MOLECULAR DYNAMICS AND VIRTUAL SCREENING APPROACHES IN DRUG DISCOVERY

Elmeri Jokinen

TURUN YLIOPISTO
UNIVERSITY OF TURKU

MOLECULAR DYNAMICS AND VIRTUAL SCREENING APPROACHES IN DRUG DISCOVERY

Elmeri Jokinen

University of Turku

Faculty of Medicine
Institute of Biomedicine
Pharmacology, Drug Development and Therapeutics
Drug Research Doctoral Programme

Supervised by

Professor, Olli Pentikäinen, PhD
Institute of Biomedicine
University of Turku
Turku, Finland

Reviewed by

Professor, Philip Biggin, PhD
CChem, FRSC
Department of Biochemistry
University of Oxford
Oxford, United Kingdom

Professor, Gerrit Groenhof, PhD
Nanoscience Center and
Department of Chemistry
University of Jyväskylä
Jyväskylä, Finland

Opponent

Professor, Stefano Moro, PhD
Molecular Modeling Section
Department of Pharmaceutical and
Pharmacological Sciences
University of Padova
Padova, Italy

"It was not easy…"

An unknown scientist from the old days

UNIVERSITY OF TURKU
Faculty of Medicine
Institute of Biomedicine
Pharmacology, Drug Development and Therapeutics
ELMERI JOKINEN: Molecular dynamics and virtual screening approaches
in drug discovery
Doctoral Dissertation, 127 pp.
Drug Research Doctoral Programme (DRDP)
October 2022

ABSTRACT

Computer-aided drug discovery (CADD) methods are now routinely used in the preclinical phase of drug development. Powerful high-performance computing facilities and the extremely fast CADD methods constantly scale up the coverage of drug-like chemical space achievable in rational drug development. In this thesis, CADD approaches were applied to address several early-phase drug discovery problems. Namely, small molecule binding site detection on a novel target protein, virtual screening (VS) of molecular databases, and characterization of small molecule interactions with metabolic enzymes were studied. Various CADD methods, including molecular dynamics (MD) simulations in mixed solvents, molecular docking, and binding free energy calculations, were employed. Co-solvent MD simulations detected biologically relevant binding sites and provided guidance for screening potential protein-protein interaction inhibitors for a crucial protein of the SARS-CoV-2. VS with fragment- and negative image-based (F-NIB) models identified three active and structurally novel inhibitors of the putative drug target phosphodiesterase 10A. MD simulations and docking provided detailed insights on the effects of active site structural flexibility and variation on the binding and resultant metabolism of small molecules with the cytochrome P450 enzymes. The results presented in this thesis contribute to the increasing evidence that supports employment and further development of CADD approaches in drug discovery. Ultimately, rational drug development coupled with CADD may enable higher quality drug candidates to the human studies in the future, reducing the risk of financially and temporally costly clinical failure.

KEYWORDS: Structure-based drug development, Computer-aided drug discovery (CADD), Molecular dynamics (MD) simulation, Virtual screening (VS), Fragment- and negative image-based (F-NIB) model, Structure-activity relationship (QSAR), Cytochrome P450 ligand binding prediction

TIIVISTELMÄ

Tietokoneavusteista lääkeaine-etsintää käytetään nykyisin yleisesti prekliinisessä lääketutkimuksessa. Suurteholaskenta ja äärimmäisen nopeat tietokoneavusteiset lääkeaine-etsintämenetelmät mahdollistavat jatkuvasti kattavamman lääkkeenkaltaisten molekyylien kemiallisen avaruuden seulonnan. Tässä väitöskirjassa tietokonepohjaisia menetelmiä hyödynnettiin lääketutkimuksen prekliiniseen vaiheeseen liittyvissä tyypillisissä tutkimusongelmissa. Työhön kuului pienmolekyylien sitoutumisalueiden tunnistus uuden kohdeproteiinin rakenteesta, molekyylitietokantojen virtuaaliseulonta sekä pienmolekyylien ja metabolian entsyymien välisten vuorovaikutusten tietokonemallinnus. Työssä käytettiin useita tietokoneavusteisen lääkeaine-etsinnän menetelmiä, sisältäen molekyylidynamiikkasimulaatiot (MD-simulaatiot) vaihtuvissa liuottimissa, molekulaarisen telakoinnin ja sitoutumisenergian laskennan. Orgaanisen liuottimen ja veden sekoituksessa tehdyt MD-simulaatiot tunnistivat biologisesti merkittäviä sitoutumisalueita SARS-CoV-2:n tärkeästä proteiinista ja ohjasivat infektioon liittyvän proteiini-proteiinivuorovaikutuksen potentiaalisten estäjien etsintää. Virtuaaliseulonnalla tunnistettiin kolme rakenteellisesti uudenlaista tunnetun lääkekehityskohteen, fosfodiesteraasi 10A:n, estäjää hyödyntäen fragmentti- ja negatiivikuvamalleja. MD-simulaatiot ja telakointi tuottivat yksityiskohtaista tietoa sytokromi P450 entsyymien aktiivisen kohdan rakenteen jouston ja muutosten vaikutuksesta pienmolekyylien sitoutumiseen ja metaboliaan. Tämän väitöskirjan tulokset tukevat kasvavaa todistusaineistoa tietokoneavusteisen lääkeaine-etsinnän käytön ja kehityksen hyödyllisyydestä prekliinisessä lääketutkimuksessa. Tietokoneavusteinen lääkeaine-etsintä voi lopulta mahdollistaa korkeampilaatuisten lääkekandidaattien päätymisen ihmiskokeisiin, pienentäen taloudellisesti ja ajallisesti kalliin kliinisen tutkimuksen epäonnistumisen riskiä.

AVAINSANAT: Rakennepohjainen lääkeainekehitys, Tietokoneavusteinen lääkeaine-etsintä, Molekyylidynamiikkasimulaatio (MD-simulaatio), Virtuaaliseulonta, Fragmentti- ja negatiivikuvamalli, Rakenne-aktiivisuussuhde, Sytokromi P450 ligandien sitoutumisen ennustus

Table of Contents

Abbreviations

ACE2	Angiotensin-converting enzyme 2
aMD	Accelerated molecular dynamics
ANF	alpha-naphthoflavone
CADD	Computer-aided drug discovery
CoMFA	Comparative molecular field analysis
CoMSIA	Comparative molecular similarity indices analysis
CYP	Cytochrome P450
DCPCC	N-(3,5-dichlorophenyl)cyclopropanecarboxamide
FBDD	Fragment-based drug discovery
FEP	Free energy perturbation
FQSAR	Field-based quantitative structure-activity relationship
F-NIB	Fragment- and negative image-based
LA	Linoleic acid
MD	Molecular dynamics
MixMD	Mixed solvent molecular dynamics
MM/GBSA	Molecular mechanics/Generalized born surface area
NIB	Negative image-based
PDB	Protein Data Bank
PDE10A	Phosphodiesterase 10A
PLS	Partial least squares
PPI	Protein-protein interaction
PROTAC	Proteolysis targeting chimera
RBD	Receptor-binding domain
RMSD	Root mean square deviation
SARS-CoV-2	Severe acute respiratory syndrome coronavirus 2
SBDD	Structure-based drug development
SOM	Site of metabolism
S protein	Spike protein
VS	Virtual screening

List of Original Publications

This dissertation is based on the following original publications, which are referred to in the text by their Roman numerals:

I   Jokinen E.M., Postila P.A., Ahinko M., Niinivehmas S.P., Pentikäinen O.T. Fragment- and Negative Image-Based Screening of Phosphodiesterase 10A Inhibitors. *Chemical Biology & Drug Design*, 2019; 94(4):1799-1812: https://doi.org/10.1111/cbdd.13584.

II   Juvonen R.O., Jokinen E.M., Javaid A., Lehtonen M., Raunio H., Pentikäinen O.T. Inhibition of Human CYP1 Enzymes by a Classical Inhibitor α-Naphthoflavone and Novel Inhibitor N-(3, 5-Dichlorophenyl)Cyclopropanecarboxamide: An in Vitro and in Silico Study. *Chemical Biology & Drug Design*, 2020; 95(5):520–33: https://doi.org/10.1111/cbdd.13669.

III   Juvonen R.O., Jokinen E.M., Huuskonen J., Kärkkäinen O., Raunio H., Pentikäinen O.T. Molecular docking and oxidation kinetics of 3-phenyl coumarin derivatives by human CYP2A13. *Xenobiotica,* 2021; 51(11):1207-1216: https://doi.org/10.1080/00498254.2021.1898700.

IV   Jokinen E.M., Krishnasamy G., Kurkinen S.T., Pentikäinen O.T. Detection of binding sites on SARS-CoV-2 Spike protein receptor binding domain by molecular dynamics simulations in mixed solvents. *IEEE/ACM Transactions in Computational Biology and Bioinformatics*, 2021; 18(4):1281-1289: https://doi.org/10.1109/TCBB.2021.3076259.

The original publications have been reproduced with the permission of the copyright holders. Contribution of the thesis author into each article is described in a separate appendix.

1   Introduction

The proteome of an organism constitutes a multi-tool for maintaining biological systems that can be a single cell or a highly complex multi-tissue life form. Proteins mediate practically all biological processes within living matter by interactions with ligands, such as other proteins, DNA, or endo- or exogenous small molecules. Therefore, it is not surprising that diseases very often relate to the malfunction of specific proteins. Drug discovery and development aim to find substances that can be used to modulate the function of such proteins or counteract their effects indirectly. The development of novel drugs is, however, extremely laborious and expensive. The financial and time investments required for developing a new drug are currently estimated at over 2 billion USD and more than ten years, respectively (DiMasi et al., 2016). Moreover, a drug development project always includes significant risk of failure. The common root cause for clinical failure is that predicting the behavior of novel substances in the human body is still extremely challenging (Sun et al., 2022). On the other hand, many promising drug development campaigns are halted right at the start due to the target protein, despite being medicinally interesting, seeming technically too difficult to be targeted with drugs.

Initiating a drug development project with a novel target protein requires solving the following problems: where into the protein's structure should the drug bind, and what kind of compound will an effective drug be? The traditional approach to address these issues is to test large quantities of molecules experimentally by high-throughput screening and see if some of them modulate the function of the protein in a pharmacologically desirable manner (Szymański et al., 2012). However, to find compounds that affect a specific protein, segments of the astronomically sized chemical space of small-sized organic molecules must be sieved through exhaustively. Despite having shown many accomplishments, highly automated high-throughput screening can test only a tiny fraction of the possible drug-like compounds. Limited sampling of the relevant chemical space may lead to the best candidates remaining hidden, while those who enter the clinical trials may be doomed to fail if suitable candidates are found at all.

Structure-based drug development (SBDD) aims to make drug discovery more efficient by considering the structure of the target protein in the search for active and

selective compounds (Huggins et al., 2012; Lounnas et al., 2013). SBDD is strongly coupled with computer-aided drug discovery (CADD) that can be used to simulate protein-drug interactions in great detail and with unmatched speed (Sliwoski et al., 2014). Rapid and continuous development of the CADD methods constantly brings novel and improved drug discovery strategies available for investigators. Already today, CADD tools exist for detecting novel binding sites on target proteins, extremely fast evaluation of compound affinity to a drug target, and prediction of drug candidate metabolism. Employment and development of better CADD methods may bring forward novel drug targets and help improve the quality of the compounds that enter the clinical trials. In this thesis, the utility of current CADD workflows is demonstrated by focusing on binding site identification with a novel target protein, virtual screening (VS) with a novel approach, and prediction of small molecule binding to metabolic enzymes.

2 Review of the Literature

2.1 Structure-based drug discovery

SBDD relies on the availability of high-resolution structural models of the target protein, obtained, for example, by X-ray crystallography. The structure is then used as a template where molecules in enormously sized molecular databases are fit to evaluate their shape and electrostatic complementarity with the binding site and estimate their suitability for binding to the target (Huggins et al., 2012; Lounnas et al., 2013; Figure 1). Evaluating up to billions of compounds within just days or weeks is enabled by the extremely fast CADD approaches (Gorgulla et al., 2020). While the ability of the CADD methods to provide predictions within the chemical accuracy still cannot be guaranteed, they can be used to prioritize compounds or compound classes for experimental studies. In addition, VS has been argued to increase the chemical diversity of the obtained active hit compounds due to its ability to cover a more significant portion of the drug-like chemical space than what is possible for any experimental high-throughput screening protocol (Shoichet, 2004). Having multiple active structural scaffolds could be paramount for developing compounds with a chance of succeeding in the clinical trials by fulfilling the strict pharmacodynamics and –kinetics criteria of potent and safe drugs (Sun et al., 2022). SBDD and CADD approaches have been pivotal in developing many of the drugs currently on the market (Sabe et al., 2021).
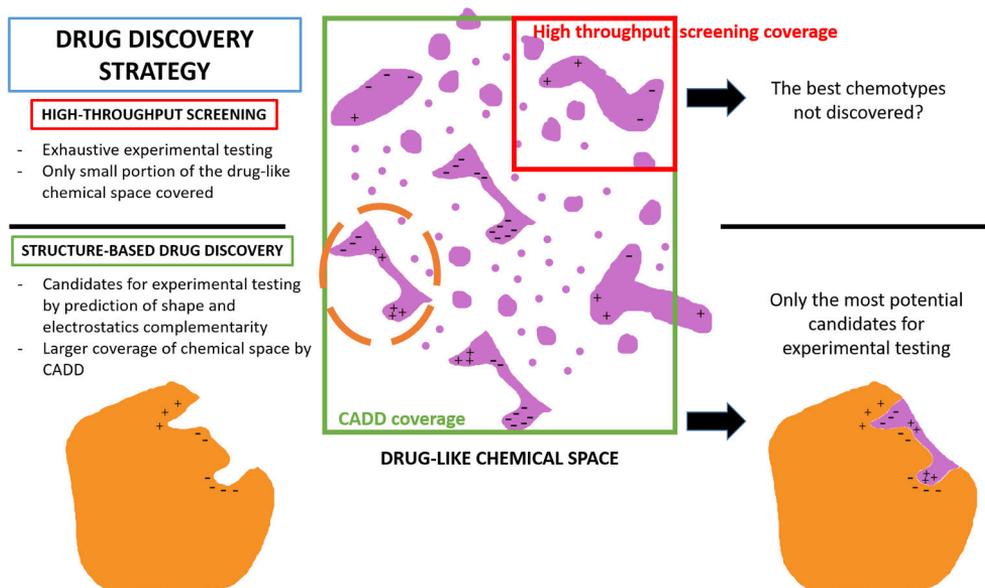
Figure 1. The rationale behind SBDD. In the illustrative SBDD flow, the target protein structure (orange) is used to find the compounds that could most potentially bind to the designated binding site. A potential hit compound is highlighted with an orange dashed circle. While experimental confirmation for the compounds predicted active by CADD is needed, the testing can be directed only to the most potential candidates.

Various *in silico* approaches exist that offer tools for the preclinical phase of the drug development process. Bioinformatics tools are generally used to identify putative drug targets (Jiang and Zhou, 2005). In contrast, simulations of varying chemical detail can be utilized to detect target binding sites for drugs and the drug screening itself. Optimization of binding affinity, selectivity, and pharmacokinetics is also supported by numerous CADD-based predictive approaches (Heifetz et al., 2016; Sliwoski et al., 2014). Computer hardware and software are constantly evolving, and high-performance computing facilities and large compound databases are becoming more readily available for CADD investigators (Puertas-Martín et al., 2020; Sterling and Irwin, 2015). Due to this development, the application of SBDD and CADD methods for solving problems that are more complex than before is drawing attention in both academia and industry.

Inadequate consideration of the structural flexibility of the target protein has been one traditional limitation of the CADD methods (Alvarez-Garcia and Barril, 2014). It is well known that many proteins undergo structural adaptation upon ligand recognition (Buonfiglio et al., 2015; Mobley and Dill, 2009). Such ligand-bound conformations may be non-visible in the protein's non-liganded (*apo*) form. Especially, cryptic binding sites that require ligand occupation to be observable in an open state may be extremely difficult to detect with experimental methods

(Bowman and Geissler, 2012; Johnson and Karanicolas, 2015). Obtaining information on conformational changes that facilitate small molecule binding could bring up novel strategies for SBDD. Specifically, many disease-related signaling pathways are mediated by protein-protein interactions (PPIs), making them a promising drug discovery and development target. PPI interfaces are also infamously difficult targets for developing selective small molecule drugs. Protein-protein contacts are generally made through large, flat, and hydrophobic interfaces that lack suitable binding sites for small molecules (Wells and McClendon, 2007). Many examples have emerged where a PPI interface accepts a small molecule binder through structural adaptation, inhibiting a specific PPI (Scott et al., 2016). Several drugs targeting PPIs have already entered the market, and many more are being evaluated in clinical trials, demonstrating the feasibility of the approach (Arkin et al., 2014; Lu et al., 2020). Modern CADD methods offer valuable tools for simulating the flexibility of such technically challenging target proteins, possibly revealing conformations where drug discovery could be targeted (Scott et al., 2016; Shan et al., 2022). Computer simulations may bring forward novel SBDD approaches considering also known drug targets. Structural rearrangements at known binding sites may reveal additional sub-pockets or conformations that could be utilized by novel ligand chemotypes with possibly improved potency and drug-like properties (Amaral et al., 2017; Shiau et al., 1998).

Fragment-based drug discovery (FBDD) has become an essential strategy within the SBDD field. A significant benefit of screening molecular fragments (MW < 300 Da) is that much smaller compound library can sample much broader chemical diversity than with the small molecules (Scott et al., 2012). Due to low affinity, binding fragments usually need to utilize efficiently the available binding interactions. Fragment binding may thus recognize interaction sites that significantly contribute to the binding energy. Ideally, active fragments binding to different high-affinity regions of the binding are fused into a tight-binding small molecule. CADD methods are having a growing impact on FBDD (Bian and Xie, 2018; Bissaro et al., 2020; Chen and Shoichet, 2009; Xiong et al., 2016), especially interesting being the recent attempts to derive fragment dissociation constants from molecular dynamics (MD) simulations (Pan et al., 2017). An experimental FBDD approach was recently combined with MD simulations to detect a cryptic binding site for Heat Shock Protein 70 (O'Connor et al., 2022). Inspection of aromatic residue fluctuations of cryptic sites in MD simulations provided structural insight for FBDD (Iida et al., 2020). The following sections present current CADD approaches utilized in SBDD for druggable binding site identification and small molecule affinity, binding mode and metabolism prediction.

2.2   Computer-aided target site identification

Especially with emerging drug targets, locations of the possible orthosteric or allosteric target sites or their ligand-bound conformations are often unknown. Even with established target sites, improved consideration of the site's flexibility may enable the identification of active compounds that contain novel structural scaffolds. The first computational approaches addressing this issue were developed already in the 1980s. The GRID program placed a protein inside a grid and probed the protein surface with various functional groups, labelling grid points with favourable energetics as potential binding hotspots (Goodford, 1985). Since then, various binding site detection algorithms have evolved that analyze protein surfaces using a single static structure, enabling rapid detection of potential binding sites from an experimental model. FPOCKET is a geometry-based method where suitable sites are located by detecting curvature or cavities within the protein structure by placing and clustering alpha spheres on the protein surfaces (Le Guilloux et al., 2009). Further calculations considering the pocket volume and cavity-lining amino acids allow scoring of the detected pockets according to their suitability for small molecule binding. Other approaches include probing the protein surface with chemical moieties, similarly to GRID, that generates energy potentials by forming non-covalent interactions on putative binding sites, or similarity-based prediction using known binding sites of structurally homologous proteins (Ngan et al., 2012; Wass et al., 2010). Moreover, machine learning is increasingly utilized in binding site detection (Jiménez et al., 2017; Stepniewska-Dziubinska et al., 2020). Recently, the performance of geometry-, energy- and machine-learning-based binding site detection algorithms was investigated using *apo* and *holo* protein structures (Clark et al., 2020).

As druggable binding pockets are not always present in the available experimental models, the potential opening of such pockets can be investigated using MD simulations. MD simulation is a tool for modelling the structural dynamics of molecules at the atomic level. The method can be used to explore the time-dependent conformational fluctuations and solvent interactions of a protein in water. Motion of all atoms in the system is integrated within femtosecond time steps. Forces that affect each atom and determine their motion are mathematically approximated using a potential energy function based on classical molecular mechanics. Commonly used force field families for proteins include AMBER, CHARMM, and OPLS (Best et al., 2012; Maier et al., 2015; Robertson et al., 2015). Continuous progress in computer hardware and simulation algorithms has made it possible to simulate bio molecular systems consisting of millions of atoms in reasonable time scales. Using specialized computer hardware, processes up to millisecond scale, for example, protein folding, are accessible for MD simulations (Shaw et al., 2009). Generally, medicinally interesting nanosecond to microsecond scale events, such as ligand binding, plasma membrane effect to ligand binding, force propagation along proteins, and protein-

protein interactions, can be routinely simulated in atomic detail (Heifetz et al., 2016; Lolicato et al., 2020; Postila et al., 2010; Seppälä et al., 2017; Shan et al., 2022; Sliwoski et al., 2014; Veeramachaneni et al., 2021). This development has made MD simulation an extraordinarily versatile and exciting tool for basic research, discovery, and optimization related to drug development (Durrant and McCammon, 2011). In the following two sections, an increasingly popular technique for binding site detection, the co-solvent MD simulation, is presented.

2.2.1   Binding pocket detection by co-solvent MD simulations

Co-solvent MD simulation is a method for detecting high-affinity interaction sites on a protein structure (Defelipe et al., 2018; Ghanakota and Carlson, 2016a). The basic idea in co-solvent MD is to probe the surface of the simulated protein with small-sized organic molecules that express drug-like chemical properties (**Figure 2**). The detected interaction sites may contain novel regions where drug discovery efforts could be targeted. In addition, the organic probe molecules may induce the opening of occluded or cryptic binding pockets that open only upon ligand recognition.
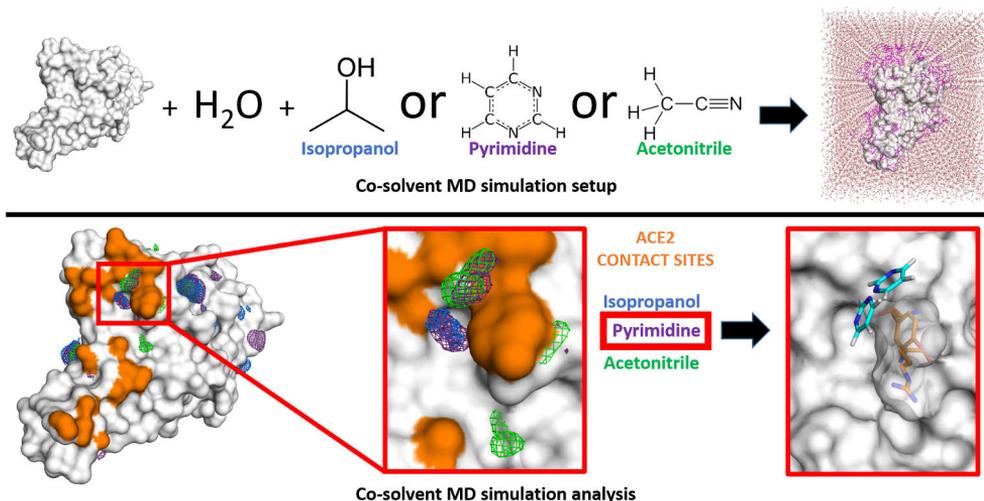
Figure 2. Overview of a co-solvent MD simulation. In co-solvent MD, the protein is solvated into a mixture of water and an organic probe molecule (upper row). In the analysis phase, the sites of high probe occupancy on the protein surface are detected and visualized as volume density (lower row, left, and middle). Here, probe occupancy was detected on a PPI site (lower row, middle; protein surface is colored white, while protein-protein contact areas are colored orange). The most common binding modes of the probe molecules on specific binding hotspots can be extracted from the simulation trajectories (lower row, right). The figure is arranged according to the co-solvent MD protocol described in Ghanakota and Carlson, 2016b.

