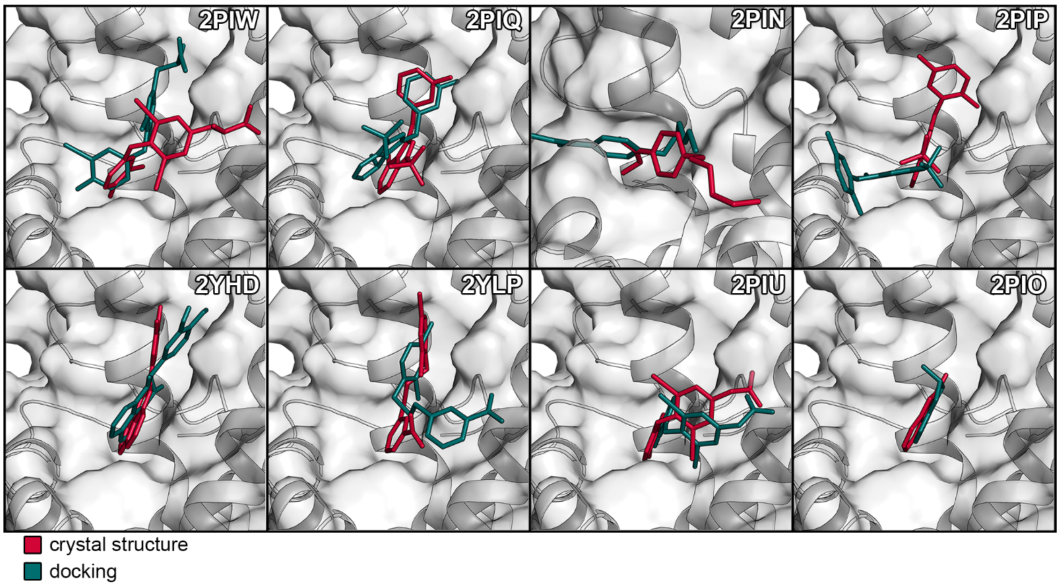


**Figure S 11** Poses obtained from redocking known crystallographic ligands: Glide SP for the AF-2 site.

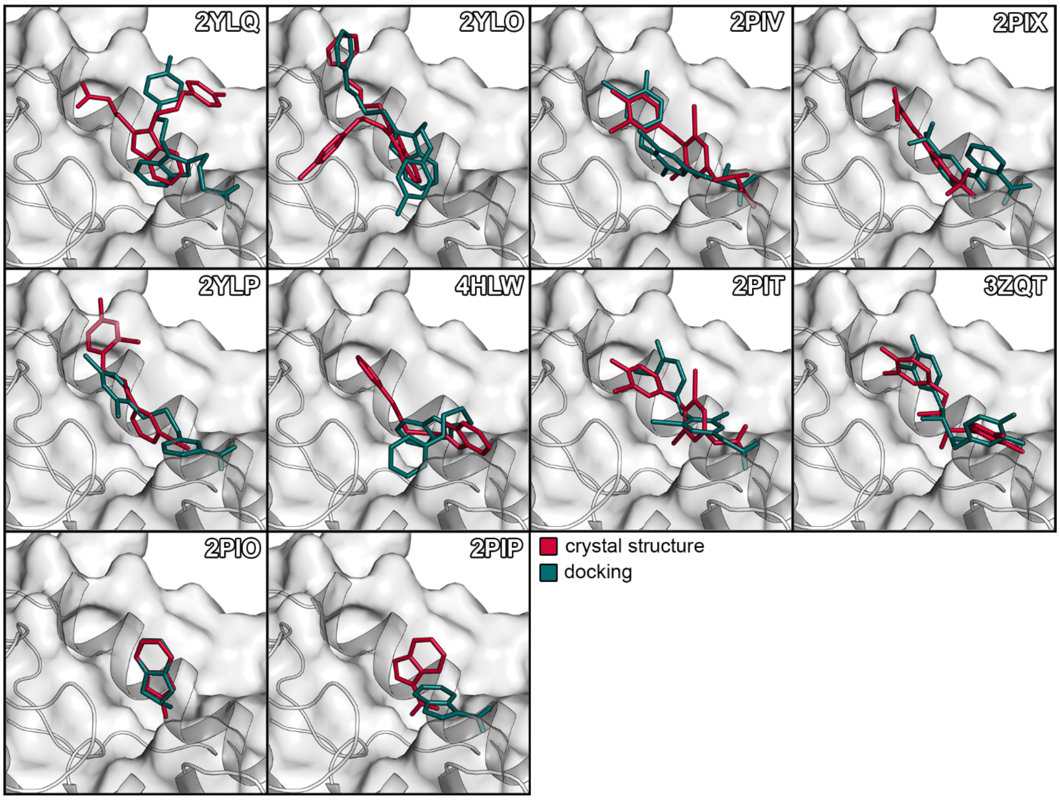


**Figure S 12** Poses obtained from redocking known crystallographic ligands: Glide SP for the BF-3 site.

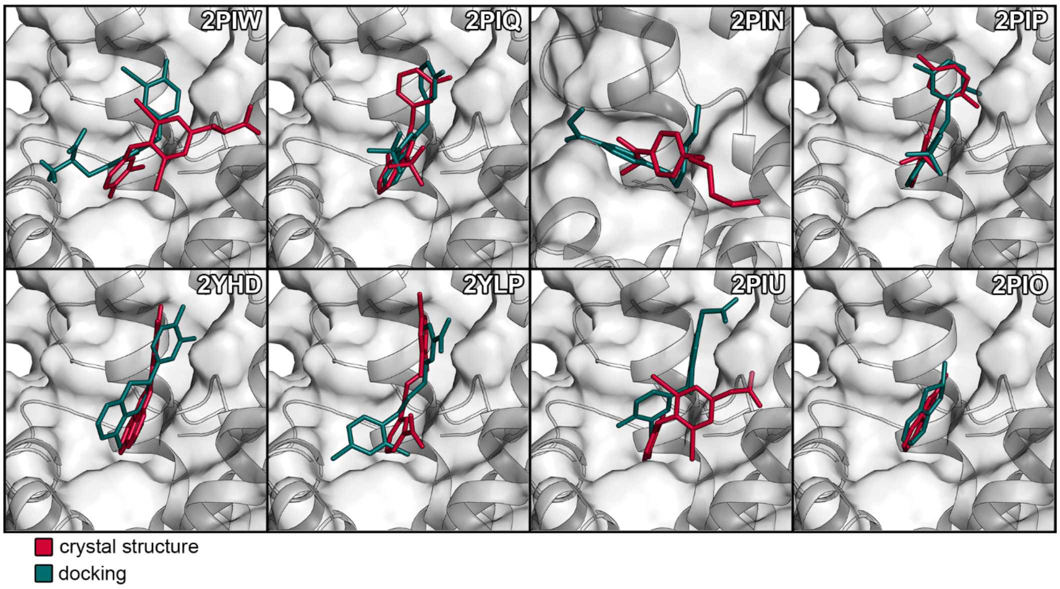


**Figure S 13** Poses obtained from redocking known crystallographic ligands: Glide XP for the AF-2 site.

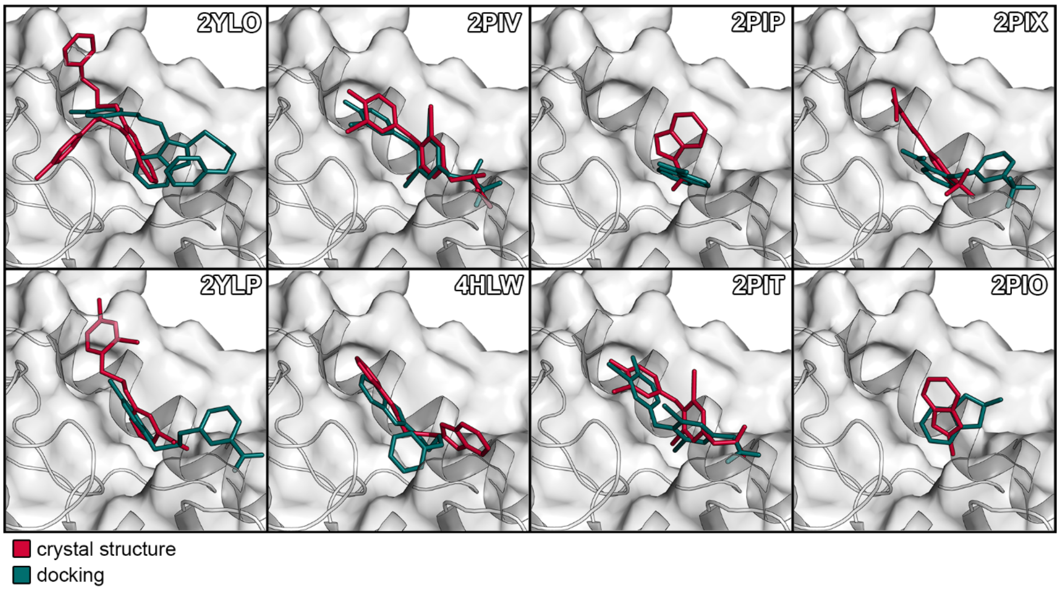


**Figure S 14** Poses obtained from redocking known crystallographic ligands: Glide XP for the AF-2 site.

**Table S 17** RMSD obtained from redocking known crystallographic ligands.

| *RMSD* | *Site* | *RMSD SP*$^a$ (Å) | *RMSD XP* $^b$ *(Å)* |
|---|---|---|---|
| 2PIQ | AF-2 | 2.11 | 1.19 |
| 2YHD | AF-2 | 1.47 | 1.96 |
| 2PIW | AF-2 | 4.90 | 7.92 |
| 2PIP | AF-2 | 7.10 | 0.84 |
| 2YLP | AF-2 | 4.67 | 7.24 |
| 2PIU | AF-2 | 2.24 | 4.28 |
| 2PIO | AF-2 | 1.67 | 0.92 |
| 2YLQ | BF-3 | 6.76 | n/a$^c$ |
| 2PIX | BF-3 | 7.46 | 5.06 |
| 2YLP | BF-3 | 5.22 | 7.10 |
| 2PIP | BF-3 | 5.15 | 4.75 |
| 4HLW | BF-3 | 4.78 | 7.00 |
| 2PIO | BF-3 | 1.25 | 3.56 |
| 2PIV | BF-3 | 2.11 | 1.07 |
| 2YLO | BF-3 | 3.83 | 8.22 |
| 2PIT | BF-3 | 2.11 | 1.90 |
| 3ZQT | BF-3 | 1.76 | n/a$^c$ |
| 2PIN | AF-2 | 4.83 | 4.91 |

[a] Results obtained using SP docking protocol. [b] Results obtained using XP docking protocol. [c] No pose was obtained by the applied protocol and specifications.

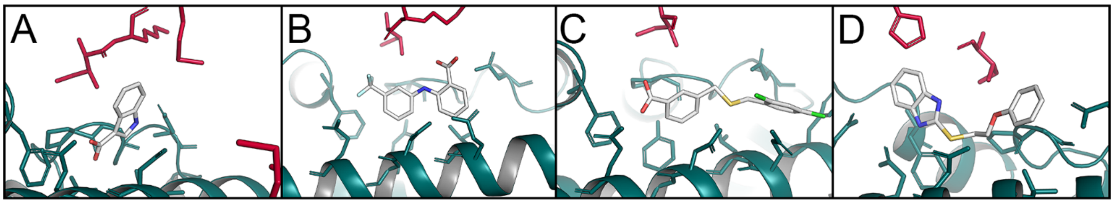


**Figure S 15** Crystal mates around the BF-3 site. Crystal mates in the 4 A radius of cocrystallized ligands at the BF-3. Nearby mates were colored red.

**Table S 18** RMSD obtained from redocking known crystallographic ligands.

| | *AR*<br>*AF-2* | *AR* (Å)<br>*BF-3* | *ER*$\alpha$<br>*AF-2* | *ER*$\beta$<br>*AF-2* | *TR*$\alpha$<br>*AF-2* | *TR*$\beta$<br>*AF-2* | *GR*<br>*AF-2* |
|---|---|---|---|---|---|---|---|
| Actives | 44 | 87 | 65 | 3 | 28 | 99 | 8 |
| Decoys | 2650 | 4350 | 4957 | 200 | 1450 | 5350 | 450 |
| ROC AUC | 0.75 | 0.76 | 0.71 | 0.85 | 0.45 | 0.55 | 0.87 |

