1

Introduction
Most drugs act by means of binding to (and activation or inhibition of) their target protein, while many side-effects are related to unwanted complexation with off-target proteins. Hence, evaluating and predicting the **affinity** of protein binding of drug-like compounds is directly relevant to pharmaceutical scientists. This thesis describes the development, evaluation, comparison and application of linear interaction energy (LIE) models for efficient computation of protein–ligand binding affinities in terms of their **free energy** of binding. In the current introductory chapter, a short overview is given of binding free energy calculations in the more general context of computer–aided drug design. The next chapter is devoted to a more detailed review of LIE and subsequent chapters present studies on the domains of applicability of (automated) LIE models, a rigorous approach to extend them, and a comparison with other efficient free energy calculations. Different (flexible) proteins of pharmaceutical relevance will act as test case, which are shortly introduced in the current chapter as well. The thesis ends with a summary and future perspectives, while the current chapter concludes with a detailed overview of the outline and aims of this thesis.
1.1. The computer in the world of biomolecules and drugs

The combination of information on chemical structures, their properties, and their interactions with biological macromolecular targets, and the translation of such data into knowledge intended to help in making decisions for drug discovery and development are on the basis of the discipline of computer-aided drug design [1]. This research field studies the physical, chemical, and biological responses to small molecules in living organisms with the aid of compute resources, by means of modeling and simulation in the contexts of virtual screening, de novo design, evaluation of drug-likeness, and drug–receptor binding determination [2]. Computational methodologies play many roles and form an important part of many drug discovery programs, from hit identification to lead optimization, and methods spanning from ligand-based to structure-based approaches have been frequently used in drug design efforts [3].

Increasing computational power [4] and availability of in silico resources have allowed and facilitated the application and development of algorithms, methods, models, and software over the last few decades [5, 6]. With simulation, microscopic aspects and properties studied in experiments are mimicked as representative as possible and can even be described at the electronic level, making that the contribution of any aspect to the outcome could in principle be inspected by removing or modifying them [7]. An example is computational alchemy in free energy calculations, in which atoms are generated and annihilated in simulations of nonphysical states. In addition, the use of computational modeling greatly facilitates the visualization of abstract concepts such as atomic interaction between molecules. Modeling can also reduce costs, time and/or risks and increase safety of experimental studies, e.g. in the case of extreme conditions or handling of toxic compounds. Moreover, computers can be used to automate tedious tasks and thereby help in avoiding human errors, and simulations can be performed to interpret and rationalize ambiguous experimental results or to answer research questions that require time-scale and/or system-size resolutions that cannot be reached in experiment [8], e.g., when monitoring the transition between open and closed conformations of proteins [7]. Computers can also play an essential role in predicting and rationalizing the interaction between compounds with lower molecular weight (ligands) and their macromolecular biological (off-)targets or receptors as in protein-binding affinity calculation.

1.2. An arsenal of computational methods

For the purpose of computer-aided drug design in general and binding affinity computation in particular, an appropriate selection of techniques from a broad field of available in silico methods is necessary. To select the most proper method(s) for a particular case, usually several aspects should be considered [9], such as:

- **the research question.** Are the efforts intended to filter potential binders from decoys of a large dataset or to improve the potency and/or safety of (lead) compounds? Does the research aim to rationalize, or to predict experimental results? What are the characteristics of the system of interest?

- **the availability of receptor and/or ligand data.** How abundant and diverse
are the available ligands? Is the target protein already known and characterized as well? Is the abundance of the data sufficient to generate a training set?

- available resources. Is the task accomplishable with the preferred method in a timely manner? Can the current accessible computer infrastructure cope with the computational demand of the method? How much supervision and technical knowledge is required to perform such approach?

- (possibilities for) method validation. Can we cross-validate the employed method? Is the applicability domain of the model defined and can it be quantified?

From a technical point of view, three levels of detail can be considered in methods to predict protein–ligand binding, as described in more detail below and as schematically represented in Figure 1.1: (i) ligand–based methods, employed in attempts to directly and cost–efficiently correlate molecular properties with end–points for prediction, which are not necessarily binding free energies but can also be other properties related to biological activity, (ii) (protein) structure–based methods without (bulk) solvent to predict protein–ligand interaction based on docking and scoring, (iii) structure–based methods with solvent included to predict affinity from sampling of dynamic and thermodynamic properties of the system of interest. In general, the choice for the level or tier of interest is governed by finding a balance between accuracy and feasibility, which will depend on aspects such as the ones listed above.