The experimental background of the method is in the multiple solvent crystal structure studies where protein crystal structures were solved with various organic co-solvents (Allen et al., 1996). The Barril group presented the first co-solvent MD simulation technique where the simulations were performed in 20% isopropanol-water solvent, and localization of the probe was compared to multiple solvent crystal structures with three target proteins (Seco et al., 2009). The probe occupancies were analyzed by a grid-based approach where binding free energy change associated with a probe binding to a specific spot was evaluated. Ligand competitive saturation, SILCS, is another early co-solvent MD method characterized by the use of artificial repulsion between the hydrophobic probes to avoid aggregation due to high probe concentrations (Guvench and MacKerell, 2009). SILCS has found applications related to pharmacophore modelling, free energy calculations, and sampling of buried binding pockets (Lakkaraju et al., 2015; Raman et al., 2012; Yu et al., 2015). The mixed solvent molecular dynamics (MixMD) protocol, developed by the Carlson group, was applied in this thesis. In the current MixMD, the simulations are performed with binary solvents expressing 5% probe-to-water ratio and fully flexible target protein (Lexa and Carlson, 2011; Ung et al., 2016). The commonly used probes include pyrimidine, acetonitrile and isopropanol. Usage of charged probes has also been demonstrated (Ghanakota and Carlson, 2016b). MixMD and its related parameters have been calibrated and benchmarked using various target proteins (Lexa et al., 2014). Generally, the method has been shown to successfully identify orthosteric and allosteric binding sites and PPI sites in both retrospective and prospective studies (DasGupta et al., 2022; Ghanakota and Carlson, 2016b; Ghanakota et al., 2018; Makley et al., 2021; Smith and Carlson, 2021). In addition to the co-solvent MD approaches discussed above, many others do exist and they have been comprehensively reviewed recently (Ghanakota and Carlson, 2016a).

Benchmarking studies have shown that the co-solvent MD simulations typically map additional sites on the protein structure that do not overlap with the known binding sites. While some of such sites could be relevant for drug discovery through allosteric effects, prospective usage of the co-solvent MD methods requires metrics for ranking and prioritization of the sites based on their druggability. In the MixMD method, a true binding hotspot is defined as a site with strong occupancy by at least two different probe types (Ghanakota and Carlson, 2016b). Several studies have employed pocket detection algorithms, such as Sitemap or PocketAnalyzer, to assess the druggability of the sites identified and sampled by the co-solvent MD methods (Bakan et al., 2012; Craig et al., 2011; Halgren, 2007, 2009; Kalenkiewicz et al., 2015; Schmidt et al., 2019; Yang, 2015). Usage of probe mixtures and conversion of probe occupancies to maximal binding free energy values at specific hotspots provided affinities that corresponded to the best-known ligands for those binding sites with several targets (Bakan et al., 2012). Improved docking performance with

co-crystal ligands was reported using interaction site information obtained from co-solvent MD simulations using ethanol as a probe (Arcon et al., 2017).

2.2.2   Conformational sampling of binding pockets by co-solvent MD

A crucial feature of co-solvent MD is to enrich simulation structures where the possible occluded and cryptic binding pockets are in an open state due to probe occupation. Such a phenomenon has been reported in multiple co-solvent MD studies (Bakan et al., 2012; Kalenkiewicz et al., 2015; Kimura et al., 2017; Schmidt et al., 2019). The effect has been attributed to either induced fit or conformational selection effects during the production simulation or to probes occupying specific pockets that open during an annealing stage of the equilibration protocol. MD simulations in pure water have also been reported to sample *holo*-like binding pocket states, but very infrequently or only partially (Gao et al., 2017; Prakash et al., 2015). Kimura et al. demonstrated complete or partial opening and mapping of cryptic pockets during co-solvent MD simulations for multiple targets when starting the simulations from their *apo* structures (Kimura et al., 2017). Cryptic pockets opened readily when they were associated with minor structural changes such as sidechain movements. In contrast, pockets related to more significant changes, such as helix displacement, were more often only partially mapped. The co-solvent MD simulations were shown to sample pocket states structurally close to experimentally determined ligand-bound structures. Schmidt et al. presented a highly automatized protocol for detecting cryptic pockets that become accessible during co-solvent MD simulations (Schmidt et al., 2019). Co-solvent MD induced cryptic pocket conformations that resembled their experimentally determined ligand-bound forms with several targets, although the simulations were started from the *apo* forms. Crystal structure poses of known ligands were reproduced by docking to the simulation structures, indicating the feasibility of the extracted structures in VS studies.

Integrating enhanced sampling techniques with co-solvent simulations has become an intriguing option to facilitate the opening of binding sites that depend on large backbone movements (Oleinikovas et al., 2016). Accelerated MD (aMD) has been combined with co-solvent MD using different probe types to investigate the opening of several allosteric or occluded binding sites and provide druggable conformations for docking (Kalenkiewicz et al., 2015; Smith and Carlson, 2021; Tze-Yang Ng and Tan, 2022; Yang, 2015). In aMD, a boost potential is added to the system's potential energy if the potential energy is under a threshold value. The approach enhances the system's ability to overcome energy barriers, allowing efficient sampling of conformations separated by such barriers (Hamelberg et al.,

2004). Comparison of pure water aMD simulations showed less frequent opening of small molecule binding pockets on the PPI interface (Kalenkiewicz et al., 2015). Systems where cryptic pocket opening required helix displacement were challenging for an aMD-MixMD approach, as *holo*-like conformation was not observed during 100 ns standard or aMD-MixMD simulations (Smith and Carlson, 2021). Running longer simulations was speculated to improve the situation as helix displacement was sometimes observed in individual simulation trajectories. Interestingly, molecular docking of known ligands to the cryptic sites showed similar performance when using structures from water or co-solvent simulations. This result was somewhat inconsistent with earlier work where co-solvent simulations had been reported to improve the generation of *holo*-like structures and docking compared to classical MD (Kimura et al., 2017; Schmidt et al., 2019). Sampling water interfaces through scaled Hamiltonians (SWISH) showed promising performance when combined with co-solvent MD (Comitani and Gervasio, 2018). In SWISH, nonpolar water-protein interactions are scaled so water gets a higher affinity for the nonpolar protein atoms. As a result, occluded or cryptic pockets open faster and are more easily mapped by the probe molecules. SWISH showed excellent performance in the detection of cryptic pockets and also allowed the binding of a known cryptic site ligand to a crystal structure-like conformation during a simulation.

2.3 Computational prediction of small molecule activity and binding mode

One of the central goals of SBDD is to predict molecular interactions and their resultant affinity between small molecules and proteins with high precision. The swift VS workflows enable the efficient identification of potentially active compounds from large molecular libraries (Figure 3). Consequently, knowing the exact binding mode of the active hit molecules is crucial for subsequent structure-based optimization to obtain even higher binding affinity and selectivity. In addition, binding mode information is critical for predicting transformations the compound will undergo when bound and processed by metabolic enzymes. Undisputed speed of the CADD methods together, with the implementation of parallel computing, allows for much broader coverage of the drug-like chemical space than what is achievable by the experimental methods (Bender et al., 2021; Grebner et al., 2020). For example, using the VirtualFlow VS platform, a billion compounds can be processed in approximately two weeks (Gorgulla et al., 2020). In the following sections, several widely used structure- and ligand-based CADD methodologies are presented.
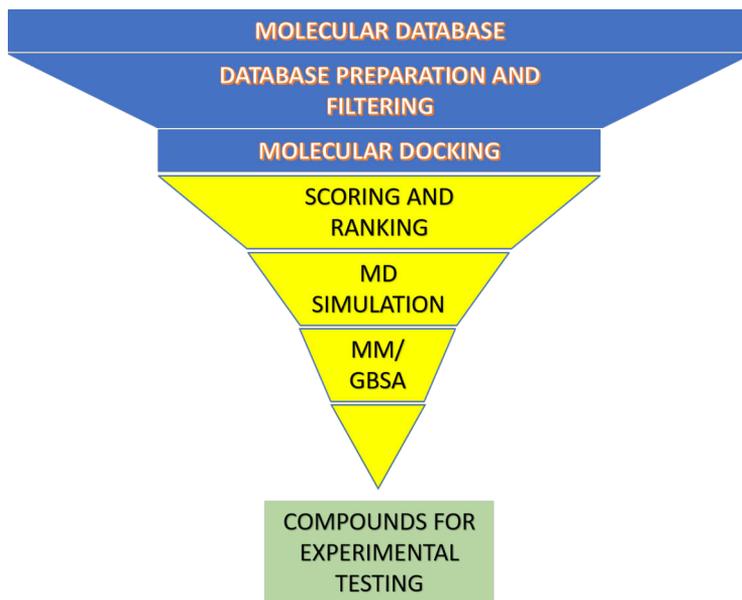
Figure 3. A structured-based virtual screening workflow. A molecular database is prepared and filtered, for example, by molecular weight, to contain drug-like compounds. The high-throughput VS is performed by evaluating the potential of each compound to bind to the target protein. The compounds are scored and ranked by their predicted affinity, possibly by multiple scoring functions. The most potential VS hits can be subjected to computationally more intensive analyses, such as MD simulations and free energy calculations. Finally, the compounds identified as the most promising will be tested experimentally.

2.3.1   Molecular docking

Molecular docking is one of the most widely applied methods in SBDD (Pinzi and Rastelli, 2019; Shoichet et al., 2002). In the general docking scheme, a conformational sampling algorithm searches low-energy configurations for a ligand and fits them into the designated binding site. A scoring function is used to assess the quality of the generated poses based on the interactions between the ligand and the targeted binding site.

In the conformational sampling phase, ligand degrees of freedom can be systematically or stochastically explored. The general aim in a systematic search is to cover the complete positional, orientational and conformational space available for a ligand. However, such a full search is often computationally too demanding, especially in the case of highly flexible ligands. GLIDE, for example, performs several filtering steps for the conformations obtained from an exhaustive sampling of ligand torsion-angle space before the computationally more expensive energy calculations (Friesner et al., 2004). A genetic algorithm is an example of a stochastic search where the conformations are produced by random bond rotations. Sharing

analogy with natural selection, the fitness of each conformation is evaluated by an energy function after going through mutations, crossovers, or recombination that alter the conformation. This optimization continues until an energy minimum has been reached (Korb et al., 2007; Morris et al., 2009). The balance between algorithm speed and exhaustiveness of the conformational search is crucial, as under sampling the ligand conformational space may decrease docking performance (Bender et al., 2021).

The scoring functions utilized in docking programs are traditionally classified as empirical, knowledge-based or force field-based (Guedes et al., 2018). Empirical scoring functions are trained with known binding affinity data to correlate the binding free energy to factors that describe the binding event. Knowledge-based scoring function development utilizes 3-D structures of protein-ligand complexes. A statistical analysis is conducted with the available data to describe favoured geometries for specific interacting atom pairs. Force field- or physics-based functions are built on classical molecular mechanics force fields and include energy terms for non-bonded interactions between a protein and a ligand and internal ligand energy for bonded and non-bonded interactions. Implicit solvation models are employed with the force field-based scoring functions to account for solvation energy (Liu and Wang, 2015). Some widely used empirical, knowledge-based, and force field-based scoring functions include PLANTS$_{CHEMPLP}$, ITScore and DOCK, respectively (Huang and Zou, 2006; Korb et al., 2009; Meng et al., 1992). In addition, a new class of scoring functions, labelled descriptor or machine learning-based functions, has emerged lately (Liu and Wang, 2015). Rescoring docking solutions with different scoring functions, as well as using docking in combination with other methods, such as similarity screening or pharmacophore modelling, has been reported to provide satisfying results in benchmarking VS studies (Miller et al., 2021; Palacio-Rodríguez et al., 2019; Sastry et al., 2013).

A great variety of docking programs is available both academically and commercially. Due to the varying scoring functions and data sets used to train them, many docking programs have shown a tendency for a highly case-specific performance in identifying active compounds and producing their bioactive binding modes. Performance of different docking programs in affinity and binding pose prediction has been compared in multiple studies, such as the CSAR benchmarking exercises (Carlson et al., 2016; Cross et al., 2009; Parks et al., 2020; Selwa et al., 2016). The typical limitations of docking software include inadequate consideration of solvation effects and target protein flexibility. Including explicit water molecules in docking models has been reported to improve docking performance (Murphy et al., 2016). Techniques to consider protein flexibility, such as soft docking and induced fit docking, consider target site adaptation by softening the van der Waals repulsion of the binding site atoms or by performing conformational sampling for a

user-defined set of binding site residues, respectively (Ferrari et al., 2004; Lexa and Carlson, 2012; Miller et al., 2021). Ensemble docking has been used to dock to multiple structures extracted from MD simulations (Amaro et al., 2008, 2018; Salmaso and Moro, 2018). Docking to a structural ensemble of the target protein was central in explaining the drug gefitinib binding an epidermal growth factor receptor mutant more efficiently than ATP (Wan et al., 2012).

2.3.2 Negative image-based methods

Negative image-based (NIB) modelling aims to consider the shape complementarity between ligands and the binding cavity in VS. The importance of three-dimensional shape complementarity for ligand binding has long been acknowledged (Nicholls et al., 2010). However, the shape is often underestimated in the scoring functions aimed at evaluating ligand affinity as more weight is put on the contribution of electrostatics (Virtanen and Pentikäinen, 2010). This was the primary motivation for developing PANTHER, a program used to generate a negative image of the targeted binding cavity (Niinivehmas et al., 2015). A NIB model fills the cavity with spheres that describe the pocket's shape and electrostatic interaction sites. The spheres are either charged or neutral, depending on the amino acid adjacent to each sphere. The NIB model is used in a shape and electrostatics similarity screening of molecular databases performed by the program SHAEP (Vainio et al., 2009). Resulting superimpositions of the screened ligands and the NIB model can be readily visualized with the target cavity structure to design affinity-increasing modifications. The NIB protocol combines structural information of the target protein with an extremely fast ligand-based screening phase, making the method optimal for large-scale SBDD studies. NIB screening has been demonstrated to provide significant early enrichment of active compounds and excellent separation of actives and inactives in VS benchmarking studies with many targets (Ahinko et al., 2019a; Niinivehmas et al., 2011, 2015, 2016; Virtanen and Pentikäinen, 2010). Furthermore, rescoring docking solutions with NIB models improved docking performance with multiple docking software (Kurkinen et al., 2018, 2019). Recently, a greedy search algorithm for optimizing NIB models with existing ligand activity data was published and was shown to boost the enrichment metrics even further with all tested target proteins (Kurkinen et al., 2022). NIB screening was successfully employed in a prospective drug discovery study where novel inverse agonists for the retinoic acid-related orphan receptor γt were found (Rauhamäki et al., 2018). In addition to PANTHER, the idea of generating space-filling NIB models to consider a precise description of the shape and geometric and topographic constraints of the binding cavity has been applied in, for example, SHAPE4 and DOCK (Ebalunode et al., 2008; Shoichet and Kuntz, 1993).

2.3.3   Binding free energy calculations

In theory, molecular docking programs should be able to give estimations of affinities between ligands and their target proteins by using their scoring functions. However, the predictions produced by docking typically lack accuracy due to approximate modelling of the events contributing to binding free energy, applied for speed in the screening procedure. While docking may still provide enrichment of active compounds that is sufficient for large-scale VS, hit-to-lead and lead optimization phases require a more rigorous estimation of the binding free energy change upon ligand binding. The commonly used end-point and alchemical free energy methods, Molecular Mechanics/Generalized Born Surface Area (MM/GBSA) and free energy perturbation (FEP), respectively, require extensive sampling of the studied system by MD simulations, making them computationally too demanding for straightforward use in VS. In a typical SBDD workflow, a portion of the top-ranked docking solutions is subjected to these methods (Gopinath et al., 2020).

MM/GBSA is based on calculating the energy of the free and bound states of the system (Kollman et al., 2000). Typically, the analysis is performed on an MD simulation trajectory of the protein-ligand complex. Running several simulation replicas is generally recommended to provide statistical significance for the binding free energy calculation by obtaining an ensemble average value from multiple independent simulations (Wan et al., 2020). MM/GBSA calculations are relatively fast due to the utilization of the Generalized Born continuum solvation model in the solvation energy calculation. MM/GBSA has often outperformed scoring functions of docking software in ranking ligands by binding affinity (Ahinko et al., 2019b; Hou et al., 2011; Niinivehmas et al., 2011; Sun et al., 2014a). Case-specific performance is reported considering the correlation between MM/GBSA energy and experimental binding affinity (Niinivehmas et al., 2011; Sun et al., 2014b; Virtanen et al., 2015; Ylilauri and Pentikäinen, 2013). Error in the binding energy predictions originates mainly from the lack of conformational entropy and explicit consideration of water-mediated interactions (Genheden and Ryde, 2015). The effect of various parameters, such as simulation length, dielectric constants, and inclusion of the entropy term, on the method's precision has been extensively studied recently (Hou et al., 2011).

FEP is considered a high-end approach among the binding energy calculation methods (Cournia et al., 2020). Usage of FEP in SBDD is limited by its high computational cost that originates from the necessity to sample unphysical intermediate states between the initial and final states of the system (Shirts et al., 2007). Performance of the FEP+ method developed by Schrödinger was recently assessed prospectively by predicting binding affinities in active drug development projects (Schindler et al., 2020; Wang et al., 2015). For 17 of the 19 studied chemical

series, the pairwise root-mean-square error between the relative predicted and experimental binding energy was < 2.0 kcal/mol. Such accuracy is considered sufficient for screening large libraries or hit-to-lead optimization, but not necessarily for late stage lead optimization. Various other studies have demonstrated FEP to outperform MM/GBSA and docking in ligand affinity and binding mode predictions (Kaus et al., 2015; Pu et al., 2017). One of the significant sources of error in FEP simulations is the inadequate sampling of relevant ligand binding conformations (Boyce et al., 2009; Chodera et al., 2011; Mobley et al., 2007). It has been suggested that MM/GBSA could be used to detect biologically relevant ligand binding modes that could be submitted for a more rigorous analysis by FEP (Ahinko et al., 2019b). Recently, FEP was successfully employed in the identification of the bioactive binding mode of a PPI inhibitor in a cryptic binding site (Shan et al., 2022).

2.3.4   Ligand-based methods

Ligand-based CADD methods offer a viable complement or alternative for the structure-based approaches when large amount of ligand activity data is available or if data of the target binding site structure is limited. Ligand-based approaches typically utilize structural similarity comparisons or alignment of active ligands and a subsequent search of the structure-activity relationship. 2-D fingerprinting methods can be employed with many atom-typing or bit scaling schemes and similarity metrics to obtain significant enrichment of active compounds in fast similarity-based VS (Sastry et al., 2010). Performing shape similarity screening with a 3-D structure may improve the identification of novel active scaffolds while retaining the overall shape of the query molecule (Rush et al., 2005; Sastry et al., 2011; Vainio et al., 2009). 3-D pharmacophore methods are used to find an optimal alignment for a set of active compounds where the activity-defining features, such as hydrophobic groups, aromatic rings or hydrogen bond acceptors and donors, are described by pharmacophore points in the 3-D space. The commonly used program PHASE accomplishes this task by defining pharmacophore features based on substructure detection within a set of conformations generated for each aligned molecule (Dixon et al., 2006b, 2006a). Various software for 3-D pharmacophore modelling have been developed that utilize different methods in the conformational sampling of ligands, their alignment and common pharmacophore detection (Schaller et al., 2020).

3-D quantitative structure-activity relationship (3D-QSAR) modelling is another method that relies on the correct alignment of active ligands in the 3-D space, performed, for example, by pharmacophore or similarity screening methods. The field-based QSAR (FQSAR) approach comparative molecular field analysis (CoMFA) is based on measuring van der Waals and electrostatic interactions between a probe atom and the aligned ligands in the form of Lennard-Jones and

Coulomb potentials, respectively (Cramer et al., 1988). Based on the same idea, comparative molecular similarity indices analysis (CoMSIA) was developed to provide smoother and more general distance dependence for the measured properties by using a Gaussian functional form (Klebe et al., 1994). CoMSIA allows computation of property fields for any important physicochemical property, such as hydrophobicity and hydrogen bonding features. The partial least squares (PLS) method is utilized in CoMFA and CoMSIA to define 3D-QSAR between *in vitro* activity data and the high-dimensional data in the property fields. CoMFA and CoMSIA have been commonly employed in the development of 3D-QSARs for ligands of various classes and targets and in the design of novel active compounds (Fan et al., 2018b; Ke et al., 2013; Li et al., 2017; Lorca et al., 2018; Tsai et al., 2006).

2.4 Computer-aided modelling of small molecule metabolism

Drugs are xenobiotic substances to the human body and are thus apace modified into more hydrophilic and easily excreted forms by the metabolic enzymes. The cytochrome P450 (CYP) monooxygenases comprise a significant operator in eliminating both endogenous and exogenous compounds in phase I metabolism. Mainly members of the CYP1, CYP2, and CYP3 families catalyze oxidation reactions of xenobiotics, especially in the liver. Specifically, the hepatic CYP forms 1A2, 2C9, 2C19, 2D6, and 3A4 are responsible for the majority of the metabolic reactions of the clinically used drugs (Guengerich, 2017; Rendic and Guengerich, 2015). Optimization of drug metabolism is an integral part of the drug development (Zhang and Tang, 2018). The pharmacologic effect of a drug may be decreased if it is metabolized to an inactive form too quickly. In contrast, too slow metabolism can cause toxicity by the accumulation of the substance. The matter is complicated by the genetic polymorphism of the metabolic enzymes, leading to individual differences in metabolic activities of specific compounds (Eichelbaum et al., 1992). For example, individuals expressing rapid CYP2D6-mediated metabolism of the analgesic codeine to morphine are known to suffer from opioidergic side effects that may be dangerous (Kirchheiner et al., 2007). Unmanageable toxicity, often related to toxic metabolites, is a common reason for drug failure in clinical trials (Sun et al., 2022). On the other hand, the prodrug technology is based on delivering an inactive form of the drug to its target site, where it becomes converted to an active metabolite, for example, by the CYP enzymes (Ortiz de Montellano, 2013). Statins are a well-known class of lipid-lowering drugs that employ a prodrug approach mediated by the esterase and paraoxonase enzymes. In contrast, inhibition of their CYP-mediated metabolism may lead to toxicity (Neuvonen et al., 2006). These aspects considered,

fine control of the metabolic reactions of a drug candidate is crucial to its success in the clinical phase.

Despite animal testing being used for preclinical predictions of substance toxicity in humans, unexpected toxic metabolites may still appear during the clinical phase (Zhang and Tang, 2018). SBDD and CADD methods can be used to help in minimizing the toxicity potential associated with bio activation of drug candidates. CADD methods offer the option of modelling the studied compounds with the human forms of enzymes active in drug metabolism. The usual aim in computational CYP studies is to predict a site of metabolism (SOM) on a substrate compound (Kirchmair et al., 2012). Extremely fast ligand-based methods, such as SMARTCYP and structural alert databases, predict reactive groups on the substrate structure either by knowledge-based or physics-based approaches (Claesson and Minidis, 2018; Rydberg et al., 2010). For example, SMARTCYP considers both quantum mechanics-derived activation energy and accessibility of a specific site in the SOM predictions. Recently published BioTransformer software was reported to combine knowledge-based and machine learning approaches for predicting SOMs and the resulting metabolites in human tissues, the human gut and the environment (Djoumbou-Feunang et al., 2019).