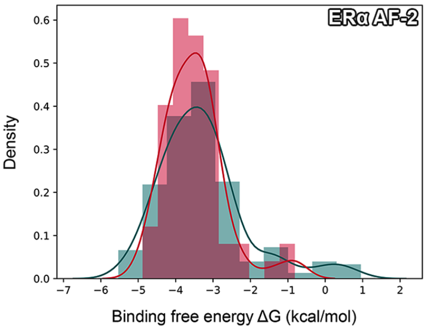


**Figure S 16** Score distributions for the ER$\alpha$.

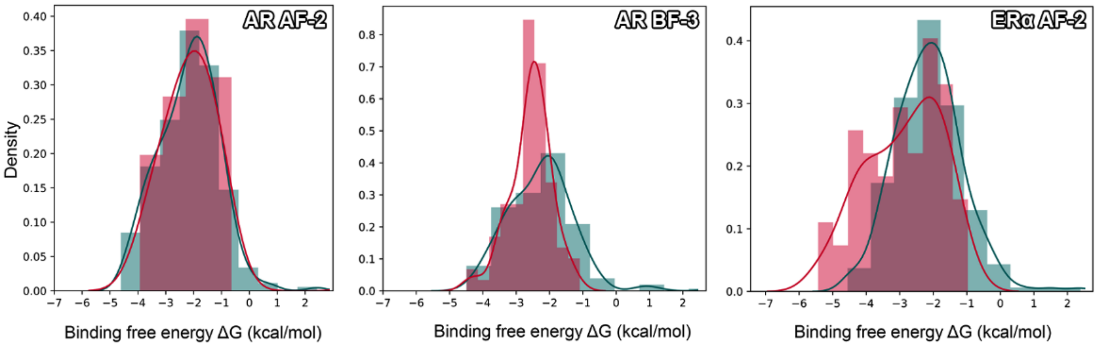


**Figure S 17** Score distributions determined by the Glide XP docking protocol.

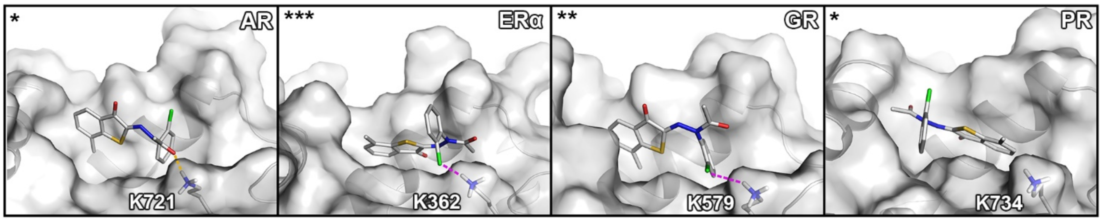


**Figure S 18** VPC16606 in docked to various NRs.

# Supporting Materials and Methods

## Sequence Similarity Among Hormonal NRs

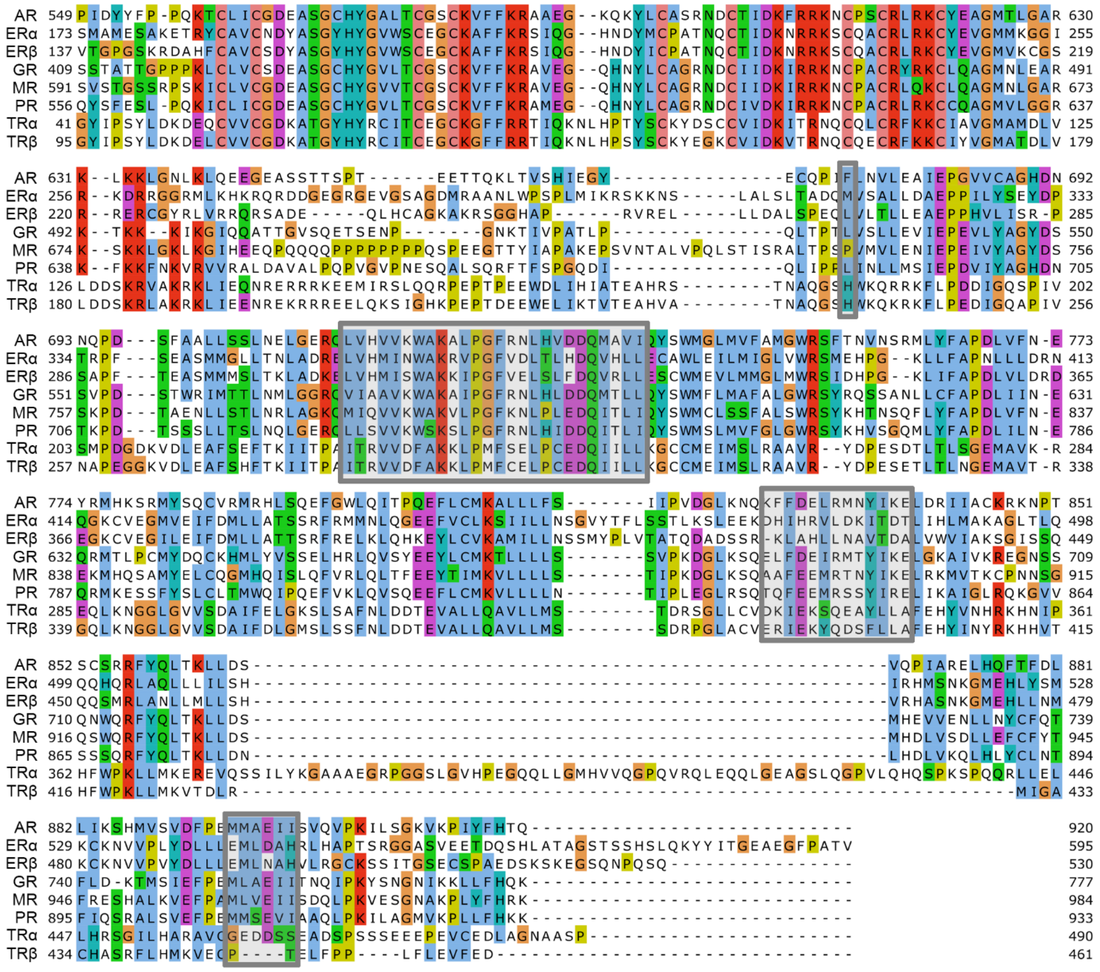


**Figure S 19** Sequence alignment of all NRs assessed in this study. Sequence alignment of the NRs considered in this study. Residues of the allosteric sites were indicated with gray boxes.

**Table S 19** RMSD obtained from redocking known crystallographic ligands.

| *Group* | *Amino acids* |
|---|---|
| 1 | A, I, L, M, F, W, V, C |
| 2 | N, Q, S, T |
| 3 | E, D |
| 4 | K, R |
| 5 | H, Y |
| 6 | P |
| 7 | G |

The amino acids groups used to determine the degree of conservation according to the ClustalW scheme are shown. The residues are given in single-letter code.

## Ligand preparation

**Table S 20** Structures prepared for molecular docking.

Please refer to the original article for this table as it contains a large amount of raw data.

## Protein preparation

**Table S 21** RMSD obtained from redocking known crystallographic ligands.

| *Receptor* | *Structures MD* | *Structures docking* | *Template structures* | *UniProt entry* |
|---|---|---|---|---|
| AR | 3L3X | 2PIT | n/a | P10275 |
| ER$\alpha$ | 5WGD | 3UUD | 1X7R | P03372 |
| ER$\beta$ | 4J24 | 2J7Y | 3OLS | Q92731 |
| GR | 5NFP | 3K22 | n/a | P04150 |
| MR | 2AA2 | 2AA2 | 2A3I | P08235 |
| PR | 1A28 | 1A28 | n/a | P06401 |
| TR$\alpha$ | 4LNW | 4LNW | n/a | P10827 |
| TR$\beta$ | 1XZX | 3GWS | 2J4A | P10828 |

The amino acids groups used to determine the degree of conservation according to the ClustalW scheme are shown. The residues are given in single-letter code.

**Table S 22** Relaxation protocol prior to MD simulation.

| *Desmond stage* | *Procedure* |
|---|---|
| 1 | Task (reading files, initializing parameters) |
| 2 | Simulate, Brownian Dynamics, NVT, T = 10 K, small time steps, and restraints on solute heavy atoms, 100 ps |
| 3 | Simulate, NVT, T = 10 K, small time steps, and restraints on solute heavy atoms, 12 ps |
| 4 | Simulate, NPT, T = 10 K, and restraints on solute heavy atoms, 12 ps |
| 5 | Solvate pocket |
| 6 | Simulate, NPT and restraints on solute heavy atoms, 12 ps |
| 7 | Simulate, NPT and no restraints, 24 ps |

**Table S 23** Mixed Solvent MD Desmond relaxation protocol.

| *Desmond stage* | *Procedure* |
|---|---|
| 1 | Brownian Dynamics, NVT, T = 10 K, 1 fs timestep, and restraints on all solute atoms, 24 ps |
| 2 | Brownian Dynamics, NVT, T = 10 K, 1 fs timestep, and restraints on solute heavy atoms, 24 ps |
| 3 | NVT, T = 10 K, 1 fs timestep, restraints on solute heavy atoms, 12 ps |
| 4 | NPT, T = 10 K, 2 fs timestep, restraints on solute heavy atoms, 12 ps |
| 5 | NPT, T= 300 K, 2 fs timestep, restraints on solute heavy atoms, 24 ps |
| 6 | NPT, T = 300 K, 2 fs timestep, 15 ps |

**Table S 24** Number input structures and minimal amount for cluster to be considered as conserved.

| *Receptor* | *Number of input structures* | *Minimal occupancy* |
|---|---|---|
| AR | 67 | 7 |
| ER$\alpha$ | 229 | 23 |
| ER$\beta$ | 32 | 3 |
| GR | 20 | 2 |
| MR | 27 | 3 |
| PR | 16 | 2 |
| TR$\alpha$ | 8 | 1 |
| TR$\beta$ | 9 | 1 |

## CHAPTER 9

# Computational Assessment of Combination Therapy of AR Targeting Compounds

Following their characterization in Chapter 8, the work presented here is focused on the assessment of potential combination therapies to modulate androgen receptor signaling with allosteric inhibitors. As classical antiandrogens induce conformational changes on the protein surface, the recognition of compounds binding the superficial allosteric sites AF-2 and BF-3 may be altered. By evaluating several microseconds of MD trajectories, this study suggests combinations that should be avoided for the treatment of prostate cancer.

**Author contributions:** Conceptualization, A.F.; formal analysis, A.F., F.H.; writing and original draft preparation, A.F.; writing, review and editing, A.F., M.A., M.S.; visualization, A.F.; supervision, M.A., M.S.

*Based on the published research article:*

Fischer, A.; Häuptli, F.; Lill, M. A.; Smieško, M. Computational Assessment of Combination Therapy of Androgen Receptor-Targeting Compounds. *J. Chem. Inf. Model.* 2021, 61, 2, 1001–1009.

## Abstract

The ligand binding domain of androgen receptor (AR) is a target for drugs against prostate cancer and offers three distinct binding sites for small-molecules. Drugs acting on the orthosteric hormone binding site suffer from resistance mechanisms that can, in the worst case, reverse their therapeutic effect. While many allosteric ligands targeting either the activation function-2 (AF-2) or the binding function-3 (BF-3) have been reported, their potential for simultaneous administration with currently prescribed antiandrogens was disregarded. Here, we report results of 40 $\mu$s molecular dynamics simulations to investigate combinations of orthosteric and allosteric AR antagonists. Our results suggest BF-3 inhibitors to be more suitable in combination with classical antiandrogens as opposed to AF-2 inhibitors based on binding free energies and binding modes. As mechanistic explanation for these observations, we deduced a structural adaptation of helix-12 involved in the formation of the AF-2 site by classical AR antagonists. Additionally, the changes were accompanied by an expansion of the orthosteric binding site. Considering our predictions, the selective combination of AR-targeting compounds may improve the treatment of prostate cancer.

## Introduction

The androgen receptor (AR) is critically involved in the development and progression of hormone-dependent prostate cancer (PC) rendering it a central target in its pharmacological treatment [1, 2]. Despite the initial success in treating the disease with the currently available AR antagonists, they frequently fail in later stages due to alterations of the receptor leading to drug resistance and a condition referred to as castrate-resistant prostate cancer (CRPC). Since single amino acid mutations near the binding site can lead to the conversion of classical AR antagonists to agonists and thus reverse their therapeutic effect, there were great efforts to develop compounds targeting alternative sites at the receptor to bypass these effects [3, 4, 5, 6]. At the ligand binding domain (LBD) shown in Figure 1A, these compounds generally target either the activation function-2 (AF-2) involved in coactivator binding essential for downstream signaling, or the binding function-3 (BF-3) suspected to allosterically modulate the AF-2 and interact with chaperones [4, 7, 8] (Figure 1B). In previous studies, it was suggested to combine AF-2 inhibitors with classical antiandrogens to improve the therapeutic outcome

[9]. However, as the association of coactivator proteins at the AF-2 site is prevented by orthosteric AR antagonists, the affinity and residence of allosteric inhibitors targeting this site might be altered in the case of concurrent therapy [10]. Similarly, it remains unknown if ligands binding to the BF-3 could be affected in the same way. Combination therapy with multiple drugs is a popular and often successful strategy in the pharmacological interventions against malignant tumors including PC. To improve PC treatment efficacy, combination of orthosteric and allosteric inhibitors offers an attractive approach as such combinations can theoretically even restore the antiandrogenic function of compounds suffering from resistance mechanisms by modifying the intrinsic residue networks or shifting the structural ensemble toward more favorable conformational states and, thus, amplify the resulting pharmacological effect. As the communication between the orthosteric site and AF-2 is thought to be bidirectional [11], such effects would be feasible. The safety and benefit of combination therapies needs to be carefully assessed regarding potential drug-drug interactions before they reach the clinics [12, 13, 14, 15, 16].

As discussed in a recent review, it remains an open question on how we can design synergistically optimized combinations of orthosteric and allosteric drugs [12]. In an experimental setting, protein crystallography, siRNA screening, affinity determination, or functional assays constitute methods to investigate drug combinations. Nevertheless, due to the high number of possible combinations, such experiments are costly, especially when large libraries of compounds are considered in early stages of drug discovery [17]. Further, resolving crystal structures of the AR bound to orthosteric antagonists was, until now, not achieved without specific binding site mutations altering the molecular mechanism of these ligands [18]. Therefore, crystallography is not feasible to explore their combination with allosteric ligands. Thus, there is a need for cost-effective methods for the development of synergistically acting combinations. In this regard, computational tools offer an inexpensive alternative to rationalize and support laboratory experiments. Several studies focused on various therapeutic targets have conducted molecular dynamics (MD) simulations with drug combinations to study the interplay of orthosteric and allosteric ligands [19, 20, 21, 22]. MD simulations enable the detailed investigation of time-evolved ligand-protein interactions by introducing structural flexibility [23]. As large share of allosteric AR inhibitors were discovered in virtual screening projects [7, 9, 24, 25], computational methods such as MD simulations constitute a suitable instrument to investigate their combination with orthosteric ligands.

In this study, we conducted microsecond MD simulations to examine combinations of compounds targeting the LBD of the AR. In particular, we assessed the combination of four allosteric inhibitors, two targeting the AF-2 and BF-3 respectively, with conventional antagonists in the orthosteric site, as well as their concurrent binding (Figure 1C). We examined their binding modes, binding free energies, and time-evolved interaction with the allosteric sites to give recommendations for combination therapies targeting the AR. Our results suggest that BF-3 inhibitors could be concurrently administered with conventional antiandrogens such as bicalutamide offering a potentially synergistic treatment effect against PC. In contrast, the AF-2 inhibitors suffered reduced shape complementarity in combination with an orthosteric antagonist. Thus, it is possible that such a combination therapy would not be beneficial. A mechanistic analysis highlighted altered plasticity and topology of the AF-2 site in combination with bicalutamide to be responsible for our observations.

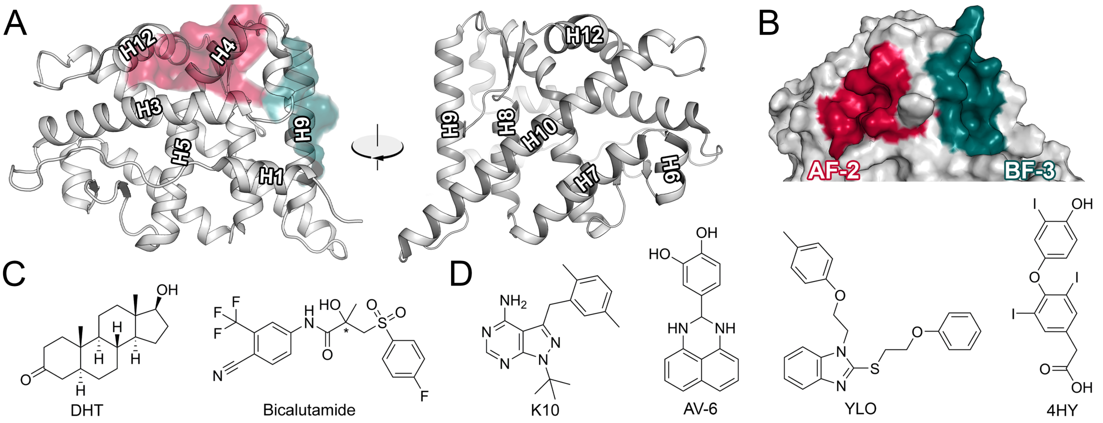


**Figure 1** Structural overview. (A) Helical architecture of the AR from two different orientations. The location of the AF-2 and BF-3 is indicated by the same colors as in subfigure B. (B) Location of the AF-2 and BF-3 allosteric sites. (C) Structures of the ligands assessed in this work.

## Results and Discussion

**Binding modes and solvent accessibility** As MD simulations provide a time-evolved ensemble of the ligand-protein complex, the area of the ligand that is exposed to solvent, relevant for hydrophobic effects and desolvation, can be determined. In ligand-protein binding, a decrease in solvent accessibility of the ligand is often associated with a gain in hydrophobic interaction and shape complementarity improving ligand affinity. Thus, its assessment can give important insights into the characteristics of a particular binding mode [26, 27]. The AF-2 and BF-3 of the AR are shallow sites located on the surface of the receptor. As the binding modes of several allosteric inhibitors are known [28], this allows their computational evaluation in combination with orthosteric or other allosteric ligands. Since we observed different degrees of ligand burying within their binding sites in our simulations, we computed the solvent-accessible surface area (SASA) of the ligands. As shown in Figure 2A, the two AF-2 inhibitors behaved differently regarding this metric. While both compounds K10 and AV-6 showed a statistically significant increase (Tables S1-S2) in SASA with bicalutamide the combination with a BF-3 inhibitor affected them differently. Compound K10 presented higher SASA values, but AV-6 improved in this metric in combination with YLO. A visual inspection of representative structures with AV-6 (Figure 2B) revealed alterations in the surface topology of the AF-2 site between structures bound to the natural hormone dihydrotestosterone (DHT) and bicalutamide. Therefore, the orthosteric antagonist decreased the capability of K10 and AV-6 to induce favorable conformational changes promoting hydrophobic contacts in the AF-2 site. As only one ligand was negatively affected when allosteric inhibitors for AF-2 and BF-3 were combined, the alterations of the AF-2 might not impact all compounds to the same degree bound to this site meaning that this effect is ligand-specific. Compounds associating with the BF-3 generally displayed increased or lower SASA values in the examined combinations compared to DHT. The values indicated improved burying within the site for YLO. A visualization of representative structures obtained from clustering based on an RMSD matrix (Figure 2C) confirmed the formation of a deeper pocket if YLO was combined with bicalutamide as opposed to DHT. Due to the induced conformational changes of the BF-3, the aromatic groups of the ligand are deeply inserted into the site leading to decreased solvent exposure. The SASA increase of 4HY amounted to 4.5 Å$^2$, which was only marginally significant and therefore, constituted a less drastic change as opposed to the decrease of up to 57.6 Å$^2$ with YLO. Notably, this analysis was only performed for simulations presenting a stable binding mode of the compounds without dissociation of the ligand, as it was done for all ligand-dependent metrics. Therefore, based on average SASA values, BF-3 inhibitors can be combined with both orthosteric antagonists and AF-2 inhibitors and could even benefit from these types of polypharmacology.

Regarding the most prevalent ligand-protein interactions, we computed their similarity across the assessed polypharmacology combinations based on binary fingerprints. The analysis consistently revealed differences in the binding modes between allosteric ligands in combination with DHT and bicalutamide (Figures 2D-F). In most cases, the allosteric ligands presented a similar interaction pattern in combinations with other allosteric ligands. These results show that alternative binding modes, which do not have to compromise the binding free energy or SASA, are present if the orthosteric site is occupied by bicalutamide. A comprehensive overview of ligand-protein contacts present in any combination is presented in Tables S3-S6. Based on the differences in ligand-protein interactions in a polypharmacology setting, future structure-based design efforts could consider the changes and improve binding in either combination.

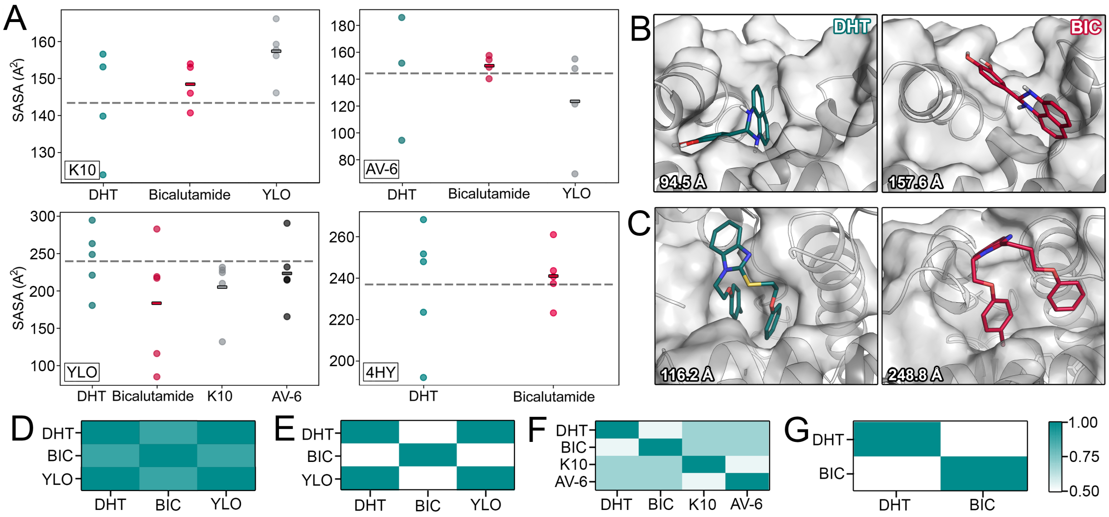


**Figure 2** Solvent accessibility and ligand-protein interactions. (A) The SASA of the respective allosteric ligand is shown in combination with orthosteric inhibitors (DHT or bicalutamide) or other allosteric inhibitors (YLO bound to the BF-3 site for K10 and AV-6 as well as K10 and AV-6 bound to the AF-2 site for YLO). Global average values of the DHT-control are indicated by a dashed line, while short lines indicate the averages of the remaining systems. (B) AV-6 in the AF-2 site is shown with DHT or bicalutamide. The average SASA of the respective simulation is indicated. (C) YLO in the BF-3 site is shown with bicalutamide and DHT. (D) Comparison of interactions of K10 systems. (E) Comparison of interactions of AV-6 systems. (F) Comparison of interactions of YLO systems. (G) Comparison of interactions of 4HY systems.

**Residence and binding energies** Besides thermodynamics, binding kinetics of ligands is an important evaluate to determine and optimize their binding efficiency. The residence of a molecule is defined as the period it resides in a particular binding site [29, 30]. Using computational methods such as MD simulations, a high number of binding events, usually obtained from several replica simulations, is required to quantify residence of ligands in accordance with experiments. As the time scale of such events is long in comparison to current simulation capabilities [30, 31], tremendous computational power would be necessary to investigate them in their full complexity. However, even though we only performed five replica simulations for each system, we could observe several unbinding events of AF-2 inhibitors, while the BF-3 inhibitors remained bound within their binding pocket (Figure 3A). As AF-2 ligand dissociation occurred independent of the orthosteric ligand, we assumed the crystallographic binding modes to be relatively unstable. Indeed, the incomplete ligand electron density of the crystal structure bound to K10 (PDB ID: 2PIP) suggests a certain degree of instability. Since the BF-3 inhibitors remained fairly stable throughout the microsecond MD simulations in all assessed combinations, they seemed to be less affected by allosterically mediated conformational changes within the receptor structure. Interestingly, one simulation in combination with DHT displayed both dissociation and reassociation of AV-6. Before and after this rare event, we could detect a highly similar binding mode with an RMSD of 1.6 Å for the best matching pair (Figure S1). However, as in other simulations, the ligands also presented alternative orientations before and after unbinding.

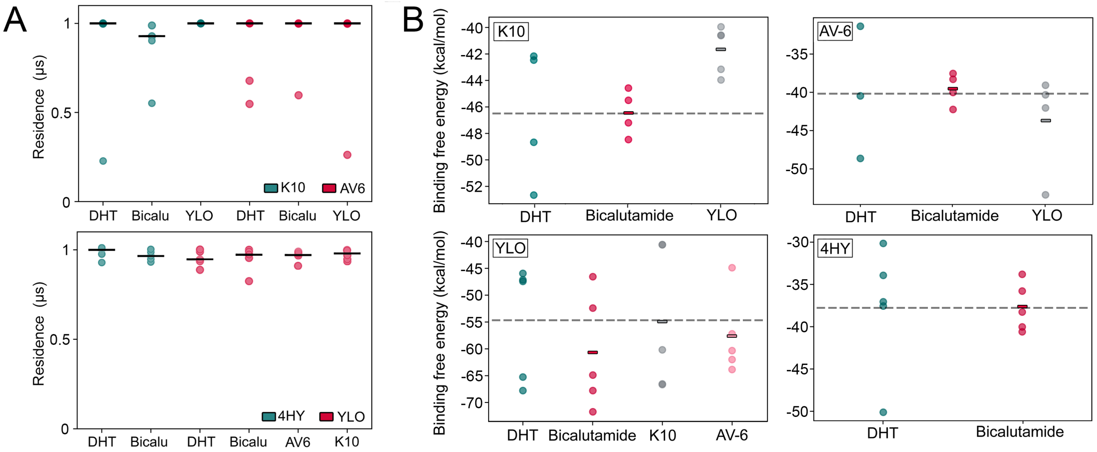


**Figure 3** Interactions of the ligands. (A) Residence of ligands within their binding sites based on distance to the respective allosteric site for all replica simulations. (B) Average ligand-protein binding free energies determined with the MM/GBSA protocol for all replicas. The global average value of the DHT-control simulations are indicated by a dashed line, while short lines indicate the averages of the remaining systems.

Except for the simulations studying the dissociation of the ligands, we determined absolute binding free energies using the molecular mechanics-generalized Born surface area (MM/GBSA) protocol of the last 100 ns of our simulations (Figure 3B). MM/GBSA calculations have been previously applied to study the interaction of orthosteric and allosteric inhibitors against the main protease of SARS-CoV-2 and have recommended potentially synergistic drug combinations [19]. Regarding AF-2 inhibitors, our results indicated that the orthosteric AR antagonist bicalutamidedid not significantly affect the binding free energy (Tables S1-S2) of both allosteric inhibitors (Figure 3B). However, while K10 suffered from a statistically significant loss in binding affinity in combination with the BF-3 inhibitor, AV-6 displayed improved energies in combination with YLO. Hence, AF-2 inhibitors could profit or suffer from a polypharmacology approach with other allosteric ligands depending on the respective ligand. This is surprising, as the rearrangement of H12 manipulates the topology of the AF-2 site as we elaborate in the following sections. Interestingly, the BF-3 inhibitors either presented unchanged or improved binding energies in either the combination with bicalutamide or AF-2 inhibitors. Especially YLO presented improved binding free energies with bicalutamide, as opposed to all other ligands. While the average energy amounted to -54.7 kcal/mol if DHT was bound concurrently, the average energies improved by -6.0 kcal/mol in combination with bicalutamide, which is the most significant difference in binding free energy we observed between different systems. The combination of YLO with AV-6 resulted in an average gain of -3.0 kcal/mol. The binding free energy of 4HY was not significantly affected by the presence of bicalutamide, even though there was a slight non-significant trend for improved energies. The number of ligand-protein hydrogen bonds remained stable among the different combinations (Figure S2). In conclusion, BF-3 inhibitors were clearly less affected by the combination with other allosteric or orthosteric inhibitors based on binding free energies, solvent accessibility and residence within their binding site (Table 1). On the other hand, combinations of AF-2 inhibitors impacted the individual compounds to various degrees suggesting that some combinations could potentially be without desirable benefit.

**Table 1** Conclusions regarding drug combinations.

| **Ligand** | **Combination** | **Binding energy**[a] | **Binding mode**[b] |
|---|---|---|---|
| K10 | BIC | unchanged | worse |
| | YLO | worse | worse |
| AV-6 | BIC | unchanged | worse |
| | YLO | improved | improved |
| YLO | BIC | improved | improved |
| | AV-6 | improved | improved |
| | K10 | unchanged | improved |
| 4HY | BIC | unchanged | worse |

Statistical significant differences of the systems compared to DHT-bound systems. [a]Binding energy refers the binding free energies obtained from the MM/GBSA results [b]Binding mode refers to the SASA data.

**Mechanistic insights** In two recent studies, it was shown that classical antiandrogens allosterically modulate the structure of the AR, in particular the AF-2 site involved in coactivator association [11, 32]. An allosteric pathway involving helix-3 (H3), H4, and H12 was proposed to induce changes of the AF-2. Another study, focused on the conformational changes caused by bicalutamide, reported alterations in H10 and H12. Since H3 and H12 are involved in the formation of this allosteric site, we studied them in more detail to understand our previous findings mechanistically. In a first step, we determined the present per-residue secondary structure of H3, H10, and H12 during all frames of our simulations. While helices 3 and 10 did not show strong alterations with different ligand combinations regarding their overall helicity (Figures S3-S5), we observed specific changes of H12 regarding its spatial location and helical architecture (Figure 4). A visual comparison of H12 between the AR bound to DHT and bicalutamide combined with an AF-2 allosteric inhibitor revealed a shift towards the AF-2 site and a stretching around its center. We observed similar conformational changes in the control simulations without allosteric inhibitors (Figures S6 and S7). A displacement of H12 as well as a break in helicity in the presence of antagonists was previously reported [32, 33]. Even though we observed a high variability among the simulations, the overall helicity was significantly decreased in all four simulations with bicalutamide bound to the orthosteric site compared to DHT-bound systems including the controls (Table S2). As we pre-equilibrated the bicalutamide systems before adding allosteric ligands, the simulations started in a low helicity state, which persisted in the subsequent period of time. Additionally, the control simulations presented a similar picture (Figure S6). Conformational changes of H12 inherently influence the topology of the AF-2 site and thus likely modulate the interaction of AF-2 inhibitors as we observed it. In the BF-3, residues P723 and F826, which are known to be crucially involved in ligand-protein interactions [34], displayed significant alterations in their mobility. Our interaction analysis confirmed the involvement of those two residues in ligand-protein contacts at the BF-3. However, the BF-3 inhibitors seemed to be more tolerant to such differences in binding pocket dynamics as indicated by their unchanged or improved predicted binding affinity and shape complementarity. To obtain better insight into the allosteric communication within the receptor, dynamic cross-correlation map (DCCM) analysis correlating collective motions among protein residues was applied in previous work [11], in which a positive correlation between H10 and the H11-H12 region, as well as the H3-H4 and H10 region was reported. Similarly, our DCCM analysis of four systems presenting distinct alterations in the positioning of H12 revealed positive correlations between H3-H4 and the terminal region of H10 (Figure S8). Additionally, we registered a positive correlation regarding the collective motions of H3 and H12. The highest negative correlation of distant sites of the protein was observed by H4-H5 and H12, which was more pronounced in simulations with DHT compared to bicalutamide. Similarly, the correlation between H3 and H10 was lower in simulations with bicalutamide. In addition to the DCCM analysis, we computed the average betweenness centrality (BC) describing the importance of a protein residue for intramolecular communication [35]. Comparing the simulations with DHT and bicalutamide, we could observe a distinct pattern for changes in BC (Figures S6 and S9). While the values for H3 were always higher in the presence of DHT, a narrow peak corresponding to the residue L811 located in H8 was higher in simulations with bicalutamide. The increased BC values of H3 are in accordance with its higher collective motions with DHT as opposed to bicalutamide. The residue L811 is located in H8 and in close spatial proximity to the C-terminus proceeding H12 and displayed a positive correlation to H3 and H12 in systems with DHT, but a negative correlation with H12 in systems with bicalutamide. In conclusion, our results suggest orthosteric AR antagonists to influence the intramolecular communication in agreement with results of a previous study [11]. Upon visual inspection of trajectories with decreased H12 helicity, we noticed a distinct conformational adaptation of F891 due to the steric pressure from the fluorinated phenyl ring of bicalutamide (Figure S10). Further, residues M895 and I899 presented differences in the respective simulations. All of these residues have been previously reported to adapt due to the presence of various antagonists and cause conformational rearrangements of H12 [10, 18, 32]. To follow up and quantify this change, we determined the per-residue RMSD of F891 (Figure S11). In all simulations with bicalutamide, F891 presented a statistically significant increase in RMSD as opposed to the remaining DHT-bound systems. The starting pose of bicalutamide for MD simulations obtained from induced-fit docking did not present an alternative rotamer of F891 compared to the DHT-bound crystal structure (Figure

S12). After one microsecond of MD simulation used to equilibrate the structure, however, a bicalutamide-induced change of F891 along with a displacement of H12 could be observed. The movement of F891 allowed a hydrophobic interaction with I899 to take place stabilizing the altered conformation of H12 along with a hydrogen bond between Q902 and E897. Thus, the altered conformation of H12 was stabilized by the bicalutamide-induced change of F891. In presence of an allosteric antagonist, the changes of F891 remained stable except one replica with AV-6. To further investigate the role of F891, we conducted a single simulation with the F891A mutation. While the helicity of H12 was intermediate between DHT-bound and bicalutamide-bound systems (Tables S7-S8), there was no displacement of the helix toward the coactivator binding site as opposed to the wild-type system (Figure S13). However, future studies will have to study the effects of this mutation, as we only performed a single simulation without any replicas.