A small molecule's metabolism is not determined solely by the presence of a chemical group subject to, for example, an oxidation reaction. Binding mode of the compound in the active site of a specific CYP enzyme is a major determinant in predicting if the oxidation reaction will occur. By principle, an active substrate needs to bind in an orientation that allows optimal placement of the SOM into the vicinity of the heme iron atom while leaving room for reactive oxygen that participates in the catalysis (Isin and Guengerich, 2008; Kirchmair et al., 2012). Hydrogen bonding and steric interactions in the other parts of the active site contribute to the overall stability of the active binding mode and thus affect rate of the metabolic reaction. Some substrates may adopt alternative binding modes by binding to multiple conformations of a specific CYP, some of which may lead to the formation of additional metabolites or inhibit the activity of the enzyme (Guengerich et al., 2019). Numerous crystal structures have been solved for the relevant CYP isoforms complexed with various xenobiotic substances (Sansen et al., 2007; Sevrioukova and Poulos, 2013; Wang et al., 2012). This structural data enables the utilization of various CADD approaches, including binding mode prediction by molecular docking and MD simulations (Hritz et al., 2008; Kirchmair et al., 2012; Nair et al., 2019; Panneerselvam et al., 2015). Ensemble-based methods, based on docking to multiple crystal structures or structures sampled by MD simulations, may provide crucial information on how the flexibility of different CYP isoforms affects their substrate selectivity (Guengerich et al., 2019; Sheng et al., 2014). Such methods are typically coupled with various metrics for defining the active binding modes,

considering distance and orientation of the substrate to the heme group, fluctuation of the active site residues, stability of the substrate orientation, and binding free energy of different binding modes (Ahinko et al., 2019b; Hritz et al., 2008; Juvonen et al., 2016; Sheng et al., 2014). Recently, the binding of a series of profluorescent coumarin derivatives to the CYP1 family was studied utilizing docking, MM/GBSA, and the former metrics (Juvonen et al., 2021; Raunio et al., 2020). In this thesis, a similar study was conducted with the CYP2A13. Comparisons were made with the CYP2A13's structural homolog, CYP2A6, to understand how slight differences in the active site composition and the resultantly varying binding modes affect the experimentally observed enzyme kinetics (Figure 4). NIB modelling has also been successfully employed in CYP substrate SOM predictions (Juvonen et al., 2019).
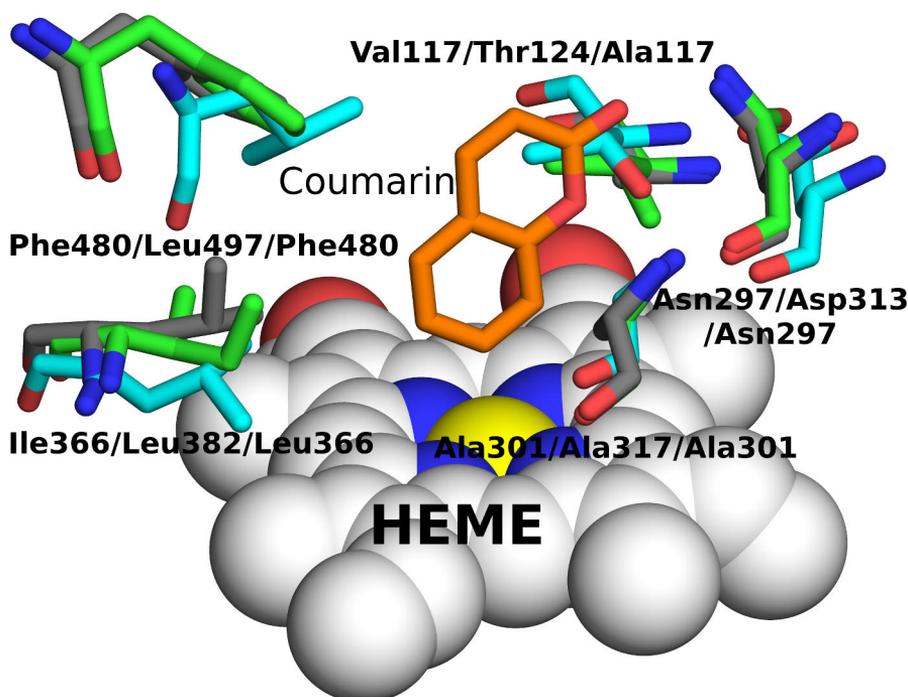
Figure 4. Superimposition of the active site residues of CYP2A6, CYP1A2, and CYP2A13. Heme group-lining residues that differ between the three CYP forms are displayed. The residue labels are in the order: CYP2A6/CYP1A2/CYP2A13. The heme group and coumarin are from the CYP2A6's PDB structure 1Z10. Carbon atoms are colored as follows: CYP2A6 amino acids green, heme white, coumarin orange, CYP1A2 cyan, and CYP2A13 grey. Other atoms are colored: oxygen red, nitrogen blue, and iron yellow.

3   Aims

The main goal of this thesis was to demonstrate the utility of current *in silico* SBDD approaches in phases of preclinical drug discovery and development. Specifically, in Study I, the aim was to test F-NIB, a novel VS approach, in a prospective drug discovery study where novel inhibitors were searched for a disease-related target protein, phosphodiesterase 10A (PDE10A). Studies II and III aimed to explain metabolism and binding selectivity of novel CYP ligands by considering the flexibility and structural differences of the active sites via computational approaches. Finally, Study IV aimed to detect potential small molecule binding sites on a crucial protein of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The detected sites were targeted with VS to investigate if any known drugs could inhibit the internalization-mediating interactions between the virus and its host cell receptor.

4   Materials and Methods

A summary of the main methods used in this thesis is given here. A more comprehensive description of the methods and their settings can be obtained from the original publications. Table 1 lists the most important methods and software used in Studies I-IV.

Table 1.   Primary methods and software used in Studies I-IV.

Method	Purpose	Software	Original publication	Software and protocol references
NIB modelling	NIB or F-NIB model generation	PANTHER	I, IV	Niinivehmas et al., 2015
Similarity screening	Virtual screening or rescoring	SHAEP	I, IV	Vainio et al., 2009
Molecular docking	Virtual screening or binding mode prediction	PLANTS	I, II, III, IV	Korb et al., 2009
FQSAR	Rescoring	MAESTRO	I	Schrödinger, LLC, New York, NY
MD simulation	Ligand binding stability	NAMD AMBER	II, IV	Phillips et al., 2005; Salomon-Ferrer et al., 2013
MixMD	Binding site detection	AMBER	IV	Ghanakota and Carlson, 2016b; Salomon-Ferrer et al., 2013
MM/GBSA	Binding free energy calculations	Prime MM-GBSA MMPBSA.py	I, II, IV	Schrödinger, LLC, New York, NY Miller et al., 2012
Data analysis	Analysis of MD simulation or docking data	CPPTRAJ SDFCONF PROBEVIEW	II, III, IV	Graham et al., 2018; Lätti et al., 2022; Roe and Cheatham, 2013
Visualization	Visualization of simulations and docking solutions	BODIL VMD PYMOL	I, II, III, IV	Humphrey et al., 1996; Lehtonen et al., 2004 Schrödinger, LLC

Elmeri Jokinen

4.1 Protein structure selection and preparation

In Studies I-IV, high-resolution (<= 3.0 Å) crystal structure models of the studied proteins were obtained from Protein Data Bank (PDB; rcsb.org; Berman et al., 2000; Table 2). In addition to resolution, structure selections were based on completeness of the amino acid sequence and variation of the bound small molecules or conformational state of the binding regions. In Study I, two crystal structures with subnanomolar inhibitors were selected. The inhibitors represented different occupations of the active site. In Study II, structures of CYP1A1, CYP1A2, and CYP1B1 in complex with the CYP1 inhibitor alpha-naphthoflavone (ANF) were obtained. In Study III, all the CYP1A13 crystal structures available in PDB at the time were employed to consider the flexibility of the active site in the binding mode predictions. In addition, the crystal structure of coumarin bound with CYP2A6 was used in comparative studies. In Study IV, a crystal structure of the Spike protein (S protein) receptor-binding domain (RBD) in complex with its host cell receptor angiotensin-converting enzyme 2 (ACE2) was chosen due to its adequate resolution and quick availability after the emergence of the novel virus.

Table 2. The crystal structures used in Studies I-IV.

PDB ID	Description	Resolution (Å)	Original publication	Reference
4HEU	PDE10A with an inhibitor	2.00	I	Rzasa et al., 2012
3SN7	PDE10A with an inhibitor	1.82	I	Malamas et al., 2011
4I8V	CYP1A1 with ANF	2.60	II	Walsh et al., 2013
2HI4	CYP1A2 with ANF	1.95	II	Sansen et al., 2007
3PM0	CYP1B1 with ANF	2.70	II	Wang et al., 2011
2P85	CYP2A13 with an indole	2.35	III	Smith et al., 2007
3T3S	CYP2A13 with pilocarpine	3.00	III	DeVore et al., 2012
4EJG	CYP2A13 with nicotine	2.50	III	DeVore and Scott, 2012
4EJH	CYP2A13 with 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone	2.35	III	DeVore and Scott, 2012
4EJI	CYP2A13 with two 4-(methylnitrosamino)-1-(3-puridyl)-1-butanones	2.10	III	DeVore and Scott, 2012
1Z10	CYP2A6 with coumarin	1.90	III	Yano et al., 2005
6M0J	SARS-CoV-2 Spike RBD with ACE2	2.45	IV	Lan et al., 2020

Before molecular docking and virtual screening, entities other than the target protein, such as crystal waters, ligands, ions, or other peptides, were removed. Hydrogens

30

were added to the protein structures using REDUCE (Word et al., 1999). Regarding MD simulations, structures with an incomplete N- or C-terminus were capped with acetyl and N-methyl groups, respectively. Histidine protonation and disulphide bridges (when relevant) were assigned by visual inspection and literature search.

4.2 Ligand preparation and parameterization

Small molecule ligands were either downloaded from public databases (e.g., DrugBank or ChEMBL; Gaulton et al., 2017; Wishart et al., 2006) in 1D or 2D format or extracted from co-crystallized protein-ligand complexes available in PDB. If ligand structures were not readily available, they were drawn in MAESTRO (Schrödinger, LLC, New York, NY; Studies II and III). For molecular docking, ligands were converted to 3D format, and partial charges were assigned with LIGPREP in MAESTRO, using the OPLS3 force field (Harder et al., 2016). Ionization states in pH 7.4 were generated with EPIK (Greenwood et al., 2010; Shelley et al., 2007). The ligands were desalted, and tautomers were generated when appropriate. Specified chiralities were retained, and other chiral centers were varied, generating at most 32 stereoisomers per ligand. Prior to performing VS, molecular databases were filtered by molecular weight (150-550 or 250-600 g/mol) and the number of rotatable bonds (=<8 or =<10) using LIGFILTER in MAESTRO. When necessary, multiple conformations were generated for each compound using CONFGEN in MAESTRO (Watts et al., 2010).

For MD simulations, atomic charges were derived using either the AM1-BCC approach, based on semi-empirical calculations and bond charge corrections (Study IV), or restrained fit to quantum mechanically calculated electrostatic potential (RESP; Study II) (Bayly et al., 1993; Jakalian et al., 2000, 2002). AM1-BCC is computationally cheaper, which is beneficial when multiple compounds need parameterization. Atomic charges derived by either AM1-BCC or RESP have been shown to perform similarly in free energy calculations (Bhati and Coveney, 2022; Manzoni and Ryde, 2018). For the RESP method, electrostatic potentials were calculated with GAUSSIAN16 (Gaussian Inc. Wallingford CT, 2016) at the Hartree-Fock/6-31*G level using the continuum solvent model. AM1-BCC charge derivation and RESP charge fitting were performed with ANTECHAMBER (Wang et al., 2006). All ligands were parameterized using General AMBER Force Field (Wang et al., 2004).

In Study III, pentacoordinate ferric high spin parameters and General AMBER Force Field were used for the cofactor heme group and its proximal cysteine (Shahrokh et al., 2012). In Study IV, the GLYCAM_06j-1 force field was used to parameterize the glycosylated asparagine (Kirschner et al., 2008).

4.3   Molecular dynamics simulations

Classical MD simulations are based on integrating the Newtonian equations of motion coupled with a potential energy function (i.e., force field) that depends on the positions of the atoms:

$$m_\alpha \ddot{\vec{r}}_\alpha = -\frac{\partial}{\partial \vec{r}_\alpha} U_{Total}(\vec{r}_1, \vec{r}_2, \ldots, \vec{r}_N), \alpha = 1, 2, \ldots N$$

where m_α is the mass of atom α, $\vec{r}_\alpha$ is the atom's position and U_{Total} is the potential energy function that couples the motions of atoms and allows exploration of the conformational space of proteins and other biomolecules (Phillips et al., 2005). The simulation time is divided into small steps (usually 1-2 fs) that define the interval at which the energy of the system and coordinates and velocities of the atoms are calculated. The basic routine performed by the standard Velocity Verlet MD simulation algorithm can be simplified into the following:

1. Calculation of potential energy of the current time step.

2. Calculation of coordinates of the next time step.

3. Calculation of energy gradient of the next time step by using the coordinates calculated in step 2.

4. Calculation of velocities of the next time step by using the coordinates and energy gradient calculated in steps 2 and 3.

These steps are repeated until the end of the simulation (Phillips et al., 2005). Additional calculations related to, for example, temperature and pressure control are performed within this basic simulation scheme.

The force field is a crucial element in the simulation as it describes the physical rules according to which the simulation is performed, significantly affecting how realistic the simulation results will be. Common molecular mechanics force fields calculate the potential energy as a sum of covalent and non-covalent interactions:

$$U_{Total} = U_{bond} + U_{angle} + U_{dihedral} + U_{vdW} + U_{Coulomb}$$

Covalently bonded atoms share bonds between neighbouring atoms, angles between atoms separated by two covalent bonds, and dihedrals between atoms separated by three covalent bonds. U_{bond}, U_{angle}, and $U_{dihedral}$ describe the energy related to stretching, bending, and torsional bonding interactions of covalently linked atoms. Non-covalent van der Waals and electrostatic interactions, incorporated by U_{vdW} and $U_{Coulomb}$, respectively, occur between all atoms in the system, their calculation being the computationally most intensive part of the simulation routine. U_{vdW} is typically described by Lennard-Jones 6-12 interaction energy truncated to zero at a specific

distance cut-off to reduce the computational cost. Coulomb law describes the pairwise short-range electrostatic interactions, whereas the particle-mesh Ewald method is used to consider long-range electrostatic in a computationally efficient manner (Darden et al., 1993; Essmann et al., 1995).

4.3.1   Simulation setup and protocol

In Studies II and IV, MD simulations were performed using the NAMD 2.12 and AMBER18 software, respectively (Götz et al., 2012; Phillips et al., 2005; Salomon-Ferrer et al., 2013). Protein was modelled by the AMBER FF14SB force field in all simulations (Maier et al., 2015). Parameter and starting coordinate files for all systems were built with TLEAP of AMBER18. Parameters for protein, ligand, cofactors, ions, and water were loaded, and each system was solvated to a cubic TIP3P water box (Jorgensen et al., 1983). When necessary, counter ions were added to neutralize the system. In Study IV, the co-solvent MD protocol developed by Carlson and colleagues, MixMD, was applied (Ghanakota and Carlson, 2016b). Carlson's method has shown excellent performance in the identification of various biological interaction sites, including orthosteric and allosteric small molecule binding sites (Ghanakota and Carlson, 2016b; Ghanakota et al., 2018; Lexa and Carlson, 2012; Ung et al., 2016). The method utilizes the pmemd MD engine of the AMBER simulation package, which is optimized for efficient usage of graphics processing units to run the simulations (Case et al., 2020; Salomon-Ferrer et al., 2013). For MixMD simulations, the "solvateshell" command was used in TLEAP to surround the protein with a layer of probe molecules before adding water. MixMD simulations were performed with pyrimidine, acetonitrile and isopropanol probes, using 5% (v:v) probe to water ratio. Probe molecule parameters validated for usage with TIP3P water were employed (Lexa et al., 2014).

All simulations were performed with periodic boundary conditions in constant temperature and pressure maintained at 300 K and 1 atm, respectively. Langevin dynamics and the Nosé-Hoover Langevin piston method were used with NAMD in Study II, and Andersen thermostat and Berendsen barostat were used with AMBER in Study IV. A time step of 2 fs was used. The SHAKE algorithm was used to restrain hydrogen motions (Ryckaert et al., 1977). Cut-off or switching for short-range non-bonded interactions was set to 10 Å and 12 Å with AMBER and NAMD, respectively. With NAMD, the switching distance was set to 10 Å. Particle-mesh Ewald was employed to treat long-range electrostatic interactions. The systems were energy minimized before applying the simulation conditions, and equilibration runs were performed with harmonic restraints on protein backbone atoms. The restraining force was gradually decreased after reaching the target temperature and pressure. Production MD simulations were run without any restraints.

4.3.2   Analysis

Simulation trajectory analyses were performed using CPPTRAJ in AMBER18 (Roe and Cheatham, 2013). In Studies II and IV, all the analysed MD simulation trajectories were superimposed with a common reference structure by atomic root mean square deviation (RMSD) fit of protein backbone Cα atoms. RMSD analyses were used to evaluate the stability of ligand binding and conformational changes at potential binding sites identified by MixMD. Common structures of the predicted binding sites were obtained by RMSD-based clustering with CPPTRAJ. Regarding MixMD simulations, the high-occupancy sites of the probe molecules were determined using the "grid" function in CPPTRAJ (Ghanakota and Carlson, 2016b). Sites occupied by multiple probe types were identified using the PYMOL-plugin PROBEVIEW (Graham et al., 2018) that combines and clusters the occupancy data from multiple simulations. The most common binding modes of the probe molecules on the identified hotspots were determined by extracting the closest probe molecules to the hotspots from each snapshot and clustering them with CPPTRAJ. The identified probe poses were incorporated into F-NIB models used in virtual screening.

4.3.3   Binding free energy calculations

MM/GBSA was employed in Studies I, II, and IV to estimate the binding affinity of small molecule ligands to the target binding site. MM/GBSA calculates the binding free energy difference between two states of a system, the bound and unbound states of solvated protein and ligand. The calculation is represented as a thermodynamic cycle that is described by:

$$\Delta G_{bind} = G_{comp} - G_{prot} - G_{lig}$$

where ΔG_{bind} is the binding free energy, calculated by subtracting the free energies of the unbound protein (G_{prot}) and ligand (G_{lig}) from the free energy of the complex (G_{comp}). Each free energy term is calculated by the equation:

$$G = E_{bond} + E_{el} + E_{vdW} + \Delta G_{GB} + \Delta G_{non\text{-}polar} - T\Delta S$$

where E_{bond}, E_{el}, and E_{vdW} are the molecular mechanics terms for gas-phase bonded, electrostatic, and van der Waals interactions. ΔG_{GB} and $\Delta G_{non\text{-}polar}$ are the polar and non-polar components of the desolvation free energy. In MM/GBSA, the Generalized Born model is used to calculate the polar component. In the binding free energy calculations performed in this thesis, the calculation of entropic change ($T\Delta S$) was omitted due to its high computational cost.

In Study I, Prime MM/GBSA (Jacobson et al., 2002) implemented in MAESTRO was employed to estimate the binding energy of the complexes generated by either F-NIB screening or molecular docking. Prime MM/GBSA performs a short energy minimization to the ligand and a user-defined part of the protein. In Study I, residues within 4 Å of the ligand were defined as flexible for the MM/GBSA. The routine was performed using the OPLS3 force field and VSGB solvation model (Li et al., 2011). In Studies II and IV, the binding energy analyses were conducted with trajectories of the simulated protein-ligand complexes. MM/GBSA analysis was performed with MMPBSA.py (Miller et al., 2012) using the igb5 model for Generalized Born calculations (Onufriev et al., 2004). Binding energy estimations were based on the ensemble averages calculated for the simulation trajectories.

4.4   Molecular docking and fragment- and negative image-based screening or rescoring

All docking experiments in Studies I-IV were performed using the PLANTS software (Korb et al., 2007, 2009). PLANTS uses a stochastic ant colony optimization method for conformational sampling and generation of docking solutions. In this work, the docking solutions were scored by the empirical CHEMPLP scoring function. CHEMPLP considers the intermolecular steric, hydrogen bonding, and metal-acceptor interactions as well as intramolecular interactions for both the protein and the ligand (Korb et al., 2009). Docking was employed in VS (Study IV) and binding mode predictions of CYP inhibitors, substrates (Studies II and III), and VS hits (Study I).

In addition to traditional molecular docking, NIB cavity models (Figure 5) were used in virtual screening (Study I) and rescoring of docked poses obtained from virtual screening (Study IV). In NIB modelling, the in-house program PANTHER was used to generate a negative image of the binding cavity (Niinivehmas et al., 2015). The program SHAEP was used to score the screened compounds based on their shape and electrostatic similarity with the NIB model (Vainio et al., 2009). In NIB rescoring, a set of docking poses was generated for each compound and they were rescored by similarity screening with the NIB model. The SHAEP option "noOptimization" was used to ensure the comparison of the exact docking poses to the NIB models without further optimizing the alignment between the compounds and the NIB model.
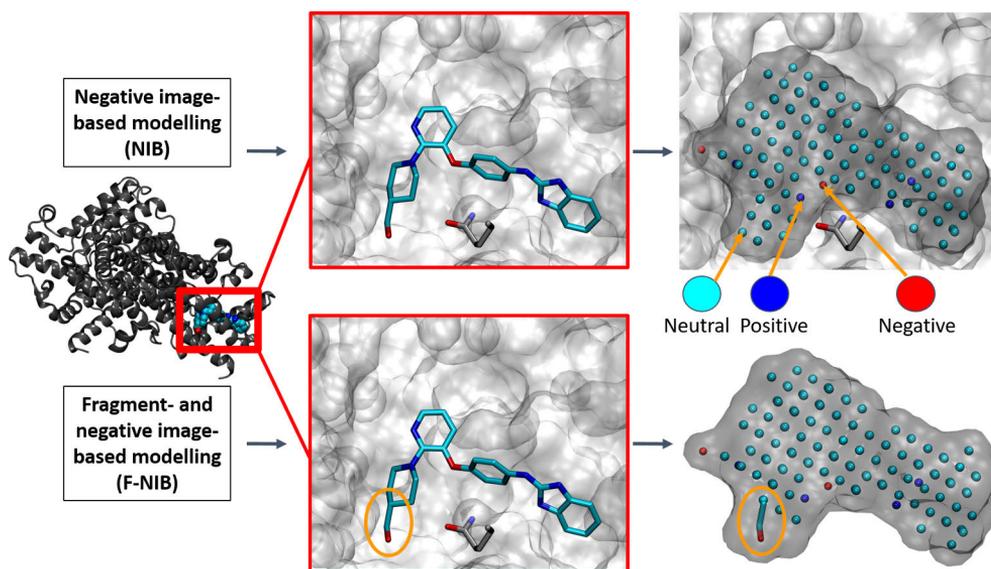
Figure 5. NIB and F-NIB model generation. A NIB model describes the shape and electrostatic properties of a binding cavity (white surface) by using neutral (cyan) and polar (positively charged blue, negatively charged red) spheres. The polar cavity spheres are placed in locations where the atoms of an adjacent amino acid are capable of forming hydrogen bonds. In F-NIB, a molecular fragment describes part of the model more precisely. The fragments can be obtained from, for example, a co-crystal structure where an active molecule is bound to the target cavity. For clarity, the F-NIB model (bottom right) is shown without the protein surface.