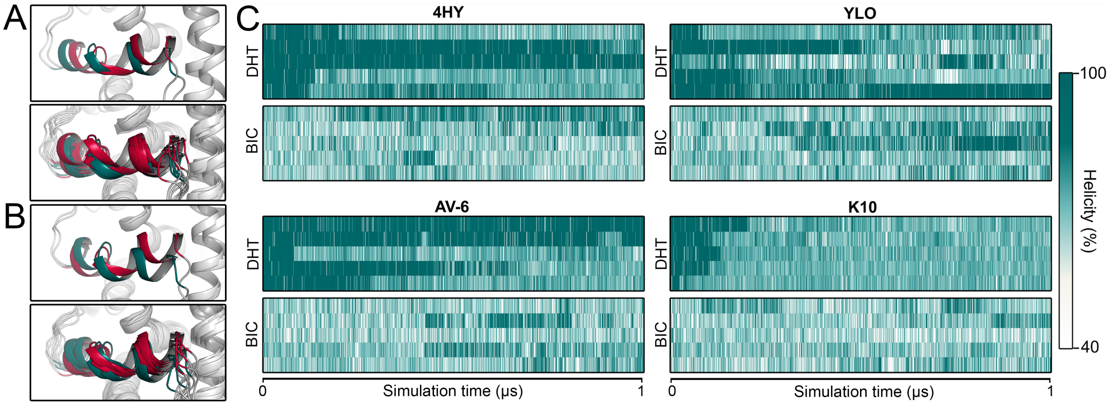


**Figure 4** Helicity of H12. (A) Representative structures show H12 in systems of K10 with DHT (pine green) and bicalutamide (red). On top, two structures with clear conformational change are shown, while the bottom part shows all representatives of the respective simulations. (B) Representative structures show H12 in systems of AV-6 with DHT (pine green) and bicalutamide (red). On top, two structures with clear conformational change are shown, while the bottom part shows all representatives of the respective simulations. (C) Helicity of H12 of all allosteric inhibitors with either DHT or bicalutamide. Combinations of allosteric inhibitors are shown in Figure S5.

We registered a decrease in backbone flexibility induced by bicalutamide in both AF-2 and BF-3 based on root-mean square fluctuation (RMSF) values compared to DHT-bound receptors (Figures 5A, 5B, and S14). Interestingly, these changes were more pronounced for the BF-3 site that did not severely suffer regarding the metrics describing ligand interactions. This indicated that BF-3 inhibitors might profit from the rigid-ification of their binding site, while it potentially penalizes some AF-2 inhibitors. In the AF-2, the most pronounced changes occurred for F725, R726, and V730. Interestingly, the F725L mutation of the AR is associated with partial androgen insensitivity syndrome and the R726L mutation was linked to PC [36]. Therefore, alterations in mobility of these residues might be involved in modifying the association of ligands and coactivator proteins at the AF-2 site as we observed it. The root-mean square deviation (RMSD) of the simulations indicated acceptable convergence of the simulations (Figure S15). In previous work on antagonist induced conformational changes of the AR, an expansion of the LBP was reported [18, 10] and the increased volume was linked to the displacement of H12 by orthosteric antagonists. In our analysis of the binding site volume, we observed a significantly higher volumes in all simulations with bicalutamide including the controls without allosteric inhibitors (Figures 5C and S5). As we likewise observed a displacement and adaptation of H12 with bicalutamide, which is bulkier than DHT, this might explain our observations. If this increase is only a consequence of the displacement due to the size and orientation of the antagonist molecule, or if it has functional consequences on the protein remains to be answered.

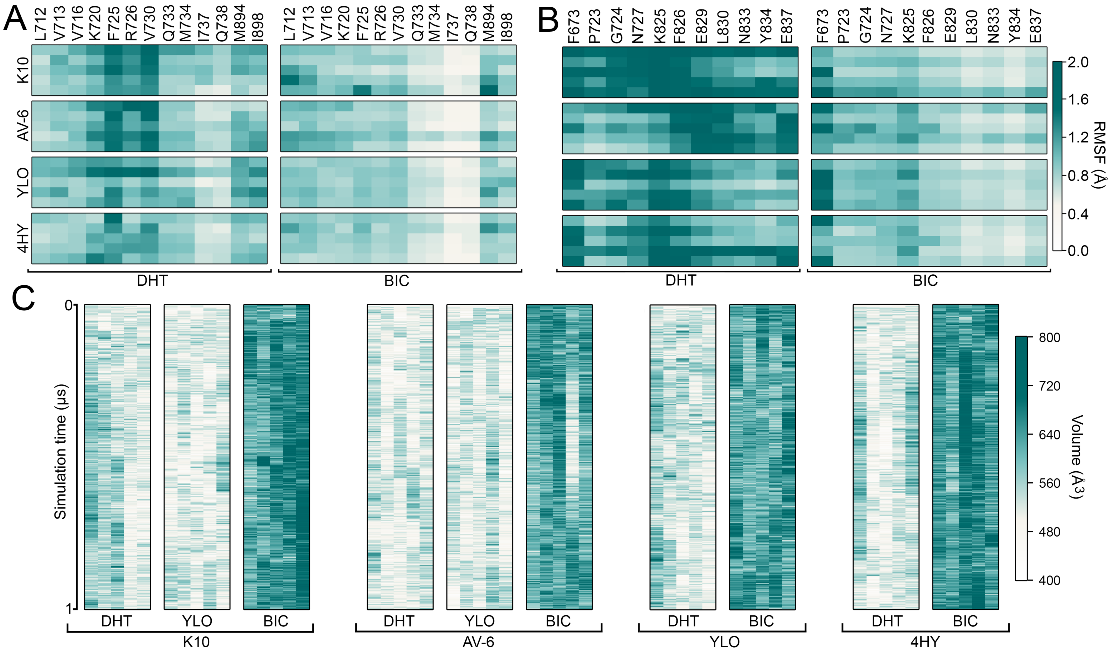


**Figure 5** Flexibility and binding site volumes. (A) Per-residue RMSF of the AF-2 site in different combinations. (B) Per-residue RMSF of the BF-3 site in different combinations. (C) Binding site volumes of the orthosteric site in simulations with different ligand combinations.

## Materials and Methods

**Model building** The crystal structure of the AR bound to DHT was retrieved from the Protein Data Bank [37] due to is excellent resolution of 1.55 Å (PDB ID: 3L3X). To assure that the receptor was modeled according to its wild-type sequence, the FASTA file of the AR was obtained from the UniProt database [38] (Uniprot ID: P10275) and compared to the amino acids of the obtained crystal structure by sequence alignment with ClustalW [39] in the UGENE (v1.32.0) toolkit [40]. Based on this assessment, we introduced a S669C mutation to restore the wild-type sequence using the 3D Builder panel within the Maestro Small-Molecule Drug Discovery Suite (v2019-3) [41]. Next, the protein was treated with the Protein Preparation Wizard [42] within Maestro to add hydrogen atoms, assign bond orders, and predict protonation states with Epik at pH 7.4. We reoriented the hydrogen bonding network using PROPKA at pH 7.4 and subjected the system to a restrained minimization using the OPLS_2005 force field with a convergence threshold of 0.3 Å for protein heavy atoms. In a final step, the cocrystallized coactivator fragment was removed from the AF-2 site.

Unfortunately, no wild-type structure of the AR cocrystallized with an antagonist is available to date. Since induced-fit is generally required in order to accommodate antagonists in the AR LBP [2], we obtained the starting conformation of bicalutamide using the DOLINA induced-fit docking software [43]. The coordinates of bicalutamide were obtained from the PubChem database (PubChem CID: 2375) [44] and treated with the LigPrep routine in Maestro to predict its protonation state at pH 7.4 and to obtain an energy-minimized conformer using the OPLS3e force field. As bicalutamide is prescribed in the racemic form [45], we selected the energetically more favorable (Table S9) (S)-isomer for our study. We validated the obtained best scored pose with crystallographic data and obtained an RMSD value of 1.75 Å to our pose despite the W741L mutation in the crystal structure (Figure S16). Further, the starting conformations of AF-2 and BF-3 allosteric ligands were obtained from their crystal structures (Table S10) by superposition to our model. We avoided steric clashes by manually adapting side chain torsions of the protein and removing interfering water molecules. These crystal structures were also treated with the Protein Preparation Wizard as described above to obtain correct bond orders and protonation according to the crystallographic information.

**MD simulations** All MD simulations in this study were performed using the Desmond (v2019-1) simulation engine [46]. The orthorhombic periodic boundary systems with a buffer of 10 Å to the next protein atom were solvated with TIP3P solvent molecules. After the default equilibration protocol of Desmond, all simulations were conducted in an NPT ensemble at a temperature of 310 K maintained by the Nose-Hoover thermostat and atmospheric pressure regulated by the Martyna-Tobias-Klein barostat. We selected the OPLS_2005 force field and a time step of 2 fs for the RESPA integrator. Long-range interactions were treated with the u-series algorithm [47] and bonds to hydrogen atoms were restrained using the M-SHAKE algorithm. Short-range interactions were cut off at 9 Å. Each system was simulated for 1 $\mu$s with five replica simulations by altering the random seed for the initial velocities with atomic coordinates recorded at an interval of 1 ns. The complex with bicalutamide was pre-equilibrated for 1 $\mu$s and clustered using the trj_cluster.py script (highest occupied cluster) that comes with Maestro before the complexes with allosteric inhibitors were built. In the clustering routine we limited the number of output clusters to 15 and information on the cluster population and the total number of clusters is given in Tables S11 and S12. Further, we ran five replicas for the AR bound to DHT and bicalutamide without any allosteric ligand present resulting in a total of 60 $\mu$s MD simulations (Table 2). In addition, we conducted a single microsecond simulation of the AR bound to bicalutamide with the F891A mutation. The mutation was introduced in the 3D Builder panel in Maestro.

**Table 2** Simulations of polypharmacology combinations.

| Orthosteric | AF-2 | BF-3 | Replicas |
|---|---|---|---|
| DHT | K10 | none | 5 |
| | AV-6 | none | 5 |
| | K10 | YLO | 5 |
| | AV-6 | YLO | 5 |
| | none | 4HY | 5 |
| | none | YLO | 5 |
| | none | none | 5 |
| BIC | K10 | none | 5 |
| | AV-6 | none | 5 |
| | none | 4HY | 5 |
| | none | YLO | 5 |
| | none | none | 5 |
| BIC F891A | none | none | 1 |

**Evaluation of the MD trajectories** The RMSD, RMSF, number of hydrogen bonds, and ligand SASA values of the protein were obtained in the Simulation Interaction Diagram panel within Maestro. We computed the binding free energies for each allosteric ligand using the thermal_mmgbsa.py script provided with Maestro. This analysis was conducted for the last 100 ns of the simulations. The residence of the ligands within their binding site was calculated with an in-house python routine measuring the distance between the centroid of each ligand and the $\alpha$-carbon of a protein residue located at the respective site (M734 at AF-2 and F673 at BF-3). Based on a visual inspection of the trajectories, the ligand was considered to be bound if the distance was below 14 Å or 16 Å to for the AF-2 site and BF-3 site respectively since a comparably buried $\alpha$-carbon atom was selected. Ligand-protein interactions were evaluated based on the data obtained from the Simulation Interaction Diagram panel in Maestro. The occurrence of a specific interaction was summed up over all replicas (except the ones presenting dissociation of the ligand) for the last 200 ns of each simulation. These counts were normalized to the number of assessed frames and listed in a table if the obtained percentage amounted to at least 20%. Interactions that occurred at least 10% were included in a binary fingerprint to compute the similarity index of contacts between different systems. To quantify the helicity of H3, H10, and H12, we used the STRIDE (v29.01.96) [48] program which assigns a particular secondary structure to each protein residue. The analysis was performed on all MD frames and 3-10 helices were regarded as helical. Residues that were defined as boundaries for each helix to determine the relative percentage of helical residues within a secondary structure element are listed in Table S13. Mechanistic insights into the allosteric communication within the protein were obtained using MD-TASK (v1.0.1) [35]. In particular, we determined the dynamic cross-correlation map (DCCM) using the calc_correlation.py script in MD-TASK for every second frame of four trajectories selected from visual inspection. Similarly, we determined the BC for every trajectory frame using the calc_network.py script in MD-TASK followed by the avg_network.py script averaging the data over the trajectories. The volume of the LBP was estimated using the POVME (v2.0) algorithm [49] and the starting points for the inclusion spheres with a radius of 12 Å were determined according to the centroid of the orthosteric ligands (DHT or bicalutamide). Further, a grid spacing of 0.5 Å and a distance cutoff of 0.8 Å were selected for volume calculations. Statistical significance of the average values was performed using the ttest_ind_from_stats routine in the python-scipy module. For these calculations, the DHT-bound systems were taken as a reference. For each state (e.g. K10 with bicalutamide), we summarized the individual replica simulations (for the last 100 ns) to a single dataset and calculated average and standard deviation as proposed by Knapp and colleagues [50]. Based on these metrics, we assessed the statistical significance with the total number of observations set to the sum of measurements over five replicas. We selected a significance level of $p=0.05$.

## Conclusion

PC remains a challenging disease regarding pharmacological intervention and displays a high mortality due to resistance mechanisms. Alternative therapeutics targeting allosteric sites offer an attractive approach to treat the disease. Until today, however, the potential combination of allosteric inhibitors with conventional antiandrogens or their concurrent administration was not considered even though it was recommended without further investigation and polypharmacology is abundant in cancer treatment. Based on a detailed computational analysis using microsecond MD simulations, we provided recommendations for the combination of these compounds and explored the mechanistic background. The structural adaptation and displacement of H12 is likely involved in the negative effects of classical AR antagonists on the binding of AF-2 inhibitors ,even though the impact was lower than expected. In contrast, the BF-3 inhibitors did not suffer in binding energy or solvent accessibility in the studied combinations. By considering our recommendations, laboratory experiments could be rationalized to optimize the treatment of PC.