In Study I, an extension to the established NIB screening protocol was presented. In fragment- and negative image-based (F-NIB) screening, molecular fragments obtained from co-crystal structures or virtual screening hits are incorporated into an existing NIB model. This way, regions of the model where important binding interactions are formed get described more accurately (Figure 5). Upon fragment incorporation, the overlapping NIB model spheres were removed, while the more general NIB cavity description was retained for the rest of the model. Atomic point charges for the fragments were either obtained from the OPLS3 force field or derived from the AMBER-based charge model of PANTHER. In Study IV, for the polar atoms of the pyrimidines included in F-NIB models, the charges assigned for MixMD simulations were retained.

In NIB- and F-NIB model generation, several parameters were adjusted case-specifically as is usually required due to differences in forms of the modelled binding cavities (Ahinko et al., 2019a). Namely, model centre coordinate, box radius, protein C atom radius, and packing method (face centred cubic or body centred cubic) were varied to obtain the best possible coverage of the essential binding areas for each binding cavity.

4.5 Ligand-based virtual screening and rescoring models

In addition to docking and NIB/F-NIB, pharmacophore modelling and FQSAR modelling were used in virtual screening and rescoring in Study I. Ligand activity data for pharmacophore and FQSAR modelling was obtained from the ChEMBL database (Gaulton et al., 2017) for specific sets of 53 and 78 active PDE10A inhibitors (Malamas et al., 2011; Rzasa et al., 2012, 2014). In the used sets, IC_{50} of the compounds varied from micromolar to subnanomolar. The pharmacophore model was built with PHASE in MAESTRO (Dixon et al., 2006b, 2006a). The FQSAR models were developed using the CoMSIA-based approach implemented in MAESTRO (Klebe et al., 1994). The ligands used for FQSAR model development were aligned either with the F-NIB models or with the pharmacophore model. Steric, electrostatic, hydrophobic, hydrogen bond donor, and hydrogen bond acceptor properties were included in the models. Ligands were randomly divided into training and test sets for training and external validation of the models. Statistical evaluation was performed to select FQSAR models with sufficient predictive power. Especially the R^2, Q^2, and R^2 Scramble values were inspected. R^2 and Q^2 values (in the range 0-1) tell how well the variance in activity data of the training and test set, respectively, is explained by the model. R^2 Scramble is an R^2 value calculated with randomized activity values. Therefore, if R^2 Scramble is comparable to R^2, the model should be considered unreliable. The FQSAR models were used to predict the activities of compounds obtained by pharmacophore or F-NIB screening.

4.6 Detection of active binding modes of CYP2A13 substrates

Scoring functions of molecular docking software often fail in the identification of a high-affinity binding mode. Besides, a catalytically active binding mode is not always equivalent to the one with the highest affinity to the target enzyme. To avoid these pitfalls in Study III, docking was used to generate a set of binding poses for each compound with each CYP2A13 crystal structure, and these solutions were filtered with specific criteria obtained from the literature. Docking conformations where the 7-hydroxylation site was located more than 6 Å from the heme iron atom were discarded using SDFCONF (Hritz et al., 2008; Lätti et al., 2022). Next, conformations where the 7-hydroxylation site was oriented inaccessibly for the heme iron atom were considered unreactive and were discarded (Sheng et al., 2014). Finally, the remaining conformations were visualized, and compounds whose polar atoms did not form any hydrogen bonds with the active site residues were discarded. The assumption was based on the crystal structure showing coumarin binding with CYP2A6, which shares a high sequence identity with CYP2A13 (Yano et al., 2005).

The remaining complexes were subjected to a brief minimization by Prime in MAESTRO. The minimized complexes were visualized to identify binding modes consistent with the experiment based on the placement of the 7-hydroxylation site. Active binding modes of different substrates were compared to explain differences in their experimentally determined kinetic constants.

4.7 Figure preparation

Figures in the Studies I-IV and Figures 1-5 in this thesis were prepared using BODIL (Lehtonen et al., 2004), VMD (Humphrey et al., 1996), MOLSCRIPT (Kraulis, 1991), RASTER3D (Merritt and Murphy, 1994) and PYMOL (The PyMOL Molecular Graphics System, Schrödinger, LLC).

5   Results

5.1   Identification of biological interaction sites on a novel target protein (IV)

Rational drug discovery requires knowledge of the sites on the target protein's structure where a drug can bind and induce the desired pharmacological effect. This information may be minimal, especially when active compounds for the target protein are not known. The SARS-CoV-2 that emerged in late 2019 presented such a challenge for the drug discovery field. Impressive efforts by the scientific community to characterize the novel pathogen provided vast amounts of data extremely quickly, including the first high-resolution structural models of the essential viral proteins within just weeks of the first reported infections (Lan et al., 2020).

Study IV was commenced shortly after the publication of the first crystal structures showing S protein of the SARS-CoV-2 in complex with the human ACE2. SARS-CoV-2 attaches to the host cell by contacting an ACE2 receptor on the cell surface via a specific RBD on the trimeric S protein (Walls et al., 2020). The contact initiates internalization of the virus, after which the host cell is subjected to produce new virions and ultimately release them to its surroundings, resulting in more infected cells. The interaction between the S protein and ACE2 was quickly regarded as one of the potential targets for medicinal intervention. Prevention of the interaction could decrease the rate at which the infection spreads and give more time for the immune system to adapt and respond to the novel pathogen. In Study IV, the goal was to detect potential small molecule binding sites on the ACE2-binding interface of S protein and to search for any known drugs that could bind to these sites and hamper the internalization process.

The three probe types used in the MixMD simulations identified multiple interaction sites for biomolecules on the S protein RBD. A probe occupancy map was derived from the simulations where the sites with high probe residence time could be observed (Study IV, Figure 1). Comparison of the probe occupancy map with the available high-resolution structural models of S protein complexes showed that many probe binding hotspots were located on antibody epitopes (Study IV, Figure S1;Ju et al., 2020; Wu et al., 2020; Yuan et al., 2020). These interaction spots

identified by the probes may contribute to the antibody binding observed on these RBD interfaces. In addition, the probes mapped the binding site of linoleic acid (LA), a cavity completely buried in the starting structure used in the simulations (Study IV, Figure S2). The site opened due to the movement of a helix adjacent to the site, allowing all the used probe types to enter the region. LA has been reported to decrease S protein binding to ACE2, possibly via an allosteric effect (Toelzer et al., 2020). Interestingly, the two sites that showed the highest probe occupancy during the simulations were located on both sides of the LA site (clusters 5 and 6; Study IV, Figures 1 and S2 and Table 1). Both sites showed a groove-like surface shape that could facilitate small molecule binding. Neither of the sites was on the ACE2 interface, due to which they were not targeted by VS in Study IV. A later report, however, indicated that a set of steroidal compounds could bind to the former sites and decrease binding between the S protein RBD and ACE2 (Carino et al., 2020). This information, together with the LA binding data, suggests that MixMD simulations identified a region where the interaction between the S protein and ACE2 could be modulated via an allosteric mechanism.

Looking at the ACE2-binding interface of the RBD, the probes mapped five sites that all have been shown to contribute energetically to the S protein – ACE2 complex formation (Shang et al., 2020; Study IV, Figure 1). The majority of the mapped sites were occupied by just pyrimidine, whereas multi-probe occupation was observed at a site that overlapped with the coordinates of ACE2 Lys353 in the original RBD – ACE2 crystal structure. RMSD-based clustering of simulation structures using the residues lining this probe hotspot revealed two main conformational states for the region. The primary conformation was analogous to the crystal structure state, having a groove-like surface shape extending from an area next to Tyr505 and below Arg403 (Study IV, Figure 2). In 10 – 15% of the MD snapshots, Tyr505 and Arg403 were observed to adopt a secondary conformation that could be inhibitory for ACE2 binding due to Tyr505 blocking the binding region of ACE2 Lys353 (Study IV, Figure 2). Interestingly, the secondary RBD state was ~5 % less prevalent when using pyrimidine as a probe than with the other two probes, acetonitrile or isopropanol. The difference in conformational sampling was reasoned to be caused by pyrimidine's tendency to form stable binding interactions at the region next to Tyr505, preventing it from turning into the alternative conformation (Study IV, Figure 2).

5.2 Molecular fragments provide binding mode information for virtual screening (I, IV)

In Study I, an extension to the well-established NIB screening methodology, F-NIB, was introduced by incorporating molecular fragments into the NIB models. The

fragment coordinates can be obtained from experimental protein-ligand co-crystal structures or other virtual screening hits (Study I, Figure 1; Study IV, Figure 2). In addition, frequently occurring probe poses within the probe hotspots observed in MixMD simulations can be considered molecular fragments and fused into the NIB models. The mentioned approaches were employed in Studies I and IV to guide the NIB screening and NIB rescoring to favour compounds with optimal chemical groups at the coordinates designated by the used fragments.

5.2.1 Simulated probe accumulation captures specific binding interactions (IV)

Of the used probes, pyrimidine showed the highest occupation at the region where Lys353 of ACE2 binds to the S protein RBD (Study IV, Figure 1 and Table 1). The most common binding poses adopted by pyrimidine with the two observed RBD conformations were extracted from the simulation trajectories by RMSD-based clustering. With the primary RBD conformation, pyrimidine binding was stabilized by hydrogen bonding with Gly502 and pi-stacking with Tyr505 of the RBD, indicating spots where similar functional groups could be placed by active small molecules (Study IV, Figure 2). Clustering of pyrimidine poses using the secondary RBD conformation showed pyrimidine forming pi-stacking and hydrogen bonding interactions with the shifted Tyr505 and Arg403, respectively (Study IV, Figure 2). A second virtual screening model was derived from this observation as it was hypothesized that small molecules could be used to stabilize such a potentially inactive RBD conformation.

The pyrimidine poses were used in the subsequent rescoring of docked compounds by incorporating them into the F-NIB models. Visual inspection of the top compounds, as scored by each F-NIB model, confirmed that the top-ranked VS hits generally placed an aromatic group next to Tyr505. Compounds that were able to form hydrogen bonds within the region adjacent to Tyr505 were analysed by MD simulations and MM/GBSA to observe their stability and binding energy in the predicted binding site.

The compounds docked to the secondary conformation of the RBD were generally not stable in MD simulations. The sole exception, DB08434, had relatively poor binding energy according to the MM/GBSA analysis (Study IV, Table 2, and Figure S5). Five compounds showed stable binding to the primary RBD conformation (DB01937, DB02651, DB03714, DB08248, and DB14826). The binding of these compounds showed similarity to pyrimidine binding in the MixMD simulations. An aromatic group occupied the area next to Tyr505, and the binding was further stabilized by hydrogen bonding mainly with the residues Arg403 Lys417, Gly496, Asn501, and Gly502 (Study IV, Figures 3 and S5). Binding free

energy values, as estimated by MM/GBSA, ranged from -19.7 to -28.9 kcal/mol for these compounds (Study IV, Table 2). Based on the MD simulations, the binding of the compounds was mediated by the interactions identified by clustering the predominant pyrimidine poses observed in the MixMD simulations.

5.2.2 Molecular fragments guide negative image-based virtual screening to detect active inhibitors (I)

PDE10A is an enzyme that functions primarily in the striatum in the brain by catalyzing the hydrolysis of the intracellular second messengers cAMP and cGMP to their inactive forms, AMP and GMP (Fujishige et al., 1999). The significant regulatory role of the PDE family, along with a record of accomplishment in PDE-targeted drug development, has drawn attention to PDE10A as a potential target for antipsychotic agents.

In Study I, the active site of PDE10A was targeted with the VS method F-NIB to find novel inhibitors that could prevent the binding of the cyclic nucleotides. The vast amount of structural and ligand activity data available for PDE10A also enabled the usage of other ligand- and structure-based VS tools such as pharmacophore and FQSAR modelling. Two crystal structures of the PDE10A complexes with two structurally different inhibitors were used as templates for the F-NIB models. The known PDE10A inhibitors occupy several specific regions of the active site that are known to contribute to affinity and selectivity, designated as "Gln interaction," "selectivity pocket," "hydrophobic clamp," and "buried waters" (Chappie et al., 2012; Study I, Figure 2). Molecular fragments for the F-NIB models were obtained from the inhibitors bound in the co-crystal structures or from a pharmacophore VS hit compound, emphasizing variation in occupation of the named regions between the different models. Specifically, the utilized fragments occupied the "Gln interaction" and "hydrophobic clamp" areas in F-NIB Models III and IV and the "buried waters" region in F-NIB Model I. Models I and III were also extended to the "selectivity pocket" using the default NIB model representation (Study I, Figure 3).

F-NIB Models I and IV were used to align and geometry optimize specific sets of active PDE10A inhibitors for FQSAR model development. The statistics of the FQSAR models based on the ligand alignment by F-NIB Models I and IV indicated good predictive power as models utilizing 1 and 3 PLS factors had R^2 values of 0.89 and 0.74 and Q^2 values of 0.89 and 0.72, respectively (Study I, Table S1). No significant overfitting was observed by inspecting the R^2 Scramble values of both models (R^2 Scramble 0.38 for Model I and 0.46 for Model IV), making the models feasible for a prospective VS study. The FQSAR models were used to rescore a portion of the molecular database that scored best in VS performed by a corresponding F-NIB model. Two of the three compounds selected for *in vitro*

testing, based on the Model IV and I FQSAR activity predictions and visual inspections, were active with IC$_{50}$ values at ~27 μM and ~49 μM, respectively (compounds **1** and **2**; Study I, Figure S6 and Tables 1 and S4). No rescoring scheme was applied in VS using the F-NIB Model III. The top-scored compounds of the Model III VS contained varying structural scaffolds that were visually evaluated to be suitable for hydrogen bonding within the "Gln interaction" region. Nine of these compounds were tested experimentally, one being found active with an IC$_{50}$ of ~67 μM (compound **3**; Study I, Figure S6 and Tables 1 and S4). Detection of active compounds with Models I, III, and IV demonstrated F-NIB's applicability in both straightforward VS and consensus approaches where the compounds aligned and scored best by F-NIB are re-evaluated with another scoring function.

Generally, the novel active compounds formed hydrogen bonds with Gln716 or Tyr683 or both and filled the "hydrophobic clamp" region with a ring structure (Study I, Figures 2 and 3 and Table S3). Compounds **2** and **3** also utilized the "Selectivity pocket" region of the active site, whereas compound **1** left it unfilled. The exact binding mode of compound **3** remained uncertain as F-NIB and molecular docking suggested different solutions with opposite compound orientations (Study I, Figures 3 and S5). No significant difference for these binding modes was observed in the binding energy calculation by Prime MM/GBSA, suggesting that both orientations could be biologically relevant. 2D fingerprint similarity comparison using the Tanimoto coefficient showed structural uniqueness for the identified compounds compared to the known PDE10A inhibitors (Study I, Table 1). The PDE10A inhibitors identified by F-NIB could therefore provide starting points for developing potent lead compounds with novel structural scaffolds.

5.3   Characterization of CYP tool compounds (II, III)

In Studies II and III, the binding of small molecules to the catalytic sites of the CYP1 and CYP2A13 enzymes was studied. CYP1A2 has a major role in metabolism of xenobiotics in the liver, whereas CYP2A13 is known to mediate bioactivation of several procarcinogens in the respiratory system (Rendic and Guengerich, 2015; Su et al., 2000). Docking and MD simulations were utilized to explain the CYP1A2-selective activity of a novel inhibitor and the binding kinetics of a set of profluorescent coumarin derivative substrates of CYP2A13. Study II focused on how structural variation between the CYP1 enzyme active sites affects the catalytic outcome of small molecule binding. Study III emphasized studying how structural modifications of the 3-phenyl-coumarins affect their binding and the resultant kinetic constants with the CYP2A13.

5.3.1 Small changes in the binding of a novel CYP1A2 inhibitor induce selectivity (II)

N-(3,5-dichlorophenyl)cyclopropanecarboxamide (DCPCC) is a novel inhibitor of CYP1A2 (Raunio et al., 2016). DCPCC inhibited CYP1A2-mediated conversion of non-fluorescent 7-ethoxyresorufin and a set of 3-phenyl-coumarins to their fluorescent metabolites with IC_{50} of 0.20 – 0.71 µM. Inhibition was 10-95 times less efficient with CYP1A1 and CYP1B1. Moreover, DCPCC itself was rapidly oxidized to a single metabolite by CYP1A1, while CYP1A2 and CYP1B1 formed the metabolite at a much slower rate. A structure-based explanation for these observations was sought by modelling the binding of DCPCC at the active sites of CYP1A1, CYP1A2 and CYP1B1.

MD simulations were performed starting with multiple docking solutions for DCPCC binding into the active site of each CYP1. Based on the simulations, stability and binding interactions of each pose were assessed, and MM/GBSA analysis was performed with the simulations trajectories to estimate the binding energy of each complex. The lowest predicted binding energies for DCPCC binding to CYP1A1, 1A2 and, 1B1 were -35.5 kcal/mol, -38.0 kcal/mol and -33.0 kcal/mol, respectively (Study II, Figures S2, S3, and S4, and Table S1). Compared to the experimental IC_{50} values, DCPCC was correctly ranked to bind CYP1A2 most efficiently, whereas the rank order for 1A1 and 1B1 differed from the experiment. MM/GBSA estimates the binding free energy of complex formation, whereas IC_{50} may be affected by factors such as the fast metabolism rate. For example, the fast oxidation of DCPCC by CYP1A1 may increase the experimental IC_{50} value due to the compound rapidly converting to a metabolite with lower inhibitory activity. Such events are not captured by the classical MD simulations, making it extremely difficult to consider their contribution to binding free energy.

The low-energy binding mode of DCPCC with CYP1A1 and 1A2 was predicted to be very alike. In both complexes, the cyclopropane group was determined to be the primary site of metabolism as it was orientated towards the heme iron (Study II, Table S1). A single hydrogen bond formed between the amide group of DCPCC and the sidechain of Ser122 and Thr124 of CYP1A1 and 1A2, respectively (Study II, Figure 7). The dichlorophenyl group formed pi-stacking interactions with phenylalanine residues lining the active site. The cyclopropane and dichlorophenyl groups were similarly orientated in the DCPCC-CYP1B1 complex. However, a hydrogen bond formed with Thr334 on the opposite side of the active site to the Ser122/Thr124 site of CYP1A1/1A2 (Study II, Figure 7). Ser122/Thr124 is replaced by Ala133 in CYP1B1, abolishing the possibility of forming a hydrogen bond at this position. In all three low-energy complexes, the binding of DCPCC was stabilized after the single hydrogen bond formed, even though the original docking solution did not contain such interaction in all cases (Study II, Figure 7).

The fast DCPCC oxidation by CYP1A1 and decreased catalytic activity of CYP1A2 were linked to subtle differences in the cyclopropane group positioning in the predicted DCPCC binding modes. Based on the simulations, the cyclopropane resided close to the Ile386 sidechain ethyl and did not completely cover the heme iron with CYP1A1 (Study II, Figure 8). With CYP1A2, the conformation of Ile386's sidechain was altered by the presence of Thr124 sidechain's methyl, shifting the cyclopropane placement right above the heme iron (Study II, Figure 8). The cyclopropane occupying too much space over the heme iron could prevent molecular oxygen from participating in the reaction and result in an inhibitory effect instead of catalysis. Regarding CYP1B1, the distance between the cyclopropane and heme iron was slightly greater than with CYP1A1 or 1A2 (average distances from simulations: 5.2 Å for 1B1, 4.8 Å for 1A1, and 4.4 Å for 1A2; Study II, Figure S5), which could be non-optimal for the oxidation reaction to proceed.

5.3.2 Alternative binding poses were identified for a classical CYP1 inhibitor (II)

The binding of the CYP1 inhibitor alpha-naphthoflavone (ANF) was evaluated using the same modelling approach as with DCPCC. The results were compared to the available crystal structures with ANF bound to CYP1A1, 1A2, and 1B1. ANF exceeds DCPCC in potency but does not show significant selectivity within the CYP1 family. These qualities were reflected in the calculated ANF binding energies as the values were within the range -48.5 to -49.4 kcal/mol with all three CYP1 enzymes, suggesting tighter but non-selective binding (Study II, Figures S6, S7, and S8, and Table S2).

The low-energy pose of ANF with CYP1A1 was similar to the binding pose observed in the crystal structure with a phenyl ring orientated towards the heme group and a single hydrogen bond formed between carbonyl oxygen of ANF and Asn222 of CYP1A1 (Study II, Figure 9). The crystallized binding pose of ANF with CYP1A2 is highly similar to the one observed with CYP1A1. Such binding pose was ranked second by binding energy calculations of the ANF-CYP1A2 complexes, with a 1.2 kcal/mol difference from the first-ranked pose (Study II, Table S2). The pose ranked best had ANF in an orientation where the phenyl group pointed away from the heme group while the space above the heme was occupied by the benzo(h)chromen-4-one system (Study II, Figure 9). This three-ring moiety has been reported to contain a site of slow CYP1-mediated metabolism of ANF, indicating that both the first and the second-ranked poses are possibly biologically relevant (Bauer et al., 1995). For ANF-CYP1B1, ANF binding pose highly similar to the ANF-CYP1A1 and 1A2 crystal structure poses was predicted to have the lowest energy. However, this differs from the ANF pose observed in the ANF-CYP1B1

crystal structure, where the compound's orientation is flipped roughly 180° around its longitudinal axis (Study II, Figure 9). A crystal structure-like binding pose was ranked sixth with a 6.2 kcal/mol binding energy difference from the best-ranked pose. Un-modelled electron density can be observed in the density difference map of the ANF-CYP1B1 crystal structure (PDB: 3PM0), possibly suggesting the presence of an alternative binding pose for ANF.

5.3.3 Binding site conformation matters in docking-based search of active CYP2A13 substrate binding modes (III)

The binding of coumarin derivatives (Study III, Figure 1) to the active site of CYP2A13 was modelled to study how structural variations affect their Michaelis-Menten kinetic constants in a 7-hydroxylation reaction. Modelling was focused on coumarin and nine other compounds that were metabolized more efficiently by CYP2A13 than coumarin, based on their measured intrinsic clearances (V_{max}/K_m; Study III, Table 1). Multiple docking solutions for each substrate were generated using all available CYP2A13 crystal structures to consider the flexibility of the active site in the binding predictions. The docking complexes were inspected and filtered to detect binding modes that would allow 7-hydroxylation of each substrate. The binding modes that passed these criteria are called 'active binding modes'.

Comparing the CYP2A13 crystal structures shows that the active site undergoes structural rearrangements in adaptation to different binders. The active site residues Phe107, Phe118, Phe209, Phe300, Thr305, Leu366, and Leu370 express relatively subtle conformational changes. In contrast, Met365 shows more significant movements that affect the space available for a binding molecule above the heme group (Study III, Figure S2). One of the crystal structures (PDB: 4EJI) differs significantly from the others due to the formation of a channel leading out of the active site. Movements of two residues lining the active site, Gln104 and Phe480, induce the channel formation. In the crystal structure, one of the two ligands present in the active site occupies the channel.