## References

[1] Takahiro Matsumoto, Matomo Sakari, Maiko Okada, Atsushi Yokoyama, Sayuri Takahashi, Alexander Kouzmenko, Shigeaki Kato, and Ph75ch09-Kato Ari. The Androgen Receptor in Health and Disease. *Annual Review of Physiology*, 75(1):201–224, 2013.

[2] Joel Wahl and Martin Smieško. Endocrine disruption at the androgen receptor: Employing molecular dynamics and docking for improved virtual screening and toxicity prediction. *International Journal of Molecular Sciences*, 19(6), 2018.

[3] Zoran Culig. Molecular Mechanisms of Enzalutamide Resistance in Prostate Cancer. *Current Molecular Biology Reports*, 3(4):230–235, 2017.

[4] André Fischer and Martin Smieško. Allosteric binding sites on nuclear receptors: Focus on drug efficacy and selectivity. *International Journal of Molecular Sciences*, 21(2):6–8, 2020.

[5] Shahriar Koochekpour. Androgen receptor signaling and mutations in prostate cancer. *Asian Journal of Andrology*, 12(5):639–657, 2010.

[6] Takuma Uo, Stephen R. Plymate, and Cynthia C. Sprenger. Allosteric alterations in the androgen receptor and activity in prostate cancer. *Endocrine-related cancer*, 24(9):R335–R348, 2017.

[7] Guillermo Martinez-Ariza and Christopher Hulme. Recent advances in allosteric androgen receptor inhibitors for the potential treatment of castration-resistant prostate cancer. *Pharmaceutical patent analyst*, 4(5):387–402, 2015.

[8] Víctor Buzón, Laia R. Carbó, Sara B. Estruch, Robert J. Fletterick, and Eva Estébanez-Perpiñ. A conserved surface on the ligand binding domain of nuclear receptors for allosteric control. *Molecular and Cellular Endocrinology*, 348(2):394–402, 2012.

[9] Peter Axerio-Cilies, Nathan A Lack, M Ravi Shashi Nayana, Ka Hong Chan, Anthony Yeung, Eric Leblanc, Emma S.Tomlinson Guns, Paul S Rennie, and Artem Cherkasov. Inhibitors of androgen receptor activation function-2 (AF2) site identified through virtual screening. *Journal of Medicinal Chemistry*, 54(18):6197–6205, 2011.

[10] D. J. Osguthorpe and A. T. Hagler. Mechanism of androgen receptor antagonism by bicalutamide in the treatment of prostate cancer. *Biochemistry*, 50(19):4105–4113, 2011.

[11] Ye Jin, Mojie Duan, Xuwen Wang, Xiaotian Kong, Wenfang Zhou, Huiyong Sun, Hui Liu, Dan Li, Huidong Yu, Youyong Li, and Tingjun Hou. Communication between the Ligand-Binding Pocket and the Activation Function-2 Domain of Androgen Receptor Revealed by Molecular Dynamics Simulations. *Journal of Chemical Information and Modeling*, 59 (2):842–857, 2019.

[12] Duan Ni, Yun Li, Yuran Qiu, Jun Pu, Shaoyong Lu, and Jian Zhang. Combining Allosteric and Orthosteric Drugs to Overcome Drug Resistance. *Trends in Pharmacological Sciences*, 41(5):336–348, 2020.

[13] Ingolf Cascorbi. Drug Interactions—Principles, Examples and Clinical Consequences. *Deutsches Ärzteblatt International*, 109(33-34):546–556, 8 2012.

[14] Reza Bayat Mokhtari, Tina S. Homayouni, Narges Baluch, Evgeniya Morgatskaya, Sushil Kumar, Bikul Das, and Herman Yeger. Combination therapy in combating cancer. *Oncotarget*, 8(23):38022–38043, 2015.

[15] Nicholas D. James, Matthew R. Sydes, Noel W. Clarke, Malcolm D. Mason, David P. Dearnaley, Melissa R. Spears, Alastair W.S. S Ritchie, Christopher C. Parker, J. Martin Russell, Gerhardt Attard, Johann De Bono, William Cross, Rob J. Jones, George Thalmann, Claire Amos, David Matheson, Robin Millman, Mymoona Alzouebi, Sharon Beesley, Alison J. Birtle, Susannah Brock, Richard Cathomas, Prabir Chakraborti, Simon Chowdhury, Audrey Cook, Tony Elliott, Joanna Gale, Stephanie Gibbs, John D. Graham, John Hetherington, Robert Hughes, Robert Laing, Fiona McKinna, Duncan B. McLaren, Joe M. O'Sullivan, Omi Parikh, Clive Peedell, Andrew Protheroe, Angus J. Robinson, Narayanan Srihari, Rajaguru Srinivasan, John Staffurth, Santhanam Sundar, Shaun Tolan, David Tsang, John Wagstaff, and Mahesh K.B. B Parmar. Addition of docetaxel, zoledronic acid, or both to first-line long-term hormone therapy in prostate cancer (STAMPEDE): Survival results from an adaptive, multiarm, multistage, platform randomised controlled trial. *The Lancet*, 387(10024):1163–1177, 2016.

[16] João Marcelo Lamim Ribeiro and Marta Filizola. Insights From Molecular Dynamics Simulations of a Number of G-Protein Coupled Receptor Targets for the Treatment of Pain and Opioid Use Disorders. *Frontiers in Molecular Neuroscience*, 12(August):1–13, 2019.

[17] A. Baldi. Computational approaches for drug design and discovery: An overview. *Systematic Reviews in Pharmacy*, 1(1):99–105, 2010.

[18] Sugunadevi Sakkiah, Rebecca Kusko, Bohu Pan, Wenjing Guo, Weigong Ge, Weida Tong, and Huixiao Hong. Structural changes due to antagonist binding in ligand binding pocket

of androgen receptor elucidated through molecular dynamics simulations. *Frontiers in Pharmacology*, 9(MAY):1–13, 2018.

[19] Zahoor Ahmad Bhat, Dheeraj Chitara, Jawed Iqbal1, Sanjeev. B.S., and Arumugam Madhumalar. Targeting Allosteric Pockets of SARS-CoV-2 Main Protease Mpro. 2020.

[20] Ying-Chih Chiang, Mabel T Y Wong, and Jonathan W Essex. Molecular Dynamics Simulations of Antibiotic Ceftaroline at the Allosteric Site of Penicillin-Binding Protein 2a (PBP2a). *Israel Journal of Chemistry*, 60(7):754–763, 7 2020.

[21] Yi Shang, Holly R Yeatman, Davide Provasi, Andrew Alt, Arthur Christopoulos, Meritxell Canals, and Marta Filizola. Proposed Mode of Binding and Action of Positive Allosteric Modulators at Opioid Receptors. *ACS Chemical Biology*, 11(5):1220–1229, 5 2016.

[22] Ahmed A El Rashedy, Fisayo A Olotu, and Mahmoud E S Soliman. Dual Drug Targeting of Mutant Bcr-Abl Induces Inactive Conformation: New Strategy for the Treatment of Chronic Myeloid Leukemia and Overcoming Monotherapy Resistance. *Chemistry and biodiversity*, 15(3):e1700533, 3 2018.

[23] Freddie R Salsbury Jr. Molecular Dynamics Simulations of Protein Dynamics and their relevance To Drug Discovery. *Current Opinion in Pharmacology*, 10(6):738–744, 2011.

[24] Nathan A. Lack, Peter Axerio-Cilies, Peyman Tavassoli, Frank Q. Han, Ka Hong Chan, Clementine Feau, Eric LeBlanc, Emma Tomlinson Guns, R. Kiplin Guy, Paul S. Rennie, and Artem Cherkasov. Targeting the binding function 3 (BF3) site of the human androgen receptor through virtual screening. *Journal of Medicinal Chemistry*, 54(24):8563–8573, 2011.

[25] Ravi Shashi Nayana Munuganti, Eric Leblanc, Peter Axerio-Cilies, Christophe Labriere, Kate Frewin, Kriti Singh, Mohamed D H Hassona, Nathan A Lack, Huifang Li, Fuqiang Ban, Emma Tomlinson Guns, Robert Young, Paul S Rennie, and Artem Cherkasov. Targeting the binding function 3 (BF3) site of the androgen receptor through virtual screening. 2. Development of 2-((2-phenoxyethyl) thio)-1H-benzimidazole derivatives. *Journal of Medicinal Chemistry*, 56(3):1136–1148, 2013.

[26] Yang Cao and Lei Li. Improved protein–ligand binding affinity prediction by using a curvature-dependent surface-area model. *Bioinformatics*, 30(12):1674–1680, 6 2014.

[27] Johan Aqvist, Victor B. Luzhkov, and Bjørn O. Brandsdal. Ligand binding affinities from MD simulations. *Accounts of Chemical Research*, 35(6):358–365, 2002.

[28] E Estebanez-Perpina, L A Arnold, P Nguyen, E D Rodrigues, E Mar, R Bateman, P Pallai, K M Shokat, J D Baxter, R K Guy, P Webb, and R J Fletterick. A surface on the androgen receptor that allosterically regulates coactivator binding. *Proceedings of the National Academy of Sciences*, 104(41):16074–16079, 2007.

[29] Mattia Bernetti, Matteo Masetti, Walter Rocchia, and Andrea Cavalli. Kinetics of Drug Binding and Residence Time. *Annual Review of Physical Chemistry*, 70:143–171, 2019.

[30] Samuel D. Lotz and Alex Dickson. Unbiased Molecular Dynamics of 11 min Timescale Drug Unbinding Reveals Transition State Stabilizing Interactions. *Journal of the American Chemical Society*, 140(2):618–628, 2018.

[31] Yibing Shan, Eric T. Kim, Michael P. Eastwood, Ron O. Dror, Markus A. Seeliger, and David E. Shaw. How does a drug molecule find its target binding site? *Journal of the American Chemical Society*, 133(24):9181–9183, 2011.

[32] Na Liu, Wenfang Zhou, Yue Guo, Junmei Wang, Weitao Fu, Huiyong Sun, Dan Li, Mojie Duan, and Tingjun Hou. Molecular Dynamics Simulations Revealed the Regulation of Ligands to the Interactions between Androgen Receptor and its Coactivator. *Journal of Chemical Information and Modeling*, 58:1652–1661, 2018.

[33] Mojie Duan, Na Liu, Wenfang Zhou, Dan Li, Minghui Yang, and Tingjun Hou. Structural Diversity of Ligand-Binding Androgen Receptors Revealed by Microsecond Long Molecular Dynamics Simulations and Enhanced Sampling. *Journal of Chemical Theory and Computation*, 12(9):4611–4619, 2016.

[34] Katja Jehle, Laura Cato, Antje Neeb, Claudia Muhle-Goll, Nicole Jung, Emmanuel W. Smith, Victor Buzon, Laia R. Carbó, Eva Estebanez-Perpina, Katja Schmitz, Ljiljana Fruk, Burkhard Luy, Yu Chen, Marc B. Cox, Stefan Brase, Myles Brown, and Andrew C.B. Cato. Coregulator control of androgen receptor action by a novel nuclear receptor-binding motif. *Journal of Biological Chemistry*, 289(13):8839–8851, 2014.

[35] David K Brown, David L Penkler, Olivier Sheik Amamuddy, Caroline Ross, Ali Rana Atilgan, Canan Atilgan, and Oezlem Tastan Bishop. MD-TASK: a software suite for analyzing molecular dynamics trajectories. *Bioinformatics (Oxford, England)*, 33(17):2768–2771, 9 2017.

[36] James Thompson, Fahri Saatcioglu, Olli A Jaenne, and Jorma J Palvimo. Disrupted Amino- and Carboxyl-Terminal Interactions of the Androgen Receptor Are Linked to Androgen Insensitivity. *Molecular Endocrinology*, 15(6):923–935, 6 2001.

[37] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, T N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 1 2000.

[38] The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1):D506–D515, 1 2019.

[39] Julie D. Thompson, Desmond G. Higgins, and Toby J. Gibson. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting,

position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22(22): 4673–4680, 1994.

[40] Konstantin Okonechnikov, Olga Golosova, Mikhail Fursov, Alexey Varlamov, Yuri Vaskin, Ivan Efremov, O. G. German Grehov, Denis Kandrov, Kirill Rasputin, Maxim Syabro, and Timur Tleukenov. Unipro UGENE: A unified bioinformatics toolkit. *Bioinformatics*, 28 (8):1166–1167, 2012.

[41] Schrödinger LCC. Maestro Small-Molecule Drug Discovery Suite 2019-3. 2019.

[42] G. Madhavi Sastry, Matvey Adzhigirey, Tyler Day, Ramakrishna Annabhimoju, and Woody Sherman. Protein and ligand preparation: Parameters, protocols, and influence on virtual screening enrichments. *Journal of Computer-Aided Molecular Design*, 27(3): 221–234, 2013.

[43] Martin Smieško. DOLINA – Docking Based on a Local Induced-Fit Algorithm: Application toward Small-Molecule Binding to Nuclear Receptors. *Journal of Chemical Information and Modeling*, 53(6):1415–1423, 6 2013.

[44] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan E Bolton. PubChem 2019 update: improved access to chemical data. *Nucleic acids research*, 47(D1):D1102–D1109, 1 2019.

[45] Henning Kaemmerer, Zoltan Horvath, Ju Weon Lee, Malte Kaspereit, Robert Arnell, Martin Hedberg, Björn Herschend, Matthew J Jones, Kerstin Larson, Heike Lorenz, and Andreas Seidel-Morgensten. Separation of Racemic Bicalutamide by an Optimized Combination of Continuous Chromatography and Selective Crystallization. *Organic Process Research and Development*, 16(2):331–342, 2 2012.

[46] Kevin Bowers, Edmond Chow, Huafeng Xu, Ron Dror, Michael Eastwood, Brent Gregersen, John Klepeis, Istvan Kolossvary, Mark Moraes, Federico Sacerdoti, John Salmon, Yibing Shan, and David Shaw. Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters. *ACM/IEEE SC 2006 Conference (SC'06)*, (November):43, 2006.

[47] David E. Shaw, J. P. Grossman, Joseph A. Bank, Brannon Batson, J. Adam Butts, Jack C. Chao, Martin M. Deneroff, Ron O. Dror, Amos Even, Christopher H. Fenton, Anthony Forte, Joseph Gagliardo, Gennette Gill, Brian Greskamp, C. Richard Ho, Douglas J. Ierardi, Lev Iserovich, Jeffrey S. Kuskin, Richard H. Larson, Timothy Layman, Li Siang Lee, Adam K. Lerer, Chester Li, Daniel Killebrew, Kenneth M. Mackenzie, Shark Yeuk Hai Mok, Mark A. Moraes, Rolf Mueller, Lawrence J. Nociolo, Jon L. Peticolas, Terry Quan, Daniel Ramot, John K. Salmon, Daniele P. Scarpazza, U. Ben Schafer, Naseer Siddique, Christopher W. Snyder, Jochen Spengler, Ping Tak Peter Tang, Michael Theobald, Horia Toma, Brian Towles, Benjamin Vitale, Stanley C. Wang, and Cliff Young. Anton 2:

Raising the Bar for Performance and Programmability in a Special-Purpose Molecular Dynamics Supercomputer. *International Conference for High Performance Computing, Networking, Storage and Analysis, SC*, 2015-Janua(January):41–53, 2014.

[48] D Frishman and P Argos. Knowledge-based protein secondary structure assignment. *Proteins*, 23(4):566–579, 12 1995.

[49] Jacob D. Durrant, Lane Votapka, Jesper Sørensen, and Rommie E. Amaro. POVME 2.0: An enhanced tool for determining pocket shape and volume characteristics. *Journal of Chemical Theory and Computation*, 10(11):5047–5056, 2014.

[50] Bernhard Knapp, Luis Ospina, and Charlotte M Deane. Avoiding False Positive Conclusions in Molecular Simulation: The Importance of Replicas. *Journal of Chemical Theory and Computation*, 14(12):6127–6138, 12 2018.

## 9.1 Supporting Information

## Supporting Results and Discussion

**Table S 1** Mean and standard deviation values used for computation of statistical significance of ligand-based metrics.

| Ligand | Combination | $\Delta G_{MM/GBSA}$ (kcal/mol) | SASA (Å$^2$) |
|---|---|---|---|
| K10 | DHT | -46.5 ± 6.0 (*n*=400) | 143.4 ± 24.5 (*n*=400) |
| | BIC | -46.4 ± 4.0 (*n*=400) | 148.4 ± 16.7 (*n*=400) |
| | YLO | -41.7 ± 4.6 (*n*=500) | 157.1 ± 26.7 (*n*=500) |
| AV-6 | DHT | -40.2 ± 9.2 (*n*=300) | 144.1 ± 51.8 (*n*=300) |
| | BIC | -39.5 ± 4.1 (*n*=400) | 150.3 ± 18.6 (*n*=400) |
| | YLO | -43.7 ± 6.8 (*n*=400) | 123.5 ± 38.7 (*n*=400) |
| YLO | DHT | -54.7 ± 12.6 (*n*=500) | 241.7 ± 57.7 (*n*=500) |
| | BIC | -60.7 ± 11.7 (*n*=500) | 184.1 ± 83.3 (*n*=500) |
| | AV-6 | -57.7 ± 9.7 (*n*=500) | 223.7 ± 54.5 (*n*=500) |
| | K10 | -54.9 ± 12.8 (*n*=500) | 205.3 ± 51.5 (*n*=500) |
| 4HY | DHT | -37.8 ± 8.0 (*n*=500) | 236.7 ± 37.4 (*n*=500) |
| | BIC | -37.7 ± 4.9 (*n*=500) | 241.2 ± 27.3 (*n*=500) |

*n* represents for the number of data points included in the statistical analysis.

**Table S 2** Statistical significance of averages at p=0.05 level.

| Comparison | $\Delta G_{MM/GBSA}$ | SASA | H12 Helicity | Site volume | F891 RMSD |
|---|---|---|---|---|---|
| K10 and BIC | no | yes | yes | yes | yes |
| K10 and YLO | yes | yes | no | yes | yes |
| AV-6 and BIC | no | yes | yes | yes | yes |
| AV-6 and YLO | yes | yes | no | no | no |
| YLO and BIC | yes | yes | yes | yes | yes |
| YLO and AV-6 | yes | yes | n/a[a] | NaN[a] | n/a[a] |
| YLO and K10 | no | yes | n/a[a] | n/a[a] | NaN[a] |
| 4HY and BIC | no | yes | yes | yes | yes |

[a]Value for this system reported above, as it is not ligand-dependent.

**Table S 3** Ligand-protein interactions of K10.

| Combination | Residue | Type | Prevalence (%) |
|---|---|---|---|
| DHT | V716 | Hydrophobic | 41.8 |
|  | V730 | Hydrophobic | 32.9 |
|  | M734 | Hydrophobic | 26.0 |
|  | Q733 | Hydrogen bond | 20.4 |
| BIC | Q733 | Hydrogen bond | 42.3 |
|  | M894 | Hydrophobic | 32.0 |
|  | V716 | Hydrophobic | 27.6 |
|  | M734 | Hydrophobic | 24.6 |
| YLO | V716 | Hydrophobic | 36.5 |
|  | M734 | Hydrophobic | 24.8 |

The most prevalent ligand-protein interactions (more than 20%) are displayed.

**Table S 4** Ligand-protein interactions of AV-6.

| **Combination** | **Residue** | **Type** | **Prevalence (%)** |
|---|---|---|---|
| DHT | E709 | Hydrogen bond | 73.3 |
| | D731 | Hydrogen bond | 54.8 |
| | V716 | Hydrophobic | 34.7 |
| | V730 | Hydrogen bond | 33.7 |
| | L712 | Hydrogen bond | 29.0 |
| | M734 | Hydrophobic | 27.7 |
| | M894 | Hydrophobic | 25.2 |
| | I898 | Hydrophobic | 21.8 |
| BIC | V716 | Hydrophobic | 43.8 |
| | V730 | Hydrogen bond | 38.8 |
| | E897 | Hydrogen bond | 31.1 |
| | V730 | Hydrophobic | 30.3 |
| | M894 | Hydrophobic | 29.0 |
| | M734 | Hydrophobic | 22.6 |
| | E893 | Hydrogen bond | 22.5 |
| | D731 | Hydrogen bond | 21.6 |
| YLO | E709 | Hydrogen bond | 100.0 |
| | V716 | Hydrophobic | 52.8 |
| | V730 | Hydrogen bond | 44.1 |
| | L712 | Hydrogen bond | 29.6 |
| | M734 | Hydrophobic | 28.3 |
| | M894 | Hydrogen bond | 25.6 |
| | I898 | Hydrophobic | 21.6 |

The most prevalent ligand-protein interactions (more than 20%) are displayed.

**Table S 5** Ligand-protein interactions of 4HY.

| Combination | Residue | Type | Prevalence (%) |
|---|---|---|---|
| DHT | R726 | Hydrogen bond | 100.0 |
| | N727 | Hydrogen bond | 57.4 |
| | G724 | Hydrogen bond | 44.8 |
| | E837 | Hydrogen bond | 43.4 |
| | R840 | Hydrogen bond | 36.3 |
| | P723 | Hydrophobic | 27.3 |
| BIC | R726 | Hydrogen bond | 100.0 |
| | N727 | Hydrogen bond | 43.2 |
| | F826 | Hydrophobic | 36.1 |
| | E829 | Hydrogen bond | 33.1 |
| | N833 | Hydrogen bond | 29.1 |

The most prevalent ligand-protein interactions (more than 20%) are displayed.

**Table S 6** Ligand-protein interactions of YLO.

| Combination | Residue | Type | Prevalence (%) |
|---|---|---|---|
| DHT | F826 | Hydrophobic | 49.9 |
| | Y834 | Hydrophobic | 43.7 |
| | F673 | Hydrophobic | 30.2 |
| BIC | Y834 | Hydrophobic | 50.9 |
| | P723 | Hydrophobic | 31.7 |
| | F826 | Hydrophobic | 24.9 |
| | L674 | Hydrophobic | 24.1 |
| | L722 | Hydrophobic | 23.5 |
| | F673 | Hydrophobic | 22.1 |
| K10 | Y834 | Hydrophobic | 97.4 |
| | F673 | Hydrophobic | 67.3 |
| | F826 | Hydrophobic | 29.4 |
| AV-6 | F673 | Hydrophobic | 32.1 |
| | Y834 | Hydrophobic | 24.0 |

The most prevalent ligand-protein interactions (more than 20%) are displayed.