The effect of the catalytic site conformation on the detection of active binding modes was tested using the different crystal structures in docking (Study III, Table S1). Except for 4EJI, at least one active binding mode was generated by docking for nine of the ten studied compounds. With 4EJI, many of the docking poses occupied the additional channel, ending too far from the heme group for 7-hydroxylation. Docking with the PDB-structure 4EJH showed the best performance as an active binding mode for five of the compounds was ranked first by docking score and for three more compounds, a suitable pose was within the top three poses. With the PDB-structure 4EJG, nine compounds had a relevant binding mode within the top

three docking poses. The PDB-structure 3T3S was the only one where docking produced a binding mode suitable for 7-hydroxylation for compound **10**. Compound **10** was the largest of the active compounds and did not fit well in the active site with any of the crystal structures.

5.3.4 Identification of active binding modes of CYP2A13 substrates (III)

Predicted active binding modes of the substrates were compared with each other to explain differences in their catalytic efficiency by CYP2A13. In the case of coumarin, a comparison was also performed with a crystal structure (PDB: 1Z10) that has coumarin bound to the active site of the structurally homologous CYP2A6 (Study III, Figure S4). Coumarin metabolism was less efficient by CYP2A13 than by CYP2A6 (Study III, Table 1). A comparison of the CYP2A6 co-crystal structure with the CYP2A13 docking solutions for coumarin showed that the orientation of the compound's 7-hydroxylation site differed with the two enzymes (Study III, Figure 5). With CYP2A13, the 7-hydroxylation site was orientated slightly away from the heme iron atom due to the sidechain of Ala301 occupying space over the heme group. In CYP2A6, the corresponding residue, Gly301, leaves more space in the area and allows more optimal orientation for coumarin. Of the modelled substrates, 7-ethoxycoumarin and scoparone were structurally the most analogous to coumarin. In their predicted active binding modes, the 7-ethoxy and 7-methoxy substituents lift the coumarin core higher from the heme group level than with coumarin, which lacks corresponding substituents (Study III, Figure 5). All three compounds form a hydrogen bond with Asn297 via their carbonyl oxygen. Elevated positioning of the coumarin core of 7-ethoxycoumarin and scoparone likely enhances the stability of the binding by pi-stacking interactions with Phe107, Phe118, Phe300, and Phe480 that line the active site. The faster metabolism of 7-ethoxycoumarin than scoparone was explained by the 7-ethoxy being able to form more compact hydrophobic packing with the surrounding residues Leu370, Leu366, Thr305, and Ala301 than the shorter 7-methoxy (Study III, Figures 5 and S3). Consistently, 7-ethoxycoumarin was less actively metabolized by CYP2A6. Gly301 and Ile300 in CYP2A6 replace CYP2A13's Ala301 and Phe300, respectively, which could cause decreased hydrophobic packing and pi-stacking interactions, and reduce the stability of the active binding mode.

Seven of the studied 3-phenyl-coumarins exceeded coumarin in oxidation efficiency by CYP2A13 (Study III, Table 1). In the active binding modes identified for the 3-phenyl-coumarins, the coumarin core was generally placed closer to the heme group than 7-ethoxycoumarin or scoparone (Study III, Figures 5, 6, and S3). A single hydrogen bond was formed to Asn297 of CYP2A13 via the carbonyl

oxygen of the coumarin core of each 3-phenyl-coumarin. The 3-phenyl resided in the phenylalanine-rich region of the active site (Study III, Figure 6).

Substituents in the 6-position of the coumarin core and in the *para-* or *meta-* positions of the 3-phenyl modulated the activity of the substrates. Compound **15** had the highest V_{max} with hydroxyl groups at both the 6- and 3-phenyl-*meta* positions. The 6-position hydroxyl was predicted to stabilize the anchoring of the compound to an optimal orientation for catalysis via electrostatic interaction with the heme nitrogens (Study III, Figure 6). The hydroxyl group at the 3-phenyl-*meta* position was estimated to form direct or water-mediated hydrogen bonds with the backbone oxygen of Leu241 or Leu296. These backbone atoms are not readily accessible for a hydroxyl group at the *para* position of the 3-phenyl, which could be the reason for decreased activity of compounds **7** and **6** compared to **15** and **4**, respectively (Study III, Figures 1 and 6). Likewise, the acetate group at the *para-* or *meta-*position of the 3-phenyl decreased activity (compounds **20** and **22**), whereas corresponding hydrophobic substituents resulted in lower K_m values (compounds **21**, **3**, and **1**). Carbonyl oxygen of Leu241 and Leu296 may be too close to allow the fitting of an acetate group to the region without causing electrostatic repulsion. In contrast, a multitude of hydrophobic sidechains, namely those of Leu110, Leu241, Leu244, and Leu296, facilitate the binding of compounds with halogen substituents at the 3-phenyl-*para* position through hydrophobic interactions (Study III, Figure 6). The low V_{max} of **5** despite carrying 6-hydroxyl at the coumarin core indicates the importance of the 3-phenyl substituents in stabilizing the compounds into the active binding modes.

Compared to the compounds carrying 6-hydroxyl (**15** and **7**), compounds with 6-methyl, 6-methoxy or 6-chloro substituents (**21**, **3**, **1**, **4**, and **6**) expressed lower V_{max} values. The docking suggested hydrophobic packing between the 6-position substituents and Met365 to orientate **21** to an active binding mode (Study III, Figure 6). Met365 can be observed in varying conformations in the CYP2A13 crystal structures (Study III, Figure S2), indicating that such stabilization may be transient. Fluctuation of Met365 occasionally creates a crevice between Thr305 and Leu366 that could be utilized by hydrophobic 6-position substituents for additional binding interactions (Study III, Figure S2). Occupation of the crevice was predicted to cause orientation of the 7-hydroxylation site that is less optimal for oxidation (Study III, Figure 6). In addition to affecting the binding mode of 3-phenyl-coumarins, a bulky group, such as 6-methoxy, may decrease the accessibility of the 7-hydroxylation site, resulting in slower oxidation.

Compound **10** was the largest 3-phenyl-coumarin and did not fit well to the CYP2A13 catalytic site in the docking solutions. Only one docking pose for **10** was retained after filtering inactive poses (Study III, Figure S3). Even with this binding mode, steric overlap between **10** and the active site residues was relieved only after

an energy minimization. Thus, the active binding mode predicted for **10** was considered questionable. Interestingly, **10** was the only metabolically active 3-phenyl-coumarin in the series with a substituent at the 7-position of the coumarin core. Compound **10** likely binds to an active site conformation that the available crystal structures have not captured.

6 Discussion

6.1 Co-solvent simulations have the potential to extend the druggable proteome

From the drug discovery perspective, a very complex problem was addressed in Study IV by targeting a protein of a novel pathogen with no information on active small molecule modulators or their binding sites. In such a scenario, the first question is that is the novel target protein druggable in the first place. If there is a high-resolution structure of the protein available, co-solvent MD simulation methods, such as MixMD, can be employed to detect potential sites for molecular interaction on the structure. Especially interesting is that the organic probe molecules have been reported to induce binding site conformations that are structurally close to known ligand-bound states, even if the simulation was started from an unbound state (Kimura et al., 2017; Schmidt et al., 2019). As was observed in Study IV, the conformational space explored by a co-solvent simulation may depend on the used probe type (Kimura et al., 2017). In Study IV, MixMD performed consistently with earlier work done by many others, including varying implementations of the co-solvent MD simulation methodology (Ghanakota and Carlson, 2016b; Ghanakota et al., 2018). MixMD identified interaction hotspots on PPI interfaces and small molecule binding sites, including a site buried in the used S protein RBD crystal structure. Detection of binding sites inaccessible in the experimental structures could enable development of drugs with novel pharmacology for known targets and proteins that have been considered undruggable. The co-solvent MD methods are looking very promising for cryptic or occluded binding site identification, mainly because such sites may be difficult to detect with experimental methods (Bakan et al., 2012; Comitani and Gervasio, 2018; Kalenkiewicz et al., 2015; Kimura et al., 2017; Oleinikovas et al., 2016; Schmidt et al., 2019; Smith and Carlson, 2021; Tze-Yang Ng and Tan, 2022). Targeting drug discovery to cryptic binding sites has been estimated to increase the size of the druggable human proteome from ~40% to ~78% (Cimermancic et al., 2016).

The discovery of interaction hotspots on PPI interfaces by co-solvent MD simulations could become a timely tool for PPI targeted drug discovery. Interestingly, MixMD seemed to perform well in identifying essential interaction

sites on the ACE2 binding interface of the S protein. In addition, sites on multiple antibody epitopes were mapped, and could be speculated to have similar significance in S protein complex formation with the antibodies. Recent work by others has similarly demonstrated the feasibility of co-solvent MD simulations in the identification of various binding sites and in inducing PPI site conformations that resemble the ligand-bound states (Ghanakota et al., 2018; Kalenkiewicz et al., 2015). The utilization of enhanced sampling techniques with the co-solvent simulations has drawn particular interest regarding PPIs where binding pocket formation requires crossing a sizeable potential energy barrier (Oleinikovas et al., 2016). Such an approach could reveal novel druggable PPI sites for drug discovery purposes.

Regarding drug development, it is not enough to know the locations of potential interaction sites on a protein structure. In addition, knowledge must be gained of which of the sites are biologically and, more specifically, pharmacologically relevant. Unfortunately, probe accumulation during co-solvent MD simulations does not tell much about which ones of the identified interaction sites should be targeted to achieve the desired pharmacological effect. In Study IV, probe hotspot coordinates were coupled with the knowledge of the S protein – ACE2 complex structure. This allowed applying a straightforward screening approach for inhibitors that could block the interaction by steric hindrance. In addition, compounds that could lock the S protein's ACE2-binding interface into a potentially inhibitory conformation, observed during the MixMD simulations, were sought. Another possibility would be to search for more distant probe hotspots with allosteric communication to the region of interest. Such communication networks were later reported for the S protein RBD (Carino et al., 2020; Toelzer et al., 2020). MD simulations have been used to identify communication networks between distant pockets within a protein structure for several targets, including the S protein (Carino et al., 2020; Chen et al., 2022; La Sala et al., 2017). Recently, potential allosteric crosstalk was identified from MixMD simulations of the SARS-CoV-2 main protease using normal mode analysis (DasGupta et al., 2022).

Co-solvent MD simulations often identify sites that cannot be related to any known biological function. Such interaction hotspots could also become helpful in the future. For example, in the proteolysis targeting chimera (PROTAC) technology, a compound with two ligand head groups attached by a flexible linker binds both the target protein and an E3 ubiquitin ligase. The E3 ubiquitin ligase tags the target protein with ubiquitin for degradation by the cellular machinery (Békés et al., 2022). In theory, a PROTAC compound could bind anywhere in the target protein structure if the binding is tight enough for the ubiquitination to occur successfully. Such an approach coupled with detecting novel binding sites by co-solvent MD simulation methods can be envisioned to extend the druggable proteome substantially.

6.2 Both experimental and computational binding data are helpful in fragment- and negative image-based screening

F-NIB, introduced in Study I, is a welcome extension to the already well-established NIB modelling and VS. Utility of the extended method was demonstrated in Study I by the discovery of three novel PDE10A inhibitors that showed structural uniqueness when compared to the previously known inhibitor compounds. Notably, the active compounds were discovered by a straightforward screening with an F-NIB model (Model III) and by a combinatory approach where the top-scored compounds of the F-NIB screen were rescored with an FQSAR model (Models I and IV). Moreover, Models III and I included molecular fragments from crystal structures of inhibitor-bound PDE10A, whereas Model IV contained a fragment extracted from a docked pose of a compound suggested by pharmacophore screening. Despite being the focus of multiple drug development campaigns against schizophrenia and neurodegenerative disorders, no PDE10A inhibitor has yet reached the market. Lack of antipsychotic activity in the clinical setting has led to the termination of several PDE10A targeted clinical trials. Compounds with different pharmacology may be required to retain the efficacy observed in preclinical studies. On the other hand, PDE10A inhibitors have been suggested to provide useful tool compounds for developing other antipsychotic drugs, such as dopamine D2 receptor antagonists (Menniti et al., 2021). Thus, an emphasis was put in Study I on finding novel structural scaffolds for PDE10A inhibition. 2D fingerprint similarity analysis confirmed that the F-NIB-based VS workflows provided structurally diverse solutions for occupying the known inhibitor binding regions of the PDE10A active site.

Study IV combined the F-NIB approach with probe binding data obtained from the MixMD simulations. Conformation-specific pyrimidine binding poses were extracted for the two different forms of the S protein RBD's ACE2-binding interface. Reports of utilizing simulated probe binding data in VS models can be found, but the idea of using probe binding mode information in space-filling VS models, such as NIB models, seems to be rarely applied (Arcon et al., 2017, 2019; Lakkaraju et al., 2015). The methodology can be seen as a form of FBDD. Running MixMD simulations with various probe types highlights interaction sites where compounds with similar structural properties could bind. Naturally, the larger the portion of chemical space covered with the probe molecules, the more computationally demanding the method becomes. Thus, it is reasonable to use probe molecules that identify sites for general aromatic, hydrophobic, hydrogen bonding, and ionic interactions, and possibly vary the size of the used probes slightly. Using larger and more complex fragments could identify more specific high-affinity scaffolds. However, the binding and dissociation of larger fragments with millimolar affinity

are much slower than with the smaller probe molecules. Pan et al. found out that co-solvent MD simulations with molecular fragments had to be extended to a microsecond scale to obtain enough binding events for the fragment affinity values to converge to a standard error within 0.5 kcal/mol (Pan et al., 2017). An alternative approach for discovering larger active fragments could be to screen fragment libraries using protein structures or stably bound probe molecules extracted from co-solvent MD simulations with simpler probes.

The MixMD-based F-NIB models developed in Study IV identified compounds that expressed stable binding to the RBD in complex MD simulations. None of the stable compounds had sufficient clinical data available for consideration of their repurposing to target the S protein – ACE2 interaction. To date, no small molecules targeting the S protein – ACE2 PPI have proceeded to the clinical trials, and no significant successes in repurposing known drugs to inhibition of the complex formation have been reported (Zhao et al., 2022). This could indicate that the magnitude of inhibition of the interaction achievable by small molecules is not strong enough for medicinal purposes. In addition, later variants of the SARS-CoV-2 have shown the S protein to be prone to mutations, which significantly complicates drug development for this target (Liu et al., 2022). Currently, small molecule drugs targeting other pivotal proteins of the SARS-CoV-2 are in medical use.

6.3   CYP active site flexibility should be considered in small molecule binding mode predictions

Studies II and III demonstrated that slight differences in amino acid composition or conformation in an enzyme's active site might affect the binding modes of small molecules and, consequently, their reaction kinetics. MD simulations showed a clear difference in positioning of DCPCC with CYP1A1, CYP1A2, and CYP1B1, explaining the compound's CYP1A2-selective inhibition and fast oxidation by CYP1A1. Residue sidechain rearrangements, caused by Ser122 in CYP1A1 being replaced by Thr124 in CYP1A2, pushed DCPCC closer to the heme iron, which could hamper optimal placement of molecular oxygen needed for the catalysis. Similarly, coumarin's lower rate of metabolism with CYP2A13 than with CYP2A6 was related to a difference in the 7-hydroxylation site orientation caused by the presence of a more bulky sidechain close to the heme group in CYP2A13.

Interestingly, in the low-energy binding modes of DCPCC with both CYP1A1 and CYP1B1, the initial docking pose resided quite far from the heme iron. This likely allowed a slight induced fit effect to optimize the active site conformation for binding of DCPCC as the compound approached its eventual binding region during the simulation. The active site of the CYP1 enzymes is relatively compact, limiting the accessible motions of small molecules during a simulation. Thus, docking

solutions where either the cyclopropane- or dichlorophenyl-group was orientated towards the heme group were not interchangeable within the MD simulations. As the scoring functions of docking programs are still often unreliable in correctly identifying the energetically best poses, comprehensive sampling of initial ligand poses should be done by, for example, starting the simulations from multiple docking solutions (Wan et al., 2020). Emerging techniques, such as supervised MD simulation, may become useful in reducing the risk of wrongly docked and motionally restricted molecules producing error to free energy calculations. In supervised MD simulations, the ligand binding pathway is simulated from the solvated state to the bound state by running the simulation in short intervals and rejecting segments where the ligand has not approached the designated binding site (Sabbadin and Moro, 2014). In addition, a recent study was reported where a PPI inhibitor's binding into a crystal structure-like pose was simulated by classical unbiased MD simulation (Shan et al., 2022).

Information on the active site flexibility was crucial in predicting active binding modes of coumarin derivatives for the CYP2A13-catalyzed 7-hydroxylation. Most significant conformational freedom was observed for Met365, which resides next to the heme group. Met365 could adopt several conformations that may help either stabilize the 7-hydroxylation site optimally for catalysis or leave room for less optimal anchoring of a 3-phenyl-coumarin via its 6-position substituent. In case Met365 is not well stabilized by the binding substrate, active binding modes may occur less frequently, resulting in less efficient metabolism. Changes in active site residue fluctuations induced by substrate binding could affect the populations of specific binding modes and alter the rate of metabolism.

Conformational sampling of the CYP2A13 active site provided by the available crystal structures was inadequate for generating a reliable active binding mode prediction for compound **10**. Generally, docking produced solutions where **10** overlapped significantly with the active site residues. This could indicate that the binding of **10** induces the formation of a sub-pocket or a channel not captured within the current crystal forms of the enzyme. Alternatively, **10** could occupy the channel formed between Gln104 and Phe480 in the PDB structure 4EJI. However, no docking poses consistent with 7-hydroxylation were generated in the docking with 4EJI. The emergence of additional channels in the CYP2A13 active site has been reported in an MD simulation study (Fan et al., 2018a). Compound **10** could occupy some of these channels in its active binding mode.

6.4 Utility of computer-based predictions in drug development

Computer-based methods exist already for substantial facilitation of preclinical drug development. Medicinally interesting target proteins can be identified by analysing genomic data with bioinformatics approaches. Target proteins for which high-resolution structural models are available can be subjected to druggability assessment and target site detection with methods such as co-solvent MD simulations. Massive molecular databases can be screened virtually to find potential active compounds or structural scaffolds if the target site is known. The initial hit molecules can be optimized into potent lead compounds and drug candidates using MD simulations and binding free energy calculations. The pharmacokinetics and metabolism of the potential candidates can be predicted computationally before taking them forward to the exhaustive experimental studies. In practice, however, the effectivity of the computational approaches is still limited by lack of accuracy, especially in the prediction of binding affinity in molecular recognition events. Despite the acknowledged need for better scoring functions, the results of this thesis contribute to the growing body of evidence regarding the feasible utilization of current and the development of novel CADD methods to facilitate preclinical drug development. Publications reporting successes in computer-based binding site detection, VS, and small molecule metabolism prediction leave no doubt that a significant impact can be made on drug development with the currently available CADD approaches.

SBDD is often limited by the low availability of high-resolution structures for specific disease-related proteins. The development of artificial intelligence-based structure prediction methods is already improving this deficiency. Alphafold, which uses experimental databases to predict protein structures based on their amino acid sequences, has shown promising performance (Jumper et al., 2021). The usability of Alphafold for systems with, for example, bound cofactors or multimeric configuration is, however, yet to be seen. As outlined in the literature review of this thesis, machine learning and artificial intelligence are being growingly applied in the development of novel CADD methods and combined with the more traditional methods. Quantum mechanics is widely applied with the classical MD simulations to study a specific part of the simulation system in quantum-level detail (Cui et al., 2021). Usage of quantum mechanics simulations for large systems is practically impossible due to their computational cost that exceeds the performance of even the most powerful supercomputers. The ongoing development of quantum computers can be envisioned to eventually become a game-changer in this regard (Zinner et al., 2021).

7   Summary/Conclusions

Computational methods are having an increasing impact on drug development. Modern high-performance computing allows the elaborate usage of physics-based methods such as MD simulations to provide highly detailed data of the studied systems. In addition, binding affinity prediction for even billions of compounds to a given target protein is achievable with the current VS methods. In this thesis, various CADD approaches, including MD simulations, molecular docking, F-NIB screening, FQSAR, and binding free energy calculations, were shown to provide biologically relevant information in multiple SBDD settings. The utility of these approaches in detecting binding sites for drugs in a novel disease-related protein, structure-based VS of small molecule inhibitors, and studying the structural determinants for small molecule metabolism was demonstrated.

The ultimate goal of developing CADD approaches is to reduce the need for experimental testing in preclinical drug development and improve drug candidates' chances to survive in the exhaustive clinical trials. Undoubtedly, the accuracy of the scoring functions, especially for affinity prediction, still needs improvement to reach higher reliability. Despite the rapid development of computation methods and infrastructure, there still is a significant trade-off between prediction accuracy and computation speed. In this thesis, a miniature model of computational preclinical drug discovery is presented in the form of separate studies. Combining, for example, target site detection and VS still suffers from uncertainties in identifying the druggable conformations of the correct binding sites and the reliability of affinity prediction by scoring functions. Nevertheless, publications are emerging where CADD methods are used to develop novel drug design strategies for identifying target proteins, target sites, and structural scaffolds of active compounds. Significant technical improvements can be expected in the future, especially by developments in machine learning and quantum computing, hopefully shifting the traditional accuracy-speed trade-off of CADD closer to a one-sided bargain.

Acknowledgements

I am deeply grateful to my supervisor, Professor Olli Pentikäinen, for giving me the opportunity to work in interesting projects and see the scientific world from both academic and commercial sides. It has been a pleasure to work in a friendly atmosphere that does not lack competence or freedom to develop and test your own ideas.

I thank Professor Stefano Moro for accepting the invitation to be my opponent, and Professors Philip Biggin and Gerrit Groenhof for the preliminary examination of my thesis. I acknowledge Tomi Airenne for his time and comments in our follow-up group meetings.

I would like to thank all my collaborators, co-authors and co-workers from the last years. Pekka Postila, thank you for your thorough approach in providing me the methodological basis needed to do this work. Risto Juvonen, thank you for the fluent collaboration in our shared publications. I thank the current and former group members Sakari Lätti, Sanna Niinivehmas, Sami Kurkinen, Mira Ahinko, Gopinath Krishnasamy, Pankaj Singh and Kseniia Petrova-Szczasiuk for excellent professional collaborations, but also for good moments in somewhat unprofessional settings.

I also thank the Farmis group, Drug Research Doctoral Program, and my fellow doctoral students within the program. Thank you to Outi Irjala for the very clear and much needed guidance through the final stages of the PhD process.

Computer-based work requires lots of abstract reasoning and, from time to time, the mind must be reset by doing something more physical. I thank the badminton community of the Varsinais-Suomi region, and especially members of the club TVS-Sulka, for providing such an excellent company with whom to detach from science and delve into the secrets of the fastest sport in the world.

Finally, I thank my family for their endless support in both good and bad times. Thank you for teaching me that it is all about attitude.

Turku, October 2022
Elmeri Jokinen

References

Ahinko, M., Kurkinen, S.T., Niinivehmas, S.P., Pentikäinen, O.T., and Postila, P.A. (2019a). A Practical Perspective: The Effect of Ligand Conformers on the Negative Image-Based Screening. Int. J. Mol. Sci. *20*. https://doi.org/10.3390/ijms20112779.

Ahinko, M., Niinivehmas, S., Jokinen, E., and Pentikäinen, O.T. (2019b). Suitability of MMGBSA for the Selection of Correct Ligand Binding Modes from Docking Results. Chem. Biol. Drug Des. *93*, 522–538. https://doi.org/10.1111/cbdd.13446.