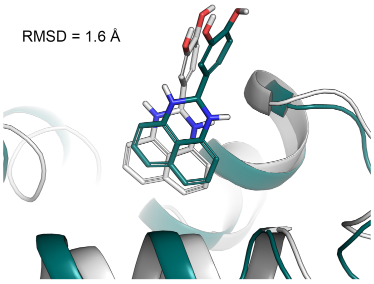


**Figure S 1** Superposition of ligand-protein complexes before and after the dissociation-reassociation event of AV-6.

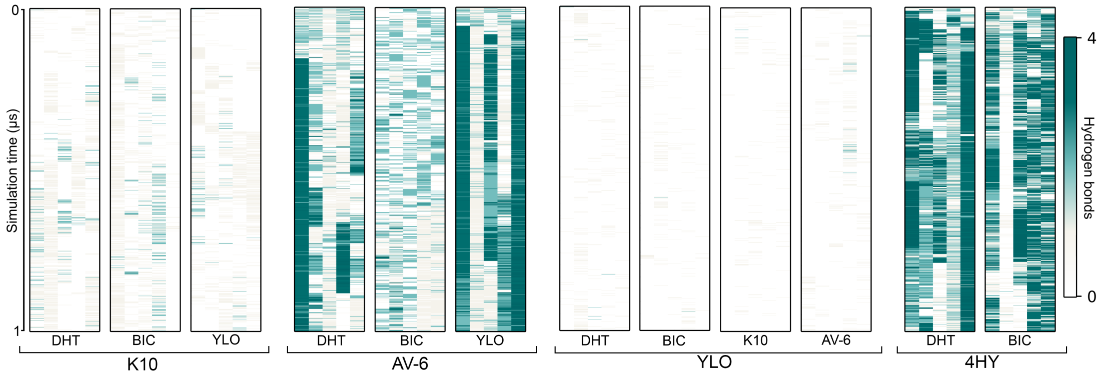


**Figure S 2** Number of ligand-protein hydrogen bonds in all simulations with allosteric ligands.

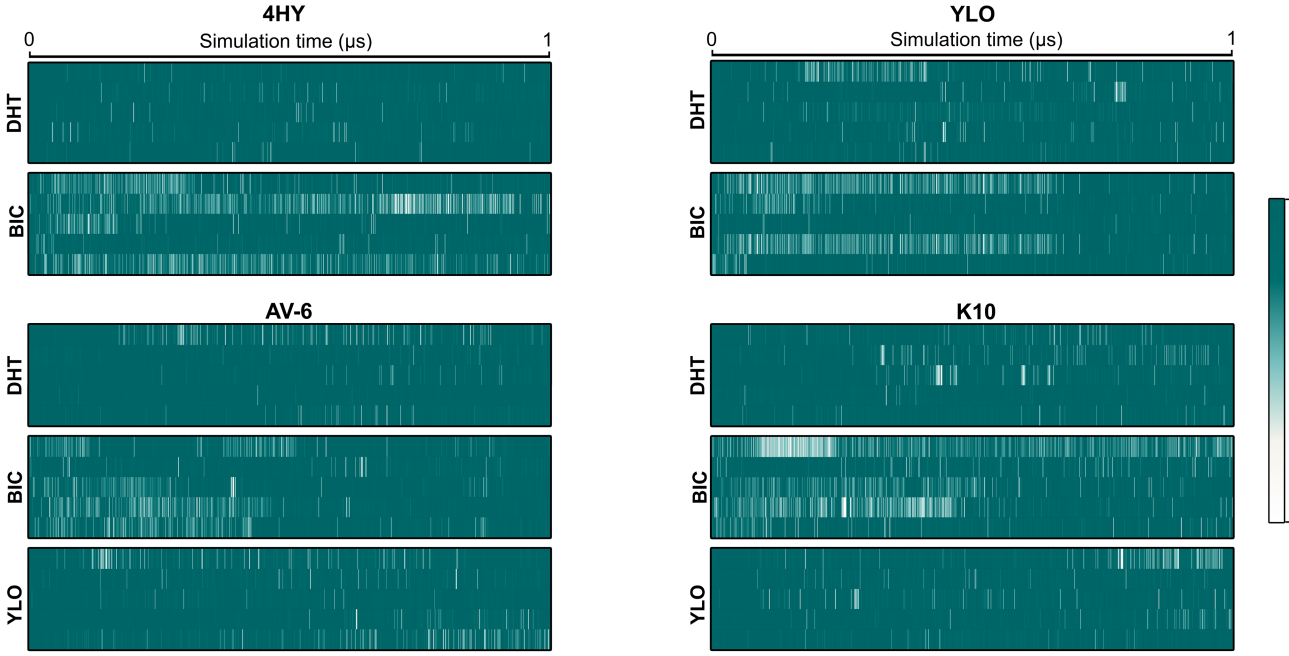


**Figure S 3** The helicity of H3 for all combination simulations.

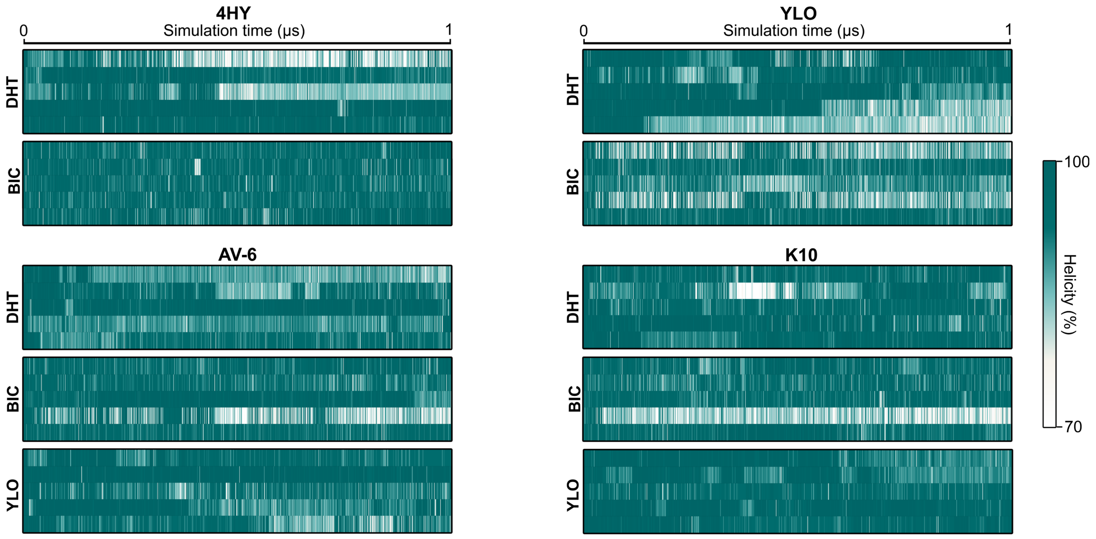


**Figure S 4** The helicity of H10 for all combination simulations.

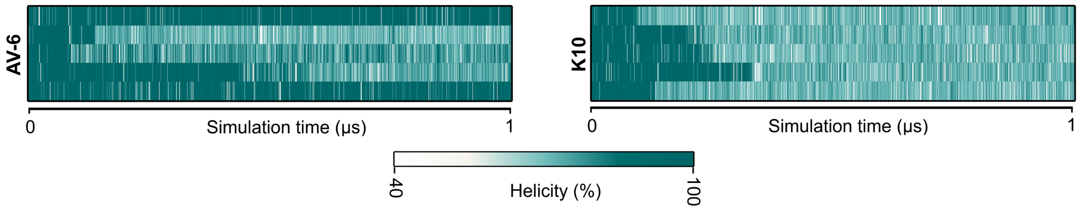


**Figure S 5** The helicity H12 in simulations combining AF-2 and BF-3 inhibitors.

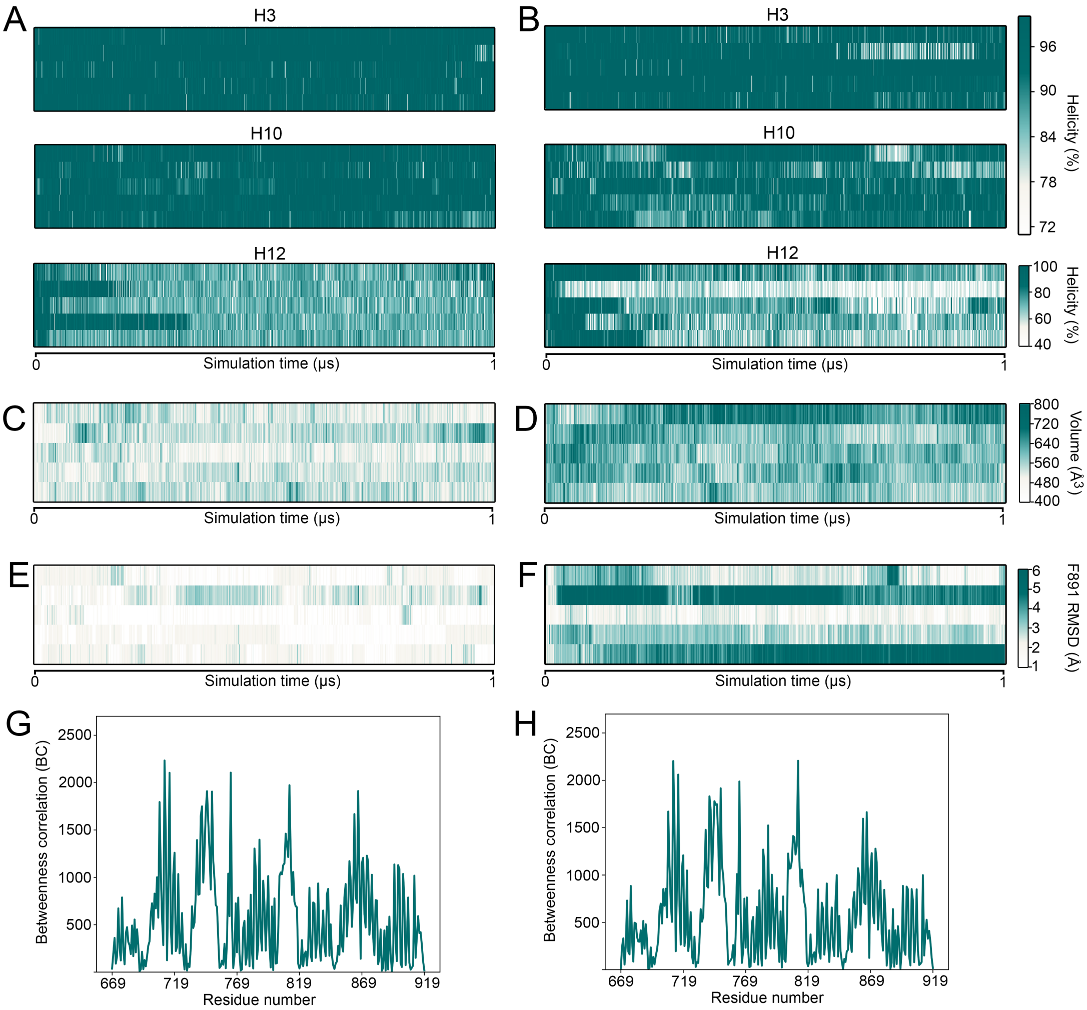


**Figure S 6** Data of the control simulations without allosteric ligands. (A) Helicity in simulations with DHT. (B) Helicity in simulations with bicalutamide. (C) Binding site volume in simulations with DHT. (D) Binding site volume in simulations with bicalutamide. (E) RMSD of F891 in simulations with DHT. (F) RMSD of F891 in simulations with bicalutamide. (G) Average BC analysis in simulations with DHT. (H) Average BC analysis in simulations with bicalutamide.

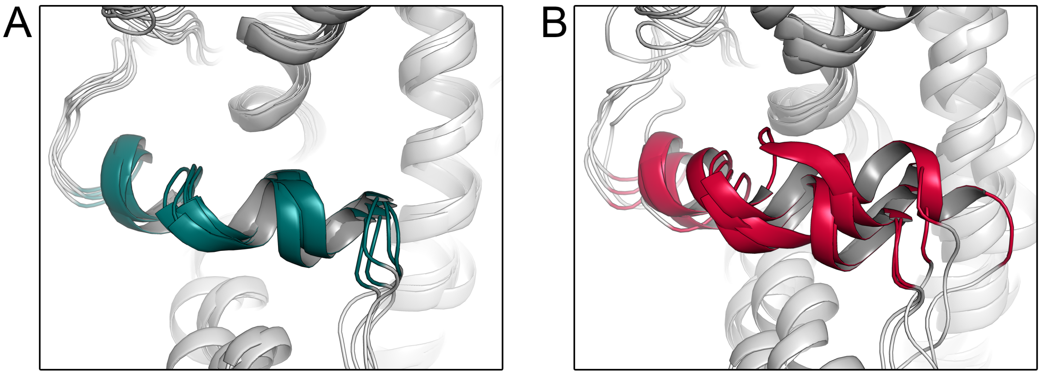


**Figure S 7** H12 in representative structures of control simulations with (A) DHT and (B) bicalutamide.

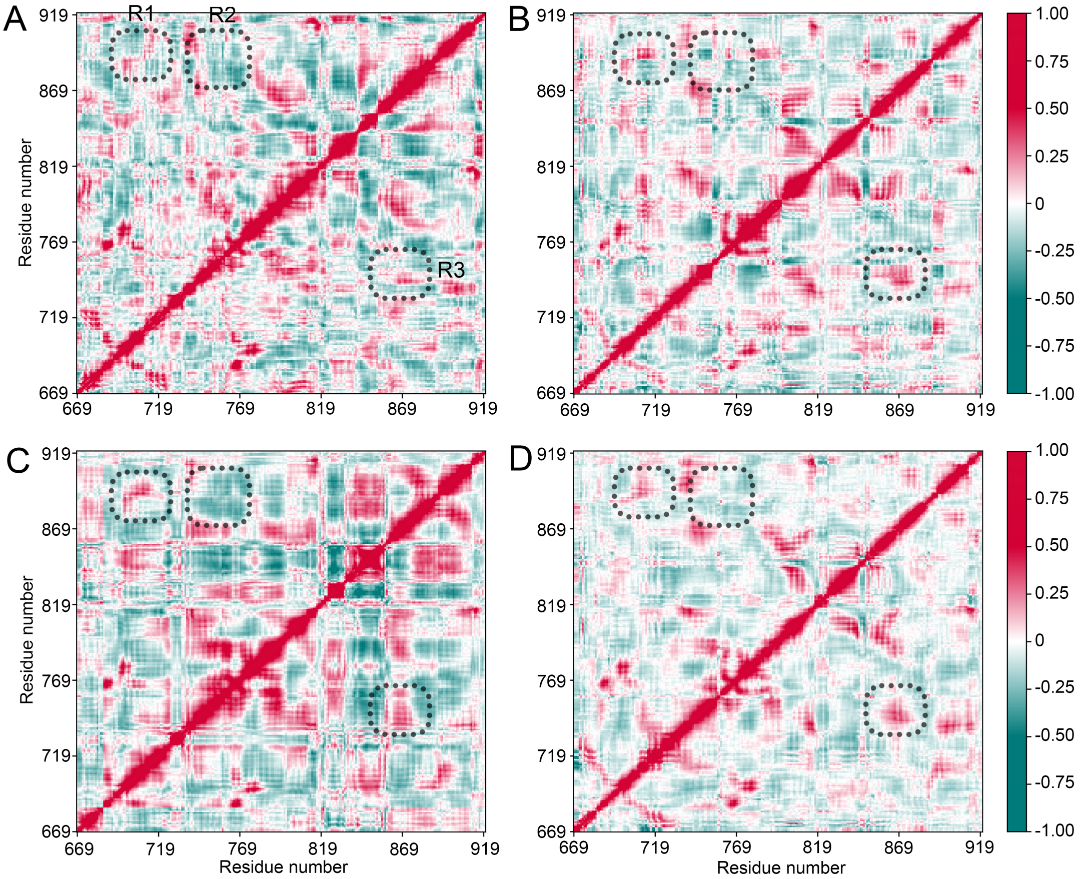


**Figure S 8** DCCM analysis of (A) K10 and DHT, (B) K10 and bicalutamide, (C) AV-6 and DHT, and (D) AV-6 and bicalutamide. Region of interest are indicated as R1 (H3 to H12), R2 (H4-H5 to H10-H12), and R3 (H3 to H10).

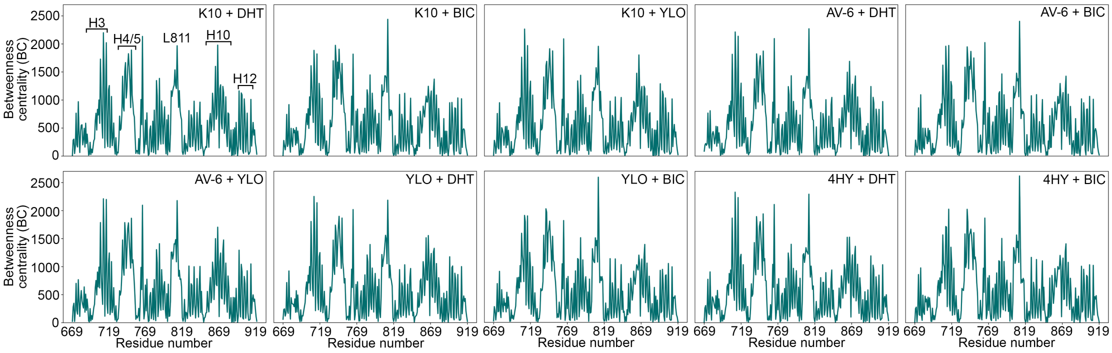


**Figure S 9** Average BC values of all simulations including the equilibration simulation with bicalutamide (BIC) bound to the AR alone. Secondary structure elements are marked in the top left subfigure.

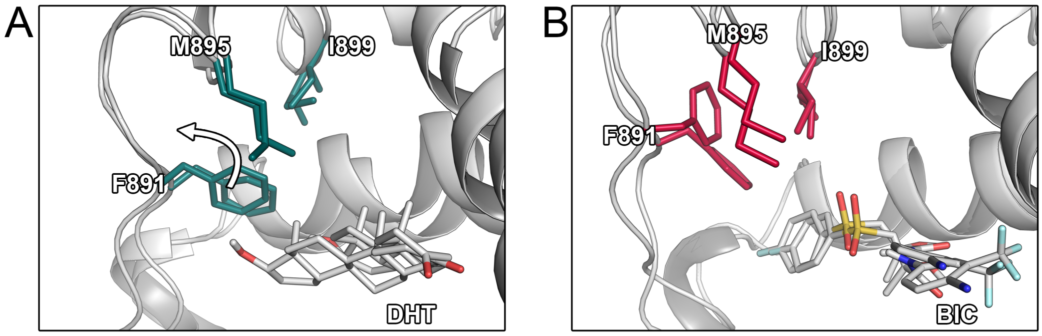


**Figure S 10** Example for conformational changes of binding site residues in (A) DHT-bound simulations (K10 with DHT, AV-6 with DHT) and (B) bicalutamide-bound simulations (K10 and bicalutamide, bicalutamide control.)

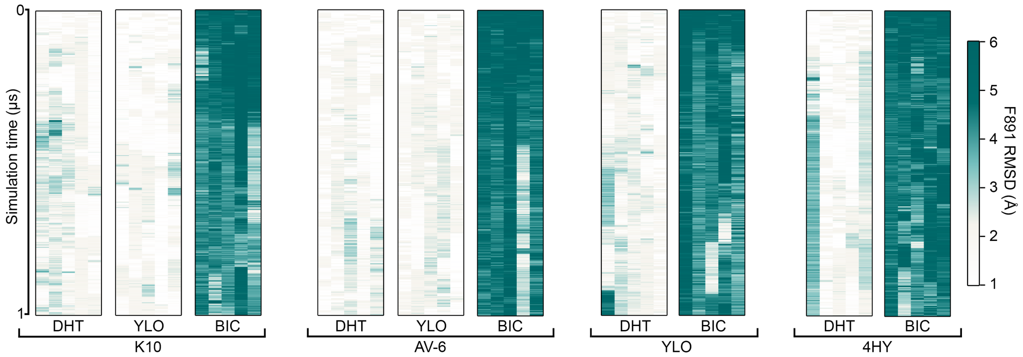


**Figure S 11** RMSD of F891 compared to crystal structure in all simulations with allosteric ligands.

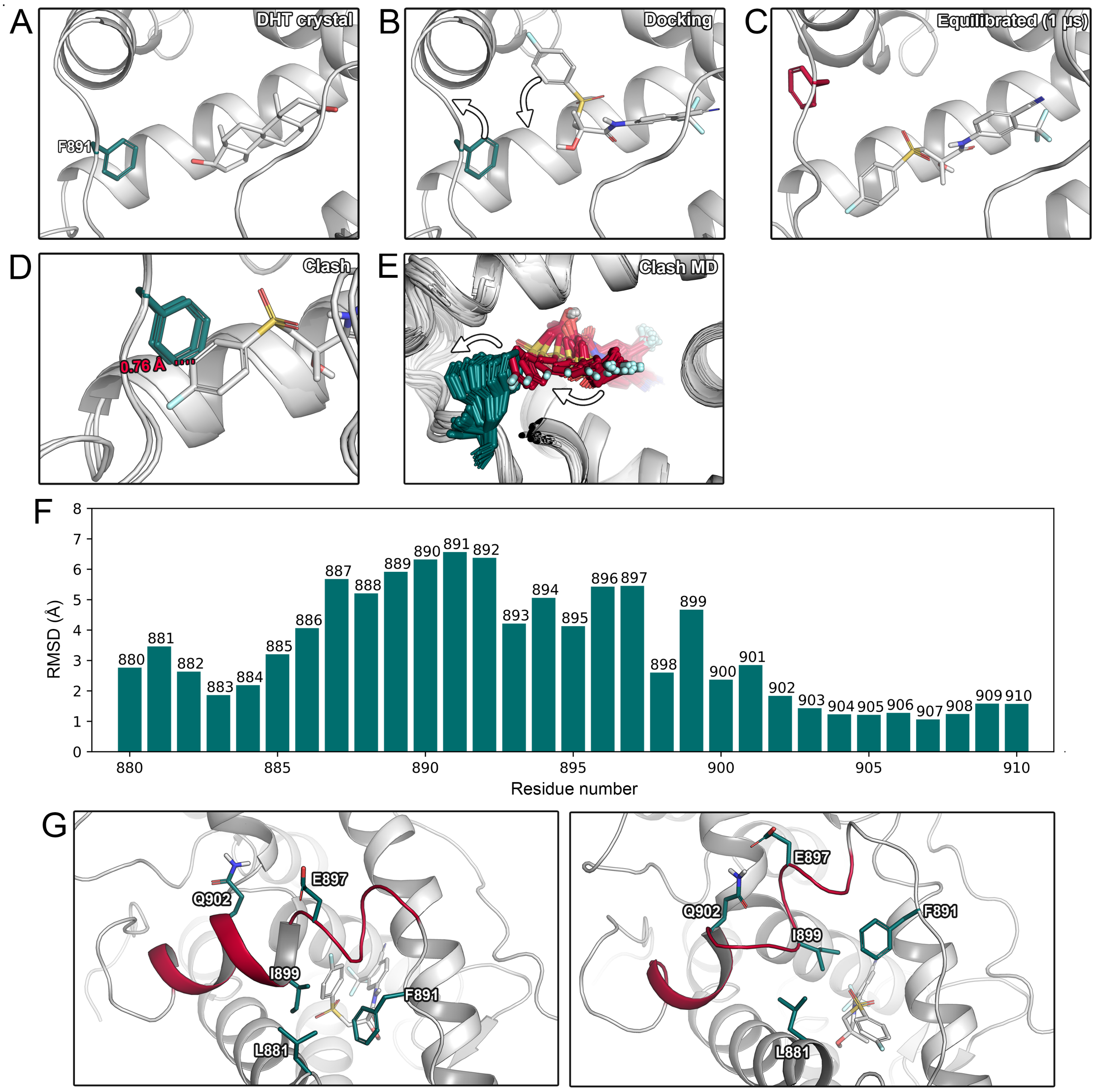


**Figure S 12** Conformational change of F891. (A) Position of F891 in native crystal structure (PDB ID: 3L3X). (B) Positioning of bicalutamide and F891 after induced-fit docking. (C) Positioning of bicalutamide and F891 after 1 $\mu$s MD simulation. (D) Close contact after superposition of MD-obtained pose of bicalutamide and F891 in its original position. (E) Concurrent movement of bicalutamide and F891 obtained from multiple consecutive MD snapshots. (F) Per-residue RMSD comparison of first and last frame of MD equilibration with bicalutamide bound. The individual residue numbers are presented above the bins. (G) Interaction network before and after conformational changes of H12.