Allen, K.N., Bellamacina, C.R., Ding, X., Jeffery, C.J., Mattos, C., Petsko, G.A., and Ringe, D. (1996). An Experimental Approach to Mapping the Binding Surfaces of Crystalline Proteins. J. Phys. Chem. *100*, 2605–2611. https://doi.org/10.1021/jp952516o.

Alvarez-Garcia, D., and Barril, X. (2014). Relationship between Protein Flexibility and Binding: Lessons for Structure-Based Drug Design. J. Chem. Theory Comput. *10*, 2608–2614. https://doi.org/10.1021/ct500182z.

Amaral, M., Kokh, D.B., Bomke, J., Wegener, A., Buchstaller, H.P., Eggenweiler, H.M., Matias, P., Sirrenberg, C., Wade, R.C., and Frech, M. (2017). Protein conformational flexibility modulates kinetics and thermodynamics of drug binding. Nat. Commun. *8*, 2276. https://doi.org/10.1038/s41467-017-02258-w.

Amaro, R.E., Baron, R., and McCammon, J.A. (2008). An improved relaxed complex scheme for receptor flexibility in computer-aided drug design. J. Comput. Aided. Mol. Des. *22*, 693–705. https://doi.org/10.1007/s10822-007-9159-2.

Amaro, R.E., Baudry, J., Chodera, J., Demir, Ö., McCammon, J.A., Miao, Y., and Smith, J.C. (2018). Ensemble Docking in Drug Discovery. Biophys. J. *114*, 2271–2278. https://doi.org/https://doi.org/10.1016/j.bpj.2018.02.038.

Arcon, J.P., Defelipe, L.A., Modenutti, C.P., López, E.D., Alvarez-Garcia, D., Barril, X., Turjanski, A.G., and Martí, M.A. (2017). Molecular Dynamics in Mixed Solvents Reveals Protein–Ligand Interactions, Improves Docking, and Allows Accurate Binding Free Energy Predictions. J. Chem. Inf. Model. *57*, 846–863. https://doi.org/10.1021/acs.jcim.6b00678.

Arcon, J.P., Defelipe, L.A., Lopez, E.D., Burastero, O., Modenutti, C.P., Barril, X., Marti, M.A., and Turjanski, A.G. (2019). Cosolvent-Based Protein Pharmacophore for Ligand Enrichment in Virtual Screening. J. Chem. Inf. Model. *59*, 3572–3583. https://doi.org/10.1021/acs.jcim.9b00371.

Arkin, M.R., Tang, Y., and Wells, J.A. (2014). Small-molecule inhibitors of protein-protein interactions: progressing toward the reality. Chem. Biol. *21*, 1102–1114. https://doi.org/10.1016/j.chembiol.2014.09.001.

Bakan, A., Nevins, N., Lakdawala, A.S., and Bahar, I. (2012). Druggability Assessment of Allosteric Proteins by Dynamics Simulations in the Presence of Probe Molecules. J. Chem. Theory Comput. *8*, 2435–2447. https://doi.org/10.1021/ct300117j.

Bauer, E., Guo, Z., Ueng, Y.F., Bell, L.C., Zeldin, D., and Guengerich, F.P. (1995). Oxidation of benzo[a]pyrene by recombinant human cytochrome P450 enzymes. Chem. Res. Toxicol. *8*, 136–142. https://doi.org/10.1021/tx00043a018.

Bayly, C.I., Cieplak, P., Cornell, W., and Kollman, P.A. (1993). A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. J. Phys. Chem. *97*, 10269–10280. https://doi.org/10.1021/j100142a004.

Békés, M., Langley, D.R., and Crews, C.M. (2022). PROTAC targeted protein degraders: the past is prologue. Nat. Rev. Drug Discov. *21*, 181–200. https://doi.org/10.1038/s41573-021-00371-6.

Bender, B.J., Gahbauer, S., Luttens, A., Lyu, J., Webb, C.M., Stein, R.M., Fink, E.A., Balius, T.E., Carlsson, J., Irwin, J.J., et al. (2021). A practical guide to large-scale docking. Nat. Protoc. *16*, 4799–4832. https://doi.org/10.1038/s41596-021-00597-z.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res. Nucleic Acids Res. *28*, 235–242. https://doi.org/gkd090 [pii].

Best, R.B., Zhu, X., Shim, J., Lopes, P.E.M., Mittal, J., Feig, M., and MacKerell, A.D. (2012). Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ, ψ and Side-Chain $\chi1$ and $\chi2$ Dihedral Angles. J. Chem. Theory Comput. *8*, 3257–3273. https://doi.org/10.1021/ct300400x.

Bhati, A.P., and Coveney, P. V (2022). Large Scale Study of Ligand–Protein Relative Binding Free Energy Calculations: Actionable Predictions from Statistically Robust Protocols. J. Chem. Theory Comput. *18*, 2687–2702. https://doi.org/10.1021/acs.jctc.1c01288.

Bian, Y., and Xie, X.-Q.S. (2018). Computational Fragment-Based Drug Design: Current Trends, Strategies, and Applications. AAPS J. *20*, 59. https://doi.org/10.1208/s12248-018-0216-7.

Bissaro, M., Sturlese, M., and Moro, S. (2020). The rise of molecular simulations in fragment-based drug design (FBDD): an overview. Drug Discov. Today *25*, 1693–1701. https://doi.org/10.1016/j.drudis.2020.06.023.

Bowman, G.R., and Geissler, P.L. (2012). Equilibrium fluctuations of a single folded protein reveal a multitude of potential cryptic allosteric sites. Proc. Natl. Acad. Sci. U. S. A. *109*, 11681–11686. https://doi.org/10.1073/pnas.1209309109.

Boyce, S.E., Mobley, D.L., Rocklin, G.J., Graves, A.P., Dill, K.A., and Shoichet, B.K. (2009). Predicting ligand binding affinity with alchemical free energy methods in a polar model binding site. J. Mol. Biol. *394*, 747–763. https://doi.org/10.1016/j.jmb.2009.09.049.

Buonfiglio, R., Recanatini, M., and Masetti, M. (2015). Protein Flexibility in Drug Discovery: From Theory to Computation. ChemMedChem *10*, 1141–1148. https://doi.org/10.1002/cmdc.201500086.

Carino, A., Moraca, F., Fiorillo, B., Marchianò, S., Sepe, V., Biagioli, M., Finamore, C., Bozza, S., Francisci, D., Distrutti, E., et al. (2020). Hijacking SARS-CoV-2/ACE2 Receptor Interaction by Natural and Semi-synthetic Steroidal Agents Acting on Functional Pockets on the Receptor Binding Domain. Front. Chem. *8*, 846. https://doi.org/10.3389/fchem.2020.572885.

Carlson, H.A., Smith, R.D., Damm-Ganamet, K.L., Stuckey, J.A., Ahmed, A., Convery, M.A., Somers, D.O., Kranz, M., Elkins, P.A., Cui, G., et al. (2016). CSAR 2014: A Benchmark Exercise Using Unpublished Data from Pharma. J. Chem. Inf. Model. *56*, 1063–1077. https://doi.org/10.1021/acs.jcim.5b00523.

Case, D.A., Belfon, K., Ben-Shalom, I.Y., Brozell, S.R., Cerutti, D.S., III, T.E.C., Cruzeiro, V.W.D., Darden, T.A., Duke, R.E., Giambasu, G., et al. (2020). AMBER2020, University of California, San Fransisco.

Chappie, T.A., Helal, C.J., and Hou, X. (2012). Current landscape of phosphodiesterase 10A (PDE10A) inhibition. J. Med. Chem. *55*, 7299–7331. https://doi.org/10.1021/jm3004976.

Chen, Y., and Shoichet, B.K. (2009). Molecular docking and ligand specificity in fragment-based inhibitor discovery. Nat. Chem. Biol. *5*, 358–364. https://doi.org/10.1038/nchembio.155.

Chen, J., Vishweshwaraiah, Y.L., and Dokholyan, N. V (2022). Design and engineering of allosteric communications in proteins. Curr. Opin. Struct. Biol. *73*, 102334. https://doi.org/10.1016/j.sbi.2022.102334.

Chodera, J.D., Mobley, D.L., Shirts, M.R., Dixon, R.W., Branson, K., and Pande, V.S. (2011). Alchemical free energy methods for drug discovery: progress and challenges. Curr. Opin. Struct. Biol. *21*, 150–160. https://doi.org/10.1016/j.sbi.2011.01.011.

Cimermancic, P., Weinkam, P., Rettenmaier, T.J., Bichmann, L., Keedy, D.A., Woldeyes, R.A., Schneidman-Duhovny, D., Demerdash, O.N., Mitchell, J.C., Wells, J.A., et al. (2016). CryptoSite: Expanding the Druggable Proteome by Characterization and Prediction of Cryptic Binding Sites. J. Mol. Biol. *428*, 709–719. https://doi.org/10.1016/j.jmb.2016.01.029.

Claesson, A., and Minidis, A. (2018). Systematic Approach to Organizing Structural Alerts for Reactive Metabolite Formation from Potential Drugs. Chem. Res. Toxicol. *31*, 389–411. https://doi.org/10.1021/acs.chemrestox.8b00046.

Clark, J.J., Orban, Z.J., and Carlson, H.A. (2020). Predicting binding sites from unbound versus bound protein structures. Sci. Rep. *10*, 15856. https://doi.org/10.1038/s41598-020-72906-7.

Comitani, F., and Gervasio, F.L. (2018). Exploring Cryptic Pockets Formation in Targets of Pharmaceutical Interest with SWISH. J. Chem. Theory Comput. *14*, 3321–3331. https://doi.org/10.1021/acs.jctc.8b00263.

Cournia, Z., Allen, B.K., Beuming, T., Pearlman, D.A., Radak, B.K., and Sherman, W. (2020). Rigorous Free Energy Simulations in Virtual Screening. J. Chem. Inf. Model. *60*, 4153–4169. https://doi.org/10.1021/acs.jcim.0c00116.

Craig, I.R., Pfleger, C., Gohlke, H., Essex, J.W., and Spiegel, K. (2011). Pocket-space maps to identify novel binding-site conformations in proteins. J. Chem. Inf. Model. *51*, 2666–2679. https://doi.org/10.1021/ci200168b.

Cramer, R.D., Patterson, D.E., and Bunce, J.D. (1988). Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. J. Am. Chem. Soc. *110*, 5959–5967. https://doi.org/10.1021/ja00226a005.

Cross, J.B., Thompson, D.C., Rai, B.K., Baber, J.C., Fan, K.Y., Hu, Y., and Humblet, C. (2009). Comparison of several molecular docking programs: pose prediction and virtual screening accuracy. J. Chem. Inf. Model. *49*, 1455–1474. .

Cui, Q., Pal, T., and Xie, L. (2021). Biomolecular QM/MM Simulations: What Are Some of the "Burning Issues"? J. Phys. Chem. B *125*, 689–702. https://doi.org/10.1021/acs.jpcb.0c09898.

Darden, T., York, D., and Pedersen, L. (1993). Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. J. Chem. Phys. *98*, 10089–10092. https://doi.org/10.1063/1.464397.

DasGupta, D., Chan, W.K.B., and Carlson, H.A. (2022). Computational Identification of Possible Allosteric Sites and Modulators of the SARS-CoV-2 Main Protease. J. Chem. Inf. Model. *62*, 618–626. https://doi.org/10.1021/acs.jcim.1c01223.

Defelipe, L.A., Arcon, J.P., Modenutti, C.P., Marti, M.A., Turjanski, A.G., and Barril, X. (2018). Solvents to Fragments to Drugs: MD Applications in Drug Design. Molecules *23*. https://doi.org/10.3390/molecules23123269.

DeVore, N.M., and Scott, E.E. (2012). Nicotine and 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone binding and access channel in human cytochrome P450 2A6 and 2A13 enzymes. J. Biol. Chem. *287*, 26576–26585. https://doi.org/10.1074/jbc.M112.372813.

DeVore, N.M., Meneely, K.M., Bart, A.G., Stephens, E.S., Battaile, K.P., and Scott, E.E. (2012). Structural comparison of cytochromes P450 2A6, 2A13, and 2E1 with pilocarpine. FEBS J. *279*, 1621–1631. https://doi.org/10.1111/j.1742-4658.2011.08412.x.

DiMasi, J.A., Grabowski, H.G., and Hansen, R.W. (2016). Innovation in the pharmaceutical industry: New estimates of R&D costs. J. Health Econ. *47*, 20–33. https://doi.org/10.1016/j.jhealeco.2016.01.012.

Dixon, S.L., Smondyrev, A.M., and Rao, S.N. (2006a). PHASE: a novel approach to pharmacophore modeling and 3D database searching. Chem. Biol. Drug Des. *67*, 370–372. https://doi.org/10.1111/j.1747-0285.2006.00384.x.

Dixon, S.L., Smondyrev, A.M., Knoll, E.H., Rao, S.N., Shaw, D.E., and Friesner, R.A. (2006b). PHASE: a new engine for pharmacophore perception, 3D QSAR model development, and 3D

database screening: 1. Methodology and preliminary results. J. Comput. Aided. Mol. Des. *20*, 647–671. https://doi.org/10.1007/s10822-006-9087-6.

Djoumbou-Feunang, Y., Fiamoncini, J., Gil-de-la-Fuente, A., Greiner, R., Manach, C., and Wishart, D.S. (2019). BioTransformer: a comprehensive computational tool for small molecule metabolism prediction and metabolite identification. J. Cheminform. *11*, 2. https://doi.org/10.1186/s13321-018-0324-5.

Durrant, J.D., and McCammon, J.A. (2011). Molecular dynamics simulations and drug discovery. BMC Biol. *9*, 71. https://doi.org/10.1186/1741-7007-9-71.

Ebalunode, J.O., Ouyang, Z., Liang, J., and Zheng, W. (2008). Novel approach to structure-based pharmacophore search using computational geometry and shape matching techniques. J. Chem. Inf. Model. *48*, 889–901. https://doi.org/10.1021/ci700368p.

Eichelbaum, M., Kroemer, H.K., and Mikus, G. (1992). Genetically determined differences in drug metabolism as a risk factor in drug toxicity. Toxicol. Lett. *64-65 Spec*, 115–122. https://doi.org/10.1016/0378-4274(92)90180-r.

Essmann, U., Perera, L., Berkowitz, M.L., Darden, T., Lee, H., and Pedersen, L.G. (1995). A smooth particle mesh Ewald method. J. Chem. Phys. *103*, 8577–8593. https://doi.org/10.1063/1.470117.

Fan, J.-R., Li, H., Zhang, H.-X., and Zheng, Q.-C. (2018a). Exploring the structure characteristics and major channels of cytochrome P450 2A6, 2A13, and 2E1 with pilocarpine. Biopolymers *109*, e23108. https://doi.org/10.1002/bip.23108.

Fan, N., Zhang, S., Sheng, T., Zhao, L., Liu, Z., Liu, J., and Wang, X. (2018b). Docking field-based QSAR and pharmacophore studies on the substituted pyrimidine derivatives targeting HIV-1 reverse transcriptase. Chem. Biol. Drug Des. *91*, 398–407. https://doi.org/10.1111/cbdd.13086.

Ferrari, A.M., Wei, B.Q., Costantino, L., and Shoichet, B.K. (2004). Soft docking and multiple receptor conformations in virtual screening. J. Med. Chem. *47*, 5076–5084. https://doi.org/10.1021/jm049756p.

Friesner, R.A., Banks, J.L., Murphy, R.B., Halgren, T.A., Klicic, J.J., Mainz, D.T., Repasky, M.P., Knoll, E.H., Shelley, M., Perry, J.K., et al. (2004). Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. J. Med. Chem. *47*, 1739–1749. https://doi.org/10.1021/jm0306430.

Fujishige, K., Kotera, J., and Omori, K. (1999). Striatum- and testis-specific phosphodiesterase PDE10A isolation and characterization of a rat PDE10A. Eur. J. Biochem. *266*, 1118–1127. .

Gao, C., Desaphy, J., and Vieth, M. (2017). Are induced fit protein conformational changes caused by ligand-binding predictable? A molecular dynamics investigation. J. Comput. Chem. *38*, 1229–1237. https://doi.org/10.1002/jcc.24714.

Gaulton, A., Hersey, A., Nowotka, M., Bento, A.P., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L.J., Cibrián-Uhalte, E., et al. (2017). The ChEMBL database in 2017. Nucleic Acids Res. *45*, D945–D954. https://doi.org/10.1093/nar/gkw1074.

Genheden, S., and Ryde, U. (2015). The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. Expert Opin. Drug Discov. *10*, 449–461. https://doi.org/10.1517/17460441.2015.1032936.

Ghanakota, P., and Carlson, H.A. (2016a). Driving Structure-Based Drug Discovery through Cosolvent Molecular Dynamics. J. Med. Chem. *59*, 10383–10399. https://doi.org/10.1021/acs.jmedchem.6b00399.

Ghanakota, P., and Carlson, H.A. (2016b). Moving Beyond Active-Site Detection: MixMD Applied to Allosteric Systems. J. Phys. Chem. B *120*, 8685–8695. https://doi.org/10.1021/acs.jpcb.6b03515.

Ghanakota, P., van Vlijmen, H., Sherman, W., and Beuming, T. (2018). Large-Scale Validation of Mixed-Solvent Simulations to Assess Hotspots at Protein-Protein Interaction Interfaces. J. Chem. Inf. Model. *58*, 784–793. https://doi.org/10.1021/acs.jcim.7b00487.

Goodford, P.J. (1985). A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. J. Med. Chem. *28*, 849–857. https://doi.org/10.1021/jm00145a002.

Gopinath, K., Jokinen, E.M., Kurkinen, S.T., and Pentikäinen, O.T. (2020). Screening of Natural Products Targeting SARS-CoV-2–ACE2 Receptor Interface – A MixMD Based HTVS Pipeline. Front. Chem. *8*, 1084. https://doi.org/10.3389/fchem.2020.589769.

Gorgulla, C., Boeszoermenyi, A., Wang, Z.-F., Fischer, P.D., Coote, P.W., Padmanabha Das, K.M., Malets, Y.S., Radchenko, D.S., Moroz, Y.S., Scott, D.A., et al. (2020). An open-source drug discovery platform enables ultra-large virtual screens. Nature *580*, 663–668. https://doi.org/10.1038/s41586-020-2117-z.

Götz, A.W., Williamson, M.J., Xu, D., Poole, D., Le Grand, S., and Walker, R.C. (2012). Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. J. Chem. Theory Comput. *8*, 1542–1555. https://doi.org/10.1021/ct200909j.

Graham, S.E., Leja, N., and Carlson, H.A. (2018). MixMD Probeview: Robust Binding Site Prediction from Cosolvent Simulations. J. Chem. Inf. Model. *58*, 1426–1433. https://doi.org/10.1021/acs.jcim.8b00265.

Grebner, C., Malmerberg, E., Shewmaker, A., Batista, J., Nicholls, A., and Sadowski, J. (2020). Virtual Screening in the Cloud: How Big Is Big Enough? J. Chem. Inf. Model. *60*, 4274–4282. https://doi.org/10.1021/acs.jcim.9b00779.

Greenwood, J.R., Calkins, D., Sullivan, A.P., and Shelley, J.C. (2010). Towards the comprehensive, rapid, and accurate prediction of the favorable tautomeric states of drug-like molecules in aqueous solution. J. Comput. Aided. Mol. Des. *24*, 591–604. https://doi.org/10.1007/s10822-010-9349-1.

Guedes, I.A., Pereira, F.S.S., and Dardenne, L.E. (2018). Empirical Scoring Functions for Structure-Based Virtual Screening: Applications, Critical Aspects, and Challenges. Front. Pharmacol. *9*. https://doi.org/10.3389/fphar.2018.01089.

Guengerich, F.P. (2017). Intersection of the Roles of Cytochrome P450 Enzymes with Xenobiotic and Endogenous Substrates: Relevance to Toxicity and Drug Interactions. Chem. Res. Toxicol. *30*, 2–12. https://doi.org/10.1021/acs.chemrestox.6b00226.

Guengerich, F.P., Wilkey, C.J., and Phan, T.T.N. (2019). Human cytochrome P450 enzymes bind drugs and other substrates mainly through conformational-selection modes. J. Biol. Chem. *294*, 10928–10941. https://doi.org/https://doi.org/10.1074/jbc.RA119.009305.

Le Guilloux, V., Schmidtke, P., and Tuffery, P. (2009). Fpocket: an open source platform for ligand pocket detection. BMC Bioinformatics *10*, 168. https://doi.org/10.1186/1471-2105-10-168.

Guvench, O., and MacKerell, A.D.J. (2009). Computational fragment-based binding site identification by ligand competitive saturation. PLoS Comput. Biol. *5*, e1000435. https://doi.org/10.1371/journal.pcbi.1000435.

Halgren, T. (2007). New method for fast and accurate binding-site identification and analysis. Chem. Biol. Drug Des. *69*, 146–148. https://doi.org/10.1111/j.1747-0285.2007.00483.x.

Halgren, T.A. (2009). Identifying and characterizing binding sites and assessing druggability. J. Chem. Inf. Model. *49*, 377–389. https://doi.org/10.1021/ci800324m.

Hamelberg, D., Mongan, J., and McCammon, J.A. (2004). Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. J. Chem. Phys. *120*, 11919–11929. https://doi.org/10.1063/1.1755656.

Harder, E., Damm, W., Maple, J., Wu, C., Reboul, M., Xiang, J.Y., Wang, L., Lupyan, D., Dahlgren, M.K., Knight, J.L., et al. (2016). OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. J. Chem. Theory Comput. *12*, 281–296. https://doi.org/10.1021/acs.jctc.5b00864.

Heifetz, A., James, T., Morao, I., Bodkin, M.J., and Biggin, P.C. (2016). Guiding lead optimization with GPCR structure modeling and molecular dynamics. Curr. Opin. Pharmacol. *30*, 14–21. https://doi.org/10.1016/j.coph.2016.06.004.

Hou, T., Wang, J., Li, Y., and Wang, W. (2011). Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations. J. Chem. Inf. Model. *51*, 69–82. https://doi.org/10.1021/ci100275a.

Hritz, J., de Ruiter, A., and Oostenbrink, C. (2008). Impact of plasticity and flexibility on docking results for cytochrome P450 2D6: a combined approach of molecular dynamics and ligand docking. J. Med. Chem. *51*, 7469–7477. https://doi.org/10.1021/jm801005m.

Huang, S.-Y., and Zou, X. (2006). An iterative knowledge-based scoring function to predict protein-ligand interactions: I. Derivation of interaction potentials. J. Comput. Chem. *27*, 1866–1875. https://doi.org/10.1002/jcc.20504.

Huggins, D.J., Sherman, W., and Tidor, B. (2012). Rational Approaches to Improving Selectivity in Drug Design. J. Med. Chem. *55*, 1424–1444. https://doi.org/10.1021/jm2010332.

Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: visual molecular dynamics. J. Mol. Graph. *14*, 27-28,33-38. https://doi.org/10.1016/0263-7855(96)00018-5.

Iida, S., Nakamura, H.K., Mashimo, T., and Fukunishi, Y. (2020). Structural Fluctuations of Aromatic Residues in an Apo-Form Reveal Cryptic Binding Sites: Implications for Fragment-Based Drug Design. J. Phys. Chem. B *124*, 9977–9986. https://doi.org/10.1021/acs.jpcb.0c04963.

Isin, E.M., and Guengerich, F.P. (2008). Substrate binding to cytochromes P450. Anal. Bioanal. Chem. *392*, 1019–1030. https://doi.org/10.1007/s00216-008-2244-0.

Jacobson, M.P., Friesner, R.A., Xiang, Z., and Honig, B. (2002). On the role of the crystal environment in determining protein side-chain conformations. J. Mol. Biol. *320*, 597–608. https://doi.org/10.1016/S0022-2836(02)00470-9.

Jakalian, A., Bush, B.L., Jack, D.B., and Bayly, C.I. (2000). Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method. J. Comput. Chem. *21*, 132–146. https://doi.org/https://doi.org/10.1002/(SICI)1096-987X(20000130)21:2<132::AID-JCC5>3.0.CO;2-P.

Jakalian, A., Jack, D.B., and Bayly, C.I. (2002). Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. J. Comput. Chem. *23*, 1623–1641. https://doi.org/10.1002/jcc.10128.

Jiang, Z., and Zhou, Y. (2005). Using bioinformatics for drug target identification from the genome. Am. J. Pharmacogenomics Genomics-Related Res. Drug Dev. Clin. Pract. *5*, 387–396. https://doi.org/10.2165/00129785-200505060-00005.

Jiménez, J., Doerr, S., Martínez-Rosell, G., Rose, A.S., and De Fabritiis, G. (2017). DeepSite: protein-binding site predictor using 3D-convolutional neural networks. Bioinformatics *33*, 3036–3042. https://doi.org/10.1093/bioinformatics/btx350.

Johnson, D.K., and Karanicolas, J. (2015). Selectivity by small-molecule inhibitors of protein interactions can be driven by protein surface fluctuations. PLoS Comput. Biol. *11*, e1004081. https://doi.org/10.1371/journal.pcbi.1004081.

Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W., and Klein, M.L. (1983). Comparison of simple potential functions for simulating liquid water. J. Chem. Phys. *79*, 926–935. https://doi.org/10.1063/1.445869.

Ju, B., Zhang, Q., Ge, J., Wang, R., Sun, J., Ge, X., Yu, J., Shan, S., Zhou, B., Song, S., et al. (2020). Human neutralizing antibodies elicited by SARS-CoV-2 infection. Nature https://doi.org/10.1038/s41586-020-2380-z.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2.

Juvonen, R.O., Kuusisto, M., Fohrgrup, C., Pitkänen, M.H., Nevalainen, T.J., Auriola, S., Raunio, H., Pasanen, M., and Pentikäinen, O.T. (2016). Inhibitory effects and oxidation of 6-methylcoumarin, 7-methylcoumarin and 7-formylcoumarin via human CYP2A6 and its mouse and pig orthologous enzymes. Xenobiotica *46*, 14–24. https://doi.org/10.3109/00498254.2015.1048327.

Juvonen, R.O., Ahinko, M., Huuskonen, J., Raunio, H., and Pentikäinen, O.T. (2019). Development of new Coumarin-based profluorescent substrates for human cytochrome P450 enzymes. Xenobiotica. *49*, 1015–1024. https://doi.org/10.1080/00498254.2018.1530399.

Juvonen, R.O., Ahinko, M., Jokinen, E.M., Huuskonen, J., Raunio, H., and Pentikäinen, O.T. (2021). Substrate Selectivity of Coumarin Derivatives by Human CYP1 Enzymes: In Vitro Enzyme Kinetics and In Silico Modeling. ACS Omega *6*, 11286–11296. https://doi.org/10.1021/acsomega.1c00123.

Kalenkiewicz, A., Grant, B.J., and Yang, C.-Y. (2015). Enrichment of druggable conformations from apo protein structures using cosolvent-accelerated molecular dynamics. Biology (Basel). *4*, 344–366. https://doi.org/10.3390/biology4020344.

Kaus, J.W., Harder, E., Lin, T., Abel, R., McCammon, J.A., and Wang, L. (2015). How To Deal with Multiple Binding Poses in Alchemical Relative Protein–Ligand Binding Free Energy Calculations. J. Chem. Theory Comput. *11*, 2670–2679. https://doi.org/10.1021/acs.jctc.5b00214.

Ke, Y.-Y., Shiao, H.-Y., Hsu, Y.C., Chu, C.-Y., Wang, W.-C., Lee, Y.-C., Lin, W.-H., Chen, C.-H., Hsu, J.T.A., Chang, C.-W., et al. (2013). 3D-QSAR-assisted drug design: identification of a potent quinazoline-based Aurora kinase inhibitor. ChemMedChem *8*, 136–148. https://doi.org/10.1002/cmdc.201200464.

Kimura, S.R., Hu, H.P., Ruvinsky, A.M., Sherman, W., and Favia, A.D. (2017). Deciphering Cryptic Binding Sites on Proteins by Mixed-Solvent Molecular Dynamics. J. Chem. Inf. Model. *57*, 1388–1401. https://doi.org/10.1021/acs.jcim.6b00623.

Kirchheiner, J., Schmidt, H., Tzvetkov, M., Keulen, J.-T., Lötsch, J., Roots, I., and Brockmöller, J. (2007). Pharmacokinetics of codeine and its metabolite morphine in ultra-rapid metabolizers due to CYP2D6 duplication. Pharmacogenomics J. *7*, 257–265. https://doi.org/10.1038/sj.tpj.6500406.

Kirchmair, J., Williamson, M.J., Tyzack, J.D., Tan, L., Bond, P.J., Bender, A., and Glen, R.C. (2012). Computational Prediction of Metabolism: Sites, Products, SAR, P450 Enzyme Dynamics, and Mechanisms. J. Chem. Inf. Model. *52*, 617–648. https://doi.org/10.1021/ci200542m.

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

Klebe, G., Abraham, U., and Mietzner, T. (1994). Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. J. Med. Chem. *37*, 4130–4146. .

Kollman, P.A., Massova, I., Reyes, C., Kuhn, B., Huo, S., Chong, L., Lee, M., Lee, T., Duan, Y., Wang, W., et al. (2000). Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. Acc. Chem. Res. *33*, 889–897. https://doi.org/10.1021/ar000033j.

Korb, O., Stützle, T., and Exner, T.E. (2007). An ant colony optimization approach to flexible protein–ligand docking. Swarm Intell. *1*, 115–134. https://doi.org/10.1007/s11721-007-0006-9.

Korb, O., Stützle, T., and Exner, T.E. (2009). Empirical scoring functions for advanced protein-ligand docking with PLANTS. J. Chem. Inf. Model. *49*, 84–96. https://doi.org/10.1021/ci800298z.

Kraulis, P.J. (1991). MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. J. Appl. Crystallogr. *24*, 946–950. https://doi.org/https://doi.org/10.1107/S0021889891004399.

Kurkinen, S.T., Niinivehmas, S., Ahinko, M., Lätti, S., Pentikäinen, O.T., and Postila, P.A. (2018). Improving Docking Performance Using Negative Image-Based Rescoring. Front. Pharmacol. *9*, 1–15. https://doi.org/10.3389/fphar.2018.00260.

Kurkinen, S.T., Lätti, S., Pentikäinen, O.T., and Postila, P.A. (2019). Getting Docking into Shape Using Negative Image-Based Rescoring. J. Chem. Inf. Model. *59*, 3584–3599. https://doi.org/10.1021/acs.jcim.9b00383.

Kurkinen, S.T., Lehtonen, J. V, Pentikäinen, O.T., and Postila, P.A. (2022). Optimization of Cavity-Based Negative Images to Boost Docking Enrichment in Virtual Screening. J. Chem. Inf. Model. *62*, 1100–1112. https://doi.org/10.1021/acs.jcim.1c01145.

Lakkaraju, S.K., Yu, W., Raman, E.P., Hershfeld, A. V, Fang, L., Deshpande, D.A., and MacKerell, A.D.J. (2015). Mapping functional group free energy patterns at protein occluded sites: nuclear

receptors and G-protein coupled receptors. J. Chem. Inf. Model. *55*, 700–708. https://doi.org/10.1021/ci500729k.

Lan, J., Ge, J., Yu, J., Shan, S., Zhou, H., Fan, S., Zhang, Q., Shi, X., Wang, Q., Zhang, L., et al. (2020). Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. Nature *581*, 215–220. https://doi.org/10.1038/s41586-020-2180-5.

Lätti, S.T., Niinivehmas, S., and Pentikäinen, O.T. (2022). Sdfconf: A Novel, Flexible, and Robust Molecular Data Management Tool. J. Chem. Inf. Model. *62*, 9–15. https://doi.org/10.1021/acs.jcim.1c01051.

Lehtonen, J. V, Still, D.-J., Rantanen, V.-V., Ekholm, J., Björklund, D., Iftikhar, Z., Huhtala, M., Repo, S., Jussila, A., Jaakkola, J., et al. (2004). BODIL: a molecular modeling environment for structure-function analysis and drug design. J. Comput. Aided. Mol. Des. *18*, 401–419. https://doi.org/10.1007/s10822-004-3752-4.

Lexa, K.W., and Carlson, H.A. (2011). Full protein flexibility is essential for proper hot-spot mapping. J. Am. Chem. Soc. *133*, 200–202. https://doi.org/10.1021/ja1079332.

Lexa, K.W., and Carlson, H.A. (2012). Protein flexibility in docking and surface mapping. Q. Rev. Biophys. *45*, 301–343. https://doi.org/10.1017/S0033583512000066.

Lexa, K.W., Goh, G.B., and Carlson, H.A. (2014). Parameter Choice Matters: Validating Probe Parameters for Use in Mixed-Solvent Simulations. J. Chem. Inf. Model. *54*, 2190–2199. https://doi.org/10.1021/ci400741u.

Li, J., Abel, R., Zhu, K., Cao, Y., Zhao, S., and Friesner, R.A. (2011). The VSGB 2.0 model: A next generation energy model for high resolution protein structure modeling. Proteins Struct. Funct. Bioinforma. *79*, 2794–2812. https://doi.org/10.1002/prot.23106.

Li, S., Fan, J., Peng, C., Chang, Y., Guo, L., Hou, J., Huang, M., Wu, B., Zheng, J., Lin, L., et al. (2017). New molecular insights into the tyrosyl-tRNA synthase inhibitors: CoMFA, CoMSIA analyses and molecular docking studies. Sci. Rep. *7*, 11525. https://doi.org/10.1038/s41598-017-10618-1.

Liu, J., and Wang, R. (2015). Classification of Current Scoring Functions. J. Chem. Inf. Model. *55*, 475–482. https://doi.org/10.1021/ci500731a.

Liu, H., Wei, P., Kappler, J.W., Marrack, P., and Zhang, G. (2022). SARS-CoV-2 Variants of Concern and Variants of Interest Receptor Binding Domain Mutations and Virus Infectivity. Front. Immunol. *13*, 825256. https://doi.org/10.3389/fimmu.2022.825256.

Lolicato, F., Juhola, H., Zak, A., Postila, P.A., Saukko, A., Rissanen, S., Enkavi, G., Vattulainen, I., Kepczynski, M., and Róg, T. (2020). Membrane-Dependent Binding and Entry Mechanism of Dopamine into Its Receptor. ACS Chem. Neurosci. *11*, 1914–1924. https://doi.org/10.1021/acschemneuro.9b00656.

Lorca, M., Morales-Verdejo, C., Vásquez-Velásquez, D., Andrades-Lagos, J., Campanini-Salinas, J., Soto-Delgado, J., Recabarren-Gajardo, G., and Mella, J. (2018). Structure-Activity Relationships Based on 3D-QSAR CoMFA/CoMSIA and Design of Aryloxypropanol-Amine Agonists with Selectivity for the Human β3-Adrenergic Receptor and Anti-Obesity and Anti-Diabetic Profiles. Molecules *23*. https://doi.org/10.3390/molecules23051191.

Lounnas, V., Ritschel, T., Kelder, J., McGuire, R., Bywater, R.P., and Foloppe, N. (2013). Current progress in Structure-Based Rational Drug Design marks a new mindset in drug discovery. Comput. Struct. Biotechnol. J. *5*, e201302011. https://doi.org/10.5936/csbj.201302011.

Lu, H., Zhou, Q., He, J., Jiang, Z., Peng, C., Tong, R., and Shi, J. (2020). Recent advances in the development of protein–protein interactions modulators: mechanisms and clinical trials. Signal Transduct. Target. Ther. *5*, 213. https://doi.org/10.1038/s41392-020-00315-3.

Maier, J.A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K.E., and Simmerling, C. (2015). ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. J. Chem. Theory Comput. *11*, 3696–3713. https://doi.org/10.1021/acs.jctc.5b00255.

Makley, L.N., Johnson, O.T., Ghanakota, P., Rauch, J.N., Osborn, D., Wu, T.S., Cierpicki, T., Carlson, H.A., and Gestwicki, J.E. (2021). Chemical validation of a druggable site on Hsp27/HSPB1 using

in silico solvent   mapping and biophysical methods. Bioorg. Med. Chem. *34*, 115990. https://doi.org/10.1016/j.bmc.2020.115990.

Malamas, M., Ni, Y., and Erdei, J. (2011). Highly potent, selective, and orally active phosphodiesterase 10A inhibitors. J. Med. Chem. *54*, 7621–7638. .

Manzoni, F., and Ryde, U. (2018). Assessing the stability of free-energy perturbation calculations by performing variations in the method. J. Comput. Aided. Mol. Des. *32*, 529–536. https://doi.org/10.1007/s10822-018-0110-5.

Meng, E.C., Shoichet, B.K., and Kuntz, I.D. (1992). Automated docking with grid-based energy evaluation. J. Comput. Chem. *13*, 505–524. https://doi.org/https://doi.org/10.1002/jcc.540130412.

Menniti, F.S., Chappie, T.A., and Schmidt, C.J. (2021). PDE10A Inhibitors—Clinical Failure or Window Into Antipsychotic Drug Action? Front. Neurosci. *14*. https://doi.org/10.3389/fnins.2020.600178.

Merritt, E.A., and Murphy, M.E. (1994). Raster3D Version 2.0. A program for photorealistic molecular graphics. Acta Crystallogr. D. Biol. Crystallogr. *50*, 869–873. https://doi.org/10.1107/S0907444994006396.

Miller, B.R., McGee, T.D., Swails, J.M., Homeyer, N., Gohlke, H., and Roitberg, A.E. (2012). MMPBSA.py: An Efficient Program for End-State Free Energy Calculations. J. Chem. Theory Comput. *8*, 3314–3321. https://doi.org/10.1021/ct300418h.

Miller, E.B., Murphy, R.B., Sindhikara, D., Borrelli, K.W., Grisewood, M.J., Ranalli, F., Dixon, S.L., Jerome, S., Boyles, N.A., Day, T., et al. (2021). Reliable and Accurate Solution to the Induced Fit Docking Problem for Protein-Ligand Binding. J. Chem. Theory Comput. *17*, 2630–2639. https://doi.org/10.1021/acs.jctc.1c00136.

Mobley, D.L., and Dill, K.A. (2009). Binding of small-molecule ligands to proteins: "what you see" is not always "what you get". Structure *17*, 489–498. https://doi.org/10.1016/j.str.2009.02.010.

Mobley, D.L., Chodera, J.D., and Dill, K.A. (2007). The Confine-and-Release Method: Obtaining Correct Binding Free Energies in the Presence of Protein Conformational Change. J. Chem. Theory Comput. *3*, 1231–1235. https://doi.org/10.1021/ct700032n.

Morris, G.M., Huey, R., Lindstrom, W., Sanner, M.F., Belew, R.K., Goodsell, D.S., and Olson, A.J. (2009). AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. J. Comput. Chem. *30*, 2785–2791. https://doi.org/10.1002/jcc.21256.

Murphy, R.B., Repasky, M.P., Greenwood, J.R., Tubert-Brohman, I., Jerome, S., Annabhimoju, R., Boyles, N.A., Schmitz, C.D., Abel, R., Farid, R., et al. (2016). WScore: A Flexible and Accurate Treatment of Explicit Water Molecules in Ligand-Receptor Docking. J. Med. Chem. *59*, 4364–4384. https://doi.org/10.1021/acs.jmedchem.6b00131.

Nair, P.C., McKinnon, R.A., and Miners, J.O. (2019). Computational Prediction of the Site(s) of Metabolism and Binding Modes of Protein Kinase Inhibitors Metabolized by CYP3A4. Drug Metab. Dispos. *47*, 616–631. https://doi.org/10.1124/dmd.118.085167.

Neuvonen, P.J., Niemi, M., and Backman, J.T. (2006). Drug interactions with lipid-lowering drugs: mechanisms and clinical relevance. Clin. Pharmacol. Ther. *80*, 565–581. https://doi.org/10.1016/j.clpt.2006.09.003.

Ngan, C.-H., Hall, D.R., Zerbe, B., Grove, L.E., Kozakov, D., and Vajda, S. (2012). FTSite: high accuracy detection of ligand binding sites on unbound protein structures. Bioinformatics *28*, 286–287. https://doi.org/10.1093/bioinformatics/btr651.

Nicholls, A., McGaughey, G.B., Sheridan, R.P., Good, A.C., Warren, G., Mathieu, M., Muchmore, S.W., Brown, S.P., Grant, J.A., Haigh, J.A., et al. (2010). Molecular shape and medicinal chemistry: a perspective. J. Med. Chem. *53*, 3862–3886. https://doi.org/10.1021/jm900818s.

Niinivehmas, S.P., Virtanen, S.I., Lehtonen, J. V., Postila, P.A., and Pentikäinen, O.T. (2011). Comparison of virtual high-throughput screening methods for the identification of phosphodiesterase-5 inhibitors. J. Chem. Inf. Model. *51*, 1353–1363. https://doi.org/10.1021/ci1004527.

Niinivehmas, S.P., Salokas, K., Lätti, S., Raunio, H., and Pentikäinen, O.T. (2015). Ultrafast protein structure-based virtual screening with Panther. J. Comput. Aided. Mol. Des. *29*, 989–1006. https://doi.org/10.1007/s10822-015-9870-3.

Niinivehmas, S.P., Manivannan, E., Rauhamäki, S., Huuskonen, J., and Pentikäinen, O.T. (2016). Identification of estrogen receptor a ligands with virtual screening techniques. J. Mol. Graph. Model. *64*, 30–39. https://doi.org/10.1016/j.jmgm.2015.12.006.

O'Connor, S., Le Bihan, Y.-V., Westwood, I.M., Liu, M., Mak, O.W., Zazeri, G., Povinelli, A.P.R., Jones, A.M., van Montfort, R., Reynisson, J., et al. (2022). Discovery and Characterization of a Cryptic Secondary Binding Site in the Molecular Chaperone HSP70. Molecules *27*. https://doi.org/10.3390/molecules27030817.

Oleinikovas, V., Saladino, G., Cossins, B.P., and Gervasio, F.L. (2016). Understanding Cryptic Pocket Formation in Protein Targets by Enhanced Sampling Simulations. J. Am. Chem. Soc. *138*, 14257–14263. https://doi.org/10.1021/jacs.6b05425.

Onufriev, A., Bashford, D., and Case, D.A. (2004). Exploring protein native states and large-scale conformational changes with a modified generalized born model. Proteins *55*, 383–394. https://doi.org/10.1002/prot.20033.

Ortiz de Montellano, P.R. (2013). Cytochrome P450-activated prodrugs. Future Med. Chem. *5*, 213–228. https://doi.org/10.4155/fmc.12.197.

Palacio-Rodríguez, K., Lans, I., Cavasotto, C.N., and Cossio, P. (2019). Exponential consensus ranking improves the outcome in docking and receptor ensemble docking. Sci. Rep. *9*, 5142. https://doi.org/10.1038/s41598-019-41594-3.

Pan, A.C., Xu, H., Palpant, T., and Shaw, D.E. (2017). Quantitative Characterization of the Binding and Unbinding of Millimolar Drug Fragments with Molecular Dynamics Simulations. J. Chem. Theory Comput. *13*, 3372–3377. https://doi.org/10.1021/acs.jctc.7b00172.

Panneerselvam, S., Yesudhas, D., Durai, P., Anwar, M.A., Gosu, V., and Choi, S. (2015). A Combined Molecular Docking/Dynamics Approach to Probe the Binding Mode of Cancer Drugs with Cytochrome P450 3A4. Molecules *20*, 14915–14935. https://doi.org/10.3390/molecules200814915.

Parks, C.D., Gaieb, Z., Chiu, M., Yang, H., Shao, C., Walters, W.P., Jansen, J.M., McGaughey, G., Lewis, R.A., Bembenek, S.D., et al. (2020). D3R grand challenge 4: blind prediction of protein-ligand poses, affinity rankings, and relative binding free energies. J. Comput. Aided. Mol. Des. *34*, 99–119. https://doi.org/10.1007/s10822-020-00289-y.

Phillips, J.C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R.D., Kale, L., and Schulten, K. (2005). Scalable molecular dynamics with NAMD. J. Comput. Chem. *26*, 1781–1802. https://doi.org/10.1002/jcc.20289.

Pinzi, L., and Rastelli, G. (2019). Molecular Docking: Shifting Paradigms in Drug Discovery. Int. J. Mol. Sci. *20*. https://doi.org/10.3390/ijms20184331.

Postila, P.A., Swanson, G.T., and Pentikäinen, O.T. (2010). Exploring kainate receptor pharmacology using molecular dynamics simulations. Neuropharmacology *58*, 515–527. https://doi.org/10.1016/j.neuropharm.2009.08.019.

Prakash, P., Hancock, J.F., and Gorfe, A.A. (2015). Binding hotspots on K-ras: consensus ligand binding sites and other reactive regions from probe-based molecular dynamics analysis. Proteins *83*, 898–909. https://doi.org/10.1002/prot.24786.

Pu, C., Yan, G., Shi, J., and Li, R. (2017). Assessing the performance of docking scoring function, FEP, MM-GBSA, and QM/MM-GBSA approaches on a series of PLK1 inhibitors. Medchemcomm *8*, 1452–1458. https://doi.org/10.1039/c7md00184c.

Puertas-Martín, S., Banegas-Luna, A.J., Paredes-Ramos, M., Redondo, J.L., Ortigosa, P.M., Brovarets', O.O., and Pérez-Sánchez, H. (2020). Is high performance computing a requirement for novel drug discovery and how will this impact academic efforts? Expert Opin. Drug Discov. *15*, 981–986. https://doi.org/10.1080/17460441.2020.1758664.

Raman, E.P., Vanommeslaeghe, K., and Mackerell, A.D.J. (2012). Site-Specific Fragment Identification Guided by Single-Step Free Energy Perturbation Calculations. J. Chem. Theory Comput. *8*, 3513–3525. https://doi.org/10.1021/ct300088r.

Rauhamäki, S., Postila, P.A., Lätti, S., Niinivehmas, S., Multamäki, E., Liedl, K.R., and Pentikäinen, O.T. (2018). Discovery of Retinoic Acid-Related Orphan Receptor γt Inverse Agonists via Docking and Negative Image-Based Screening. ACS Omega *3*, 6259–6266. https://doi.org/10.1021/acsomega.8b00603.

Raunio, H., Juvonen, R.O., Poso, A., Lahtela-Kakkonen, M., and Rahnasto-Rilla, M. (2016). Common and Distinct Interactions of Chemical Inhibitors with Cytochrome P450 CYP1A2, CYP2A6 and CYP2B6 Enzymes. Drug Metab. Lett. *10*, 56–64. https://doi.org/10.2174/1872312810666151204002456.

Raunio, H., Pentikäinen, O., and Juvonen, R.O. (2020). Coumarin-Based Profluorescent and Fluorescent Substrates for Determining Xenobiotic-Metabolizing Enzyme Activities In Vitro. Int. J. Mol. Sci. *21*. https://doi.org/10.3390/ijms21134708.