**Table S 7** Data of F891A simulation: average values

| Orthosteric ligand | Sequence | Average H12 helicity | *n*[a] |
|---|---|---|---|
| DHT | wild-type | 0.72 ± 0.11 | 1000 |
| BIC | wild-type | 0.61 ± 0.14 | 200 |
|  | F891A | 0.67 ± 0.14 | 200 |

[a] *n* represent the number of data points considered for the average calculation and following statistical assessment.

**Table S 8** Data of F891A statistical significance and outcomes.

| System 1 | System 2 | Significance[a] | Outcome |
|---|---|---|---|
| DHT wild-type | BIC F891A | yes | BIC F891A lower |
| BIC wild-type | BIC F891A | yes | BIC F891A higher |

[a]Statistical significance was evaluated at p=0.05.

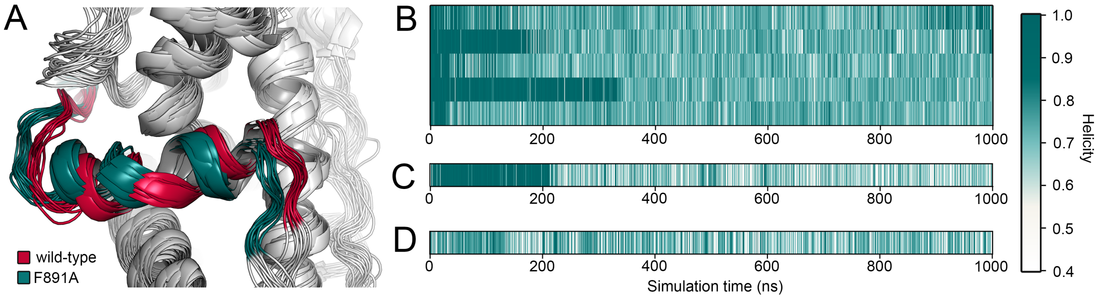


**Figure S 13** Data on the F891A mutation. (A) MD snapshots of H12 when bicalutamide was combined with wild-type and F891A receptor. H12 was colored in red and pine green respectively.

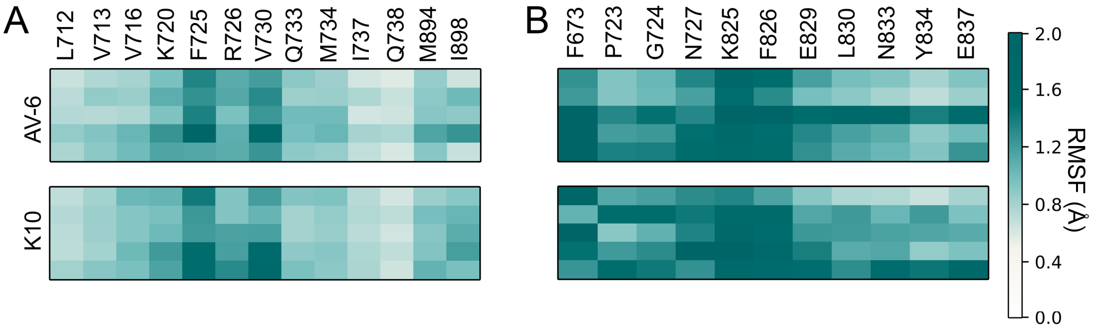


**Figure S 14** Per-residue RMSF values in simulations combining AF-2 and BF-3 inhibitors for (A) DHT-bound systems and (B) bicalutamide-bound systems.

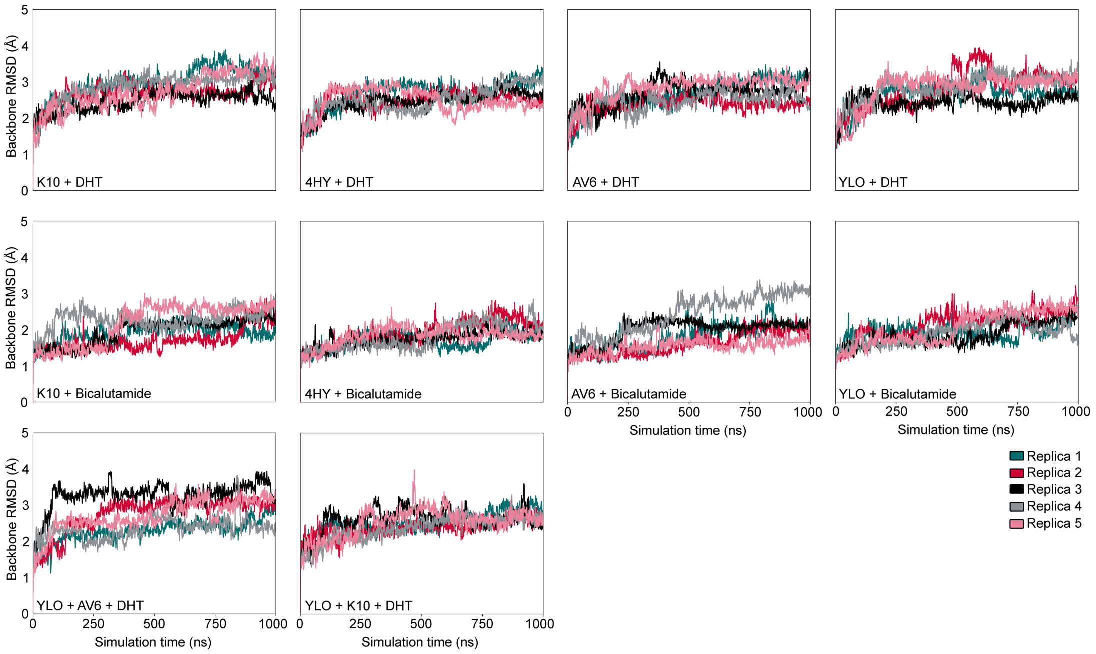


**Figure S 15** RMSD values of the simulations of different combinations.

## Supporting Materials and Methods

**Table S 9** Energy of bicalutamide isomers.

| Stereoisomer | LigPrep energy (kcal/mol) |
|---|---|
| (R)-bicalutamide | 27.9 |
| (S)-bicalutamide | 26.1 |

Lower values of the ”r_lp_energy” parameter correspond to a more favorable energy.

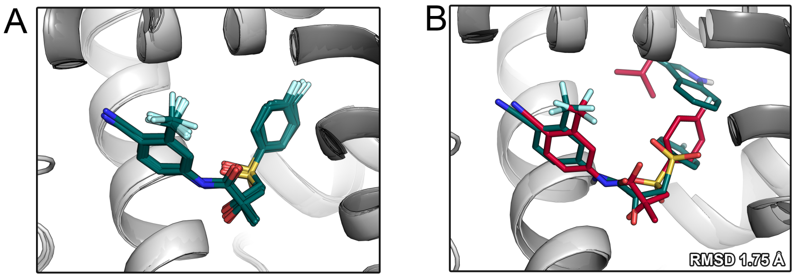


**Figure S 16** Binding mode of bicalutamide. (A) Superimposed crystal structures of AR cocrystallized with bicalutamide. (B) Comparison of bicalutamide docking pose obtained with DOLINA to corystallized structure.

**Table S 10** Origin of allosteric ligands.

| Ligand | PDB ID | Pubchem CID |
|---|---|---|
| K10 | 2PIP | 11522441 |
| AV-6 | 2YHD | 761631 |
| 4HY | 2PIT | 5803 |
| YLO | 2YLO | 3114779 |

**Table S 11** Population of clusters.

| Allosteric | Orthosteric | Replica 1 | Replica 2 | Replica 3 | Replica 4 | Replica 5 |
|---|---|---|---|---|---|---|
| K10 | DHT | 12 | 11 | 12 | 10 | 15 |
| K10 | BIC | 18 | 12 | 14 | 21 | 12 |
| K10 | YLO | 14 | 18 | 27 | 16 | 15 |
| AV-6 | DHT | 17 | 28 | 14 | 19 | 15 |
| AV-6 | BIC | 14 | 11 | 16 | 12 | 20 |
| AV-6 | YLO | 14 | 17 | 18 | 14 | 17 |
| YLO | DHT | 16 | 18 | 13 | 15 | 17 |
| YLO | BIC | 14 | 18 | 15 | 16 | 13 |
| 4HY | DHT | 11 | 15 | 19 | 13 | 12 |
| 4HY | BIC | 11 | 17 | 11 | 18 | 28 |
| none | DHT | 19 | - | - | - | - |
| none | BIC | 13 | - | - | - | - |

The population of the highest cluster is given for all systems and the respective replica simulations. A total of 101 structures were included in the clustering analysis.

**Table S 12** Number of clusters.

| **Allosteric** | **Orthosteric** | **Replica 1** | **Replica 2** | **Replica 3** | **Replica 4** | **Replica 5** |
|---|---|---|---|---|---|---|
| K10 | DHT | 15 | 14 | 13 | 15 | 11 |
| K10 | BIC | 11 | 13 | 14 | 8 | 12 |
| K10 | YLO | 14 | 14 | 11 | 12 | 11 |
| AV-6 | DHT | 15 | 11 | 9 | 8 | 12 |
| AV-6 | BIC | 13 | 15 | 12 | 12 | 13 |
| AV-6 | YLO | 13 | 13 | 12 | 14 | 13 |
| YLO | DHT | 14 | 12 | 13 | 13 | 11 |
| YLO | BIC | 13 | 13 | 12 | 12 | 13 |
| 4HY | DHT | 14 | 10 | 12 | 13 | 12 |
| 4HY | BIC | 15 | 12 | 15 | 10 | 6 |
| none | DHT | 15 | - | - | - | - |
| none | BIC | 13 | - | - | - | - |

The total number of clusters is given for all systems and the respective replica simulations.

**Table S 13** Residues flanking H3, H10, and H12.

| **Helix** | **Start Residue** | **End Residue** |
|---|---|---|
| H3 | F697 | K720 |
| H10 | S851 | I882 |
| H12 | M894 | I906 |

Residue boundaries used for the determination of helicity.

# List of Publications

1. Fischer, A.; Smieško, M. Spontaneous Ligand Access Events to Membrane-Bound Cytochrome P450 2D6 Sampled at Atomic Resolution. *Sci. Rep.* **2019**, 9, 16411.
2. Fischer, A.; Smieško, M. Ligand Pathways in Nuclear Receptors. J. Chem. Inf. Model. **2019**, 59 (7), 3100–3109.
3. Fischer, A.; Smieško, M. Allosteric Binding Sites on Nuclear Receptors: Focus on Drug Efficacy and Selectivity. *Int. J. Mol. Sci.* **2020**, 21, 6–8.
4. Fischer, A..; Sellner, M.; Neranjan, S.; Smieško, M.; Lill, M. A. Potential Inhibitors for Novel Coronavirus Protease Identified by Virtual Screening of 606 Million Compounds. *Int. J. Mol. Sci.* **2020**, 21 (10), 1–17.
5. Fischer, A.; Frehner, G.; Lill, M. A.; Smieško, M. Conformational Changes of Thyroid Receptors in Response to Antagonists. *J. Chem. Inf. Model.* **2021**, 61 (2), 1010-1019 .
6. Sellner, M.; Fischer, A.; Don, C. G.; Smieško, M. Conformational Landscape of Cytochrome P450 Reductase Interactions. *Int. J. Mol. Sci.* **2021**, 22 (3), 1–14.

7. Fischer, A.; Häuptli, F.; Lill, M. A.; Smieško, M. Computational Assessment of Combination Therapy of Androgen Receptor-Targeting Compounds. *J. Chem. Inf. Model.* **2021**, 61, 2, 1001–1009.
8. Fischer, A.; Sellner, M.; Mitusińska, K.; Bzówka, M.; Lill, M. A.; Góra, A.; Smieško, M. Computational Selectivity Assessment of Protease Inhibitors against Sars-Cov-2. *Int. J. Mol. Sci.* **2021**, 22 (4), 1–17.
9. Fischer, A.; Smieško, M.; Sellner, M.; Lill, M. A. Decision Making in Structure-Based Drug Discovery: Visual Inspection of Docking Results. *J. Med. Chem.* **2021**, 64 (5), 2489–2500.

# Curriculum Vitae

André Fischer was born on the 17th of March 1994 in Niedergösgen, Switzerland. After finishing his high school degree focused on natural sciences with a major in chemistry and biology at the Kantonsschule Olten in 2012, he enrolled in the bachelor degree course of Pharmaceutical Sciences at the University of Basel. After finishing this course, he started his master studies in Drug Sciences at the same university. During his final year as a student, he joined the Molecular Modeling group at the University of Basel headed by PD Dr. Martin Smieško as a master intern. He was awarded the GSIA price for the best master thesis of the Drug Sciences graduates in his year. Fascinated by the field of computational chemistry, he was enabled to consecutively start his Ph.D studies in the group hat was later headed by Prof. Dr. Markus A. Lill. His work includes contributions in the field of computational chemistry in regards to virtual screening, compound selectivity, and ligand-induced conformational changes with a focus on nuclear receptors, drug-metabolizing enzymes, and proteases. The ten publications he coauthored, nine of them as lead author, already yielded over 170 citations to the date of handing in his dissertation. Following his stay at the University, he continued his career as cheminformatician at DSM Nutritional Products in Kaiseraugst.