Rendic, S., and Guengerich, F.P. (2015). Survey of Human Oxidoreductases and Cytochrome P450 Enzymes Involved in the Metabolism of Xenobiotic and Natural Chemicals. Chem. Res. Toxicol. *28*, 38–42. https://doi.org/10.1021/tx500444e.

Robertson, M.J., Tirado-Rives, J., and Jorgensen, W.L. (2015). Improved Peptide and Protein Torsional Energetics with the OPLS-AA Force Field. J. Chem. Theory Comput. *11*, 3499–3509. https://doi.org/10.1021/acs.jctc.5b00356.

Roe, D.R., and Cheatham, T.E. 3rd (2013). PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. J. Chem. Theory Comput. *9*, 3084–3095. https://doi.org/10.1021/ct400341p.

Rush, T.S. 3rd, Grant, J.A., Mosyak, L., and Nicholls, A. (2005). A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. J. Med. Chem. *48*, 1489–1495. https://doi.org/10.1021/jm040163o.

Ryckaert, J.-P., Ciccotti, G., and Berendsen, H.J.C. (1977). Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. J. Comput. Phys. *23*, 327–341. https://doi.org/https://doi.org/10.1016/0021-9991(77)90098-5.

Rydberg, P., Gloriam, D.E., Zaretzki, J., Breneman, C., and Olsen, L. (2010). SMARTCyp: A 2D Method for Prediction of Cytochrome P450-Mediated Drug Metabolism. ACS Med. Chem. Lett. *1*, 96–100. https://doi.org/10.1021/ml100016x.

Rzasa, R.M., Hu, E., Rumfelt, S., Chen, N., Andrews, K.L., Chmait, S., Falsey, J.R., Zhong, W., Jones, A.D., Porter, A., et al. (2012). Discovery of selective biaryl ethers as PDE10A inhibitors: Improvement in potency and mitigation of Pgp-mediated efflux. Bioorganic Med. Chem. Lett. *22*, 7371–7375. https://doi.org/10.1016/j.bmcl.2012.10.078.

Rzasa, R.M., Frohn, M.J., Andrews, K.L., Chmait, S., Chen, N., Clarine, J.G., Davis, C., Eastwood, H.A., Horne, D.B., Hu, E., et al. (2014). Synthesis and preliminary biological evaluation of potent and selective 2-(3-alkoxy-1-azetidinyl) quinolines as novel PDE10A inhibitors with improved solubility. Bioorganic Med. Chem. *22*, 6570–6585. https://doi.org/10.1016/j.bmc.2014.10.013.

Sabbadin, D., and Moro, S. (2014). Supervised Molecular Dynamics (SuMD) as a Helpful Tool To Depict GPCR–Ligand Recognition Pathway in a Nanosecond Time Scale. J. Chem. Inf. Model. *54*, 372–376. https://doi.org/10.1021/ci400766b.

Sabe, V.T., Ntombela, T., Jhamba, L.A., Maguire, G.E.M., Govender, T., Naicker, T., and Kruger, H.G. (2021). Current trends in computer aided drug design and a highlight of drugs discovered via computational techniques: A review. Eur. J. Med. Chem. *224*, 113705. https://doi.org/10.1016/j.ejmech.2021.113705.

La Sala, G., Decherchi, S., De Vivo, M., and Rocchia, W. (2017). Allosteric Communication Networks in Proteins Revealed through Pocket Crosstalk Analysis. ACS Cent. Sci. *3*, 949–960. https://doi.org/10.1021/acscentsci.7b00211.

Salmaso, V., and Moro, S. (2018). Bridging Molecular Docking to Molecular Dynamics in Exploring Ligand-Protein Recognition Process: An Overview. Front. Pharmacol. *9*, 923. https://doi.org/10.3389/fphar.2018.00923.

Salomon-Ferrer, R., Götz, A.W., Poole, D., Le Grand, S., and Walker, R.C. (2013). Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. J. Chem. Theory Comput. *9*, 3878–3888. https://doi.org/10.1021/ct400314y.

Sansen, S., Yano, J.K., Reynald, R.L., Schoch, G.A., Griffin, K.J., Stout, C.D., and Johnson, E.F. (2007). Adaptations for the oxidation of polycyclic aromatic hydrocarbons exhibited by the structure of human P450 1A2. J. Biol. Chem. *282*, 14348–14355. https://doi.org/10.1074/jbc.M611692200.

Sastry, G.M., Dixon, S.L., and Sherman, W. (2011). Rapid shape-based ligand alignment and virtual screening method based on atom/feature-pair similarities and volume overlap scoring. J. Chem. Inf. Model. *51*, 2455–2466. https://doi.org/10.1021/ci2002704.

Sastry, G.M., Inakollu, V.S.S., and Sherman, W. (2013). Boosting virtual screening enrichments with data fusion: coalescing hits from two-dimensional fingerprints, shape, and docking. J. Chem. Inf. Model. *53*, 1531–1542. https://doi.org/10.1021/ci300463g.

Sastry, M., Lowrie, J.F., Dixon, S.L., and Sherman, W. (2010). Large-scale systematic analysis of 2D fingerprint methods and parameters to improve virtual screening enrichments. J. Chem. Inf. Model. *50*, 771–784. https://doi.org/10.1021/ci100062n.

Schaller, D., Šribar, D., Noonan, T., Deng, L., Nguyen, T.N., Pach, S., Machalz, D., Bermudez, M., and Wolber, G. (2020). Next generation 3D pharmacophore modeling. WIREs Comput. Mol. Sci. *10*, e1468. https://doi.org/https://doi.org/10.1002/wcms.1468.

Schindler, C.E.M., Baumann, H., Blum, A., Böse, D., Buchstaller, H.-P., Burgdorf, L., Cappel, D., Chekler, E., Czodrowski, P., Dorsch, D., et al. (2020). Large-Scale Assessment of Binding Free Energy Calculations in Active Drug Discovery Projects. J. Chem. Inf. Model. *60*, 5457–5474. https://doi.org/10.1021/acs.jcim.0c00900.

Schmidt, D., Boehm, M., McClendon, C.L., Torella, R., and Gohlke, H. (2019). Cosolvent-Enhanced Sampling and Unbiased Identification of Cryptic Pockets Suitable for Structure-Based Drug Design. J. Chem. Theory Comput. *15*, 3331–3343. https://doi.org/10.1021/acs.jctc.8b01295.

Scott, D.E., Coyne, A.G., Hudson, S.A., and Abell, C. (2012). Fragment based approaches in drug discovery and chemical biology. Biochemistry *51*, 4990–5003. https://doi.org/10.1021/bi3005126.

Scott, D.E., Bayly, A.R., Abell, C., and Skidmore, J. (2016). Small molecules, big targets: drug discovery faces the protein–protein interaction challenge. Nat. Rev. Drug Discov. *15*, 533. .

Seco, J., Luque, F.J., and Barril, X. (2009). Binding site detection and druggability index from first principles. J. Med. Chem. *52*, 2363–2371. https://doi.org/10.1021/jm801385d.

Selwa, E., Martiny, V.Y., and Iorga, B.I. (2016). Molecular docking performance evaluated on the D3R Grand Challenge 2015 drug-like ligand datasets. J. Comput. Aided. Mol. Des. *30*, 829–839. https://doi.org/10.1007/s10822-016-9983-3.

Seppälä, J., Bernardi, R.C., Haataja, T.J.K., Hellman, M., Pentikäinen, O.T., Schulten, K., Permi, P., Ylänne, J., and Pentikäinen, U. (2017). Skeletal Dysplasia Mutations Effect on Human Filamins' Structure and Mechanosensing. Sci. Rep. *7*. https://doi.org/10.1038/s41598-017-04441-x.

Sevrioukova, I.F., and Poulos, T.L. (2013). Dissecting cytochrome P450 3A4-ligand interactions using ritonavir analogues. Biochemistry *52*, 4474–4481. https://doi.org/10.1021/bi4005396.

Shahrokh, K., Orendt, A., Yost, G.S., and Cheatham, T.E. 3rd (2012). Quantum mechanically derived AMBER-compatible heme parameters for various states of the cytochrome P450 catalytic cycle. J. Comput. Chem. *33*, 119–133. https://doi.org/10.1002/jcc.21922.

Shan, Y., Mysore, V.P., Leffler, A.E., Kim, E.T., Sagawa, S., and Shaw, D.E. (2022). How does a small molecule bind at a cryptic binding site? PLoS Comput. Biol. *18*, e1009817. https://doi.org/10.1371/journal.pcbi.1009817.

Shang, J., Ye, G., Shi, K., Wan, Y., Luo, C., Aihara, H., Geng, Q., Auerbach, A., and Li, F. (2020). Structural basis of receptor recognition by SARS-CoV-2. Nature *581*, 221–224. https://doi.org/10.1038/s41586-020-2179-y.

Shaw, D.E., Dror, R.O., Salmon, J.K., Grossman, J.P., Mackenzie, K.M., Bank, J.A., Young, C., Deneroff, M.M., Batson, B., Bowers, K.J., et al. (2009). Millisecond-Scale Molecular Dynamics Simulations on Anton. In Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis, (New York, NY, USA: Association for Computing Machinery), p.

Shelley, J.C., Cholleti, A., Frye, L.L., Greenwood, J.R., Timlin, M.R., and Uchimaya, M. (2007). Epik: a software program for pK(a) prediction and protonation state generation for drug-like molecules. J. Comput. Aided. Mol. Des. *21*, 681–691. https://doi.org/10.1007/s10822-007-9133-z.

Sheng, Y., Chen, Y., Wang, L., Liu, G., Li, W., and Tang, Y. (2014). Effects of protein flexibility on the site of metabolism prediction for CYP2A6 substrates. J. Mol. Graph. Model. *54*, 90–99. https://doi.org/10.1016/j.jmgm.2014.09.005.

Shiau, A.K., Barstad, D., Loria, P.M., Cheng, L., Kushner, P.J., Agard, D.A., and Greene, G.L. (1998). The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen. Cell *95*, 927–937. https://doi.org/10.1016/s0092-8674(00)81717-1.

Shirts, M.R., Mobley, D.L., and Chodera, J.D. (2007). Chapter 4 Alchemical Free Energy Calculations: Ready for Prime Time? D.C. Spellmeyer, and R. Wheeler, eds. (Elsevier), pp. 41–59.

Shoichet, B.K. (2004). Virtual screening of chemical libraries. Nature *432*, 862–865. https://doi.org/10.1038/nature03197.

Shoichet, B.K., and Kuntz, I.D. (1993). Matching chemistry and shape in molecular docking. Protein Eng. *6*, 723–732. https://doi.org/10.1093/protein/6.7.723.

Shoichet, B.K., McGovern, S.L., Wei, B., and Irwin, J.J. (2002). Lead discovery using molecular docking. Curr. Opin. Chem. Biol. *6*, 439–446. https://doi.org/10.1016/s1367-5931(02)00339-3.

Sliwoski, G., Kothiwale, S., Meiler, J., and Lowe, E.W.J. (2014). Computational methods in drug discovery. Pharmacol. Rev. *66*, 334–395. https://doi.org/10.1124/pr.112.007336.

Smith, R.D., and Carlson, H.A. (2021). Identification of Cryptic Binding Sites Using MixMD with Standard and Accelerated Molecular Dynamics. J. Chem. Inf. Model. *61*, 1287–1299. https://doi.org/10.1021/acs.jcim.0c01002.

Smith, B.D., Sanders, J.L., Porubsky, P.R., Lushington, G.H., Stout, C.D., and Scott, E.E. (2007). Structure of the human lung cytochrome P450 2A13. J. Biol. Chem. *282*, 17306–17313. https://doi.org/10.1074/jbc.M702361200.

Stepniewska-Dziubinska, M.M., Zielenkiewicz, P., and Siedlecki, P. (2020). Improving detection of protein-ligand binding sites with 3D segmentation. Sci. Rep. *10*, 5035. https://doi.org/10.1038/s41598-020-61860-z.

Sterling, T., and Irwin, J.J. (2015). ZINC 15 – Ligand Discovery for Everyone. J. Chem. Inf. Model. *55*, 2324–2337. https://doi.org/10.1021/acs.jcim.5b00559.

Su, T., Bao, Z., Zhang, Q.Y., Smith, T.J., Hong, J.Y., and Ding, X. (2000). Human cytochrome P450 CYP2A13: predominant expression in the respiratory tract and its high efficiency metabolic activation of a tobacco-specific carcinogen, 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone. Cancer Res. *60*, 5074–5079. .

Sun, D., Gao, W., Hu, H., and Zhou, S. (2022). Why 90% of clinical drug development fails and how to improve it? Acta Pharm. Sin. B https://doi.org/https://doi.org/10.1016/j.apsb.2022.02.002.

Sun, H., Li, Y., Shen, M., Tian, S., Xu, L., Pan, P., Guan, Y., and Hou, T. (2014a). Assessing the performance of MM/PBSA and MM/GBSA methods. 5. Improved docking performance using high solute dielectric constant MM/GBSA and MM/PBSA rescoring. Phys. Chem. Chem. Phys. *16*, 22035–22045. https://doi.org/10.1039/c4cp03179b.

Sun, H., Li, Y., Tian, S., Xu, L., and Hou, T. (2014b). Assessing the performance of MM/PBSA and MM/GBSA methods. 4. Accuracies of MM/PBSA and MM/GBSA methodologies evaluated by

various simulation protocols using PDBbind data set. Phys. Chem. Chem. Phys. *16*, 16719–16729. https://doi.org/10.1039/c4cp01388c.

Szymański, P., Markowicz, M., and Mikiciuk-Olasik, E. (2012). Adaptation of high-throughput screening in drug discovery-toxicological screening tests. Int. J. Mol. Sci. *13*, 427–452. https://doi.org/10.3390/ijms13010427.

Toelzer, C., Gupta, K., Yadav, S.K.N., Borucu, U., Davidson, A.D., Kavanagh Williamson, M., Shoemark, D.K., Garzoni, F., Staufer, O., Milligan, R., et al. (2020). Free fatty acid binding pocket in the locked structure of SARS-CoV-2 spike protein. Science (80-.). *370*, 725–730. https://doi.org/10.1126/science.abd3255.

Tsai, K.-C., Chen, S.-Y., Liang, P.-H., Lu, I.-L., Mahindroo, N., Hsieh, H.-P., Chao, Y.-S., Liu, L., Liu, D., Lien, W., et al. (2006). Discovery of a novel family of SARS-CoV protease inhibitors by virtual screening and 3D-QSAR studies. J. Med. Chem. *49*, 3485–3495. https://doi.org/10.1021/jm050852f.

Tze-Yang Ng, J., and Tan, Y.S. (2022). Accelerated Ligand-Mapping Molecular Dynamics Simulations for the Detection of Recalcitrant Cryptic Pockets and Occluded Binding Sites. J. Chem. Theory Comput. *18*, 1969–1981. https://doi.org/10.1021/acs.jctc.1c01177.

Ung, P.M.U., Ghanakota, P., Graham, S.E., Lexa, K.W., and Carlson, H.A. (2016). Identifying binding hot spots on protein surfaces by mixed-solvent molecular dynamics: HIV-1 protease as a test case. Biopolymers *105*, 21–34. https://doi.org/10.1002/bip.22742.

Vainio, M.J., Puranen, J.S., and Johnson, M.S. (2009). ShaEP: Molecular overlay based on shape and electrostatic potential. J. Chem. Inf. Model. *49*, 492–502. https://doi.org/10.1021/ci800315d.

Veeramachaneni, G.K., Thunuguntla, V.B.S.C., Bobbillapati, J., and Bondili, J.S. (2021). Structural and simulation analysis of hotspot residues interactions of SARS-CoV 2 with human ACE2 receptor. J. Biomol. Struct. Dyn. *39*, 4015–4025. https://doi.org/10.1080/07391102.2020.1773318.

Virtanen, S.I., and Pentikäinen, O.T. (2010). Efficient virtual screening using multiple protein conformations described as negative images of the ligand-binding site. J. Chem. Inf. Model. *50*, 1005–1011. https://doi.org/10.1021/ci100121c.

Virtanen, S.I., Niinivehmas, S.P., and Pentikäinen, O.T. (2015). Case-specific performance of MM-PBSA, MM-GBSA, and SIE in virtual screening. J. Mol. Graph. Model. *62*, 303–318. https://doi.org/10.1016/j.jmgm.2015.10.012.

Walls, A.C., Park, Y.-J., Tortorici, M.A., Wall, A., McGuire, A.T., and Veesler, D. (2020). Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. Cell *181*, 281-292.e6. https://doi.org/https://doi.org/10.1016/j.cell.2020.02.058.

Walsh, A.A., Szklarz, G.D., and Scott, E.E. (2013). Human cytochrome P450 1A1 structure and utility in understanding drug and xenobiotic metabolism. J. Biol. Chem. *288*, 12932–12943. https://doi.org/10.1074/jbc.M113.452953.

Wan, S., Wright, D.W., and Coveney, P. V (2012). Mechanism of drug efficacy within the EGF receptor revealed by microsecond molecular dynamics simulation. Mol. Cancer Ther. *11*, 2394–2400. https://doi.org/10.1158/1535-7163.MCT-12-0644-T.

Wan, S., Bhati, A.P., Zasada, S.J., and Coveney, P. V (2020). Rapid, accurate, precise and reproducible ligand-protein binding free energy prediction. Interface Focus *10*, 20200007. https://doi.org/10.1098/rsfs.2020.0007.

Wang, A., Savas, U., Stout, C.D., and Johnson, E.F. (2011). Structural characterization of the complex between alpha-naphthoflavone and human cytochrome P450 1B1. J. Biol. Chem. *286*, 5736–5743. https://doi.org/10.1074/jbc.M110.204420.

Wang, A., Savas, U., Hsu, M.-H., Stout, C.D., and Johnson, E.F. (2012). Crystal structure of human cytochrome P450 2D6 with prinomastat bound. J. Biol. Chem. *287*, 10834–10843. https://doi.org/10.1074/jbc.M111.307918.

Wang, J., Wolf, R.M., Caldwell, J.W., Kollman, P.A., and Case, D.A. (2004). Development and testing of a general amber force field. J. Comput. Chem. *25*, 1157–1174. https://doi.org/10.1002/jcc.20035.

Wang, J., Wang, W., Kollman, P.A., and Case, D.A. (2006). Automatic atom type and bond type perception in molecular mechanical calculations. J. Mol. Graph. Model. *25*, 247–260. .

Wang, L., Wu, Y., Deng, Y., Kim, B., Pierce, L., Krilov, G., Lupyan, D., Robinson, S., Dahlgren, M.K., Greenwood, J., et al. (2015). Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. J. Am. Chem. Soc. *137*, 2695–2703. https://doi.org/10.1021/ja512751q.

Wass, M.N., Kelley, L.A., and Sternberg, M.J.E. (2010). 3DLigandSite: predicting ligand-binding sites using similar structures. Nucleic Acids Res. *38*, W469-73. https://doi.org/10.1093/nar/gkq406.

Watts, K.S., Dalal, P., Murphy, R.B., Sherman, W., Friesner, R.A., and Shelley, J.C. (2010). ConfGen: A conformational search method for efficient generation of bioactive conformers. J. Chem. Inf. Model. *50*, 534–546. https://doi.org/10.1021/ci100015j.

Wells, J.A., and McClendon, C.L. (2007). Reaching for high-hanging fruit in drug discovery at protein–protein interfaces. Nature *450*, 1001–1009. https://doi.org/10.1038/nature06526.

Wishart, D.S., Knox, C., Guo, A.C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z., and Woolsey, J. (2006). DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. *34*, D668-72. https://doi.org/10.1093/nar/gkj067.

Word, J.M., Lovell, S.C., Richardson, J.S., and Richardson, D.C. (1999). Asparagine and Glutamine: Using Hydrogen Atom Contacts in the Choice of Side-chain Amide Orientation. J. Mol. Biol. *285*, 1735–1747. https://doi.org/10.1006/jmbi.1998.2401.

Wu, Y., Wang, F., Shen, C., Peng, W., Li, D., Zhao, C., Li, Z., Li, S., Bi, Y., Yang, Y., et al. (2020). A noncompeting pair of human neutralizing antibodies block COVID-19 virus binding to its receptor ACE2. Science (80-.). https://doi.org/10.1126/science.abc2241.

Xiong, L., Zhu, X.-L., Gao, H.-W., Fu, Y., Hu, S.-Q., Jiang, L.-N., Yang, W.-C., and Yang, G.-F. (2016). Discovery of Potent Succinate-Ubiquinone Oxidoreductase Inhibitors via Pharmacophore-linked Fragment Virtual Screening Approach. J. Agric. Food Chem. *64*, 4830–4837. https://doi.org/10.1021/acs.jafc.6b00325.

Yang, C.-Y. (2015). Identification of potential small molecule allosteric modulator sites on IL-1R1 ectodomain using accelerated conformational sampling method. PLoS One *10*, e0118671. https://doi.org/10.1371/journal.pone.0118671.

Yano, J.K., Hsu, M.-H.H., Griffin, K.J., Stout, C.D., and Johnson, E.F. (2005). Structures of human microsomal cytochrome P450 2A6 complexed with coumarin and methoxsalen. Nat. Struct. Mol. Biol. *12*, 822–823. https://doi.org/10.1038/nsmb971.

Ylilauri, M., and Pentikäinen, O.T. (2013). MMGBSA as a tool to understand the binding affinities of filamin-peptide interactions. J. Chem. Inf. Model. *53*. https://doi.org/10.1021/ci4002475.

Yu, W., Lakkaraju, S.K., Raman, E.P., Fang, L., and MacKerell, A.D.J. (2015). Pharmacophore modeling using site-identification by ligand competitive saturation (SILCS) with multiple probe molecules. J. Chem. Inf. Model. *55*, 407–420. https://doi.org/10.1021/ci500691p.

Yuan, M., Wu, N.C., Zhu, X., Lee, C.-C.D., So, R.T.Y., Lv, H., Mok, C.K.P., and Wilson, I.A. (2020). A highly conserved cryptic epitope in the receptor binding domains of SARS-CoV-2 and SARS-CoV. Science *368*, 630–633. https://doi.org/10.1126/science.abb7269.

Zhang, Z., and Tang, W. (2018). Drug metabolism in drug discovery and development. Acta Pharm. Sin. B *8*, 721–732. https://doi.org/https://doi.org/10.1016/j.apsb.2018.04.003.

Zhao, L., Li, S., and Zhong, W. (2022). Mechanism of Action of Small-Molecule Agents in Ongoing Clinical Trials for SARS-CoV-2: A Review. Front. Pharmacol. *13*. https://doi.org/10.3389/fphar.2022.840639.

Zinner, M., Dahlhausen, F., Boehme, P., Ehlers, J., Bieske, L., and Fehring, L. (2021). Quantum computing's potential for drug discovery: Early stage industry dynamics. Drug Discov. Today *26*, 1680–1688. https://doi.org/10.1016/j.drudis.2021.06.003.

Author's contributions to articles

Article I. E.M.J. designed the study with all other authors. E.M.J. performed all the calculations and the related data analysis. All authors contributed to the manuscript and approved the submitted version.

Article II. E.M.J. designed the computational modelling studies with O.T.P. E.M.J. performed all calculations and the related data analysis. All authors contributed to the manuscript and approved the submitted version.

Article III. E.M.J. designed and performed all the computational modelling studies and the related data analysis. All authors contributed to the manuscript and approved the submitted version.

Article IV. E.M.J. designed the study with the other authors and performed all calculations and data analysis except for molecular library preparation. All authors contributed to the manuscript and approved the submitted version.

TURUN YLIOPISTO
UNIVERSITY OF TURKU