![Figure 1.1: Classification of computational drug design methods based on the level of detail considered, adapted from reference [9].](image)

1.2.1. Tier 1: Ligand only
This first tier of computational drug design approaches is represented by ligand–based methods in which only information on the (drug) compound is taken into account. This tier has been one of the earliest computer–aided drug design methods since it was pioneered by Corwin Hansch in the early 1960s [10]. It has evolved from relatively simple regressions of quantitative structure activity relationships (QSARs) and pharmacophore
modelling to sophisticated statistical and machine learning techniques [11] used to analyze huge data sets of diverse ligands [12]. With methods belonging to this tier, ADMET or other molecular properties can also explain the basis of the drug–likeliness of the ligands [2]. Traditionally, the core of ligand–based modelling is the use of molecular descriptors such as structural and physicochemical properties, or "features" as derived from machine learning to establish correlations with the ligands' biological activity or affinity towards the target protein. Especially in the case of abundantly available information on large sets of compounds, the use of ligand–based methods can be beneficial to find possible biological patterns featured in the dataset [13–15]. When ligand information is known but the structure of the target protein is not available yet, ligand–based methods can be advantageous over structure–based ones (Tiers 2 and 3) although the use of these methods remains may well be limited by the fact that the information of protein–ligand interaction cannot be retrieved.

1.2.2. Tier 2: Ligand + protein

The importance of assessing the interaction between a ligand and its receptor protein was recognized 125 years ago by Emil Fischer. In 1894 he formulated his "lock and key" hypothesis [16] stating that "enzyme and glycoside must fit together like a key and a lock in order to initiate a chemical action upon each other", assuming that the shape of the receptor active site is exactly complementary to the shape of the ligand, which fits as a key in a lock. This idea has laid the basis of rigid body docking and was many decades later adapted to the induced fit model, in which both the receptor and ligand adapt their conformations to bind to each other [17]. When the target protein is structurally characterized, these (protein) structure–based methods may well be preferred over the ligand–based methods, because of the more detailed information of protein–ligand interaction that can be retrieved. Following the rapidly enhanced availability of structures of biomolecules in the Protein Data Bank (PDB) [18, 19], the popularity and ease of using structure–based drug design approaches has increased ever since, along with the detailed information that can be obtained by docking ligands and/or building pharmacophores on the modes of binding between drug molecules and their target. Structure–based drug design methodologies have delivered several successful stories of drug candidates entering the clinical trial and regulatory approval phases [20], including examples of de novo design [21] in which the novel drug candidate is grown from a seed scaffold in order to adjust to the characteristics and properties of the protein binding site. In the 1980s, molecular docking methods to automatically explore geometrically feasible alignments of ligands and receptors were extensively pioneered [22]. Compared to de novo approaches, docking is more popular due to the availability of large ligand library databases nowadays (such as PubChem [23] and ZINC [24–26]), which can be used directly for the purpose of virtual screening campaigns. Screening ligands from a database with docking has the advantage over de novo design that ligands can be pre–selected according to desirable properties, such as ease of synthesis, high bioaccessibility and safety [27].

The performance of docking is contingent to two factors, the search algorithm and the scoring function [28]. The former generates binding poses to find minima at the potential or free energy landscape as defined by the (empirical, force field, or knowledge–
1. Introduction

based) scoring function, which assesses the potential ligand complementarity to the binding site and differentiates probable poses from the improbable ones. Hence, molecular docking can be used for binding pose prediction and to distinguish binders from nonbinders in virtual screening campaigns. For the purpose of quantitative affinity and binding free energy prediction, the applicability of docking and scoring can be limited, as discussed in Section 1.3.1.

While ligand and selected receptor side–chain flexibility can be taken into account in most docking tools, docking can typically be considered as using a rigid–body protein template [28]. Similarly, solvent effects are often neglected or incorporated by means of active site water molecules only [29]. Thus, conformational sampling and incorporation of the contribution of (de)solvation to protein–ligand binding is often limited, which can be especially an issue for flexible proteins with a malleable active site and/or that can bind their ligands in multiple orientations [30]. Although approximate, docking is often crucial to select starting points for subsequent modeling, such as simulations in the context of binding free energy computation [30–32] that take place at the level of the next tier.

1.2.3. Tier 3: Ligand + protein + solvent

When interested in appropriate conformational sampling of the system, e.g., for the purpose of (binding) free energy calculation, it is not only necessary to turn to methods such as molecular dynamics (MD) or Monte Carlo (MC) simulation [33] but also to explicitly include (bulk) solvent molecules in the system setup [7], in order to circumvent for example surface tension artifacts [8]. As a result, the current tier is the tier of choice in all chapters included in this thesis.

There are three primary interests of using molecular simulation methods such as MD or MC: to sample configurational space, to obtain a description of the system at equilibrium, and to examine the actual dynamics. For the first two, either MC or MD can be used. On the other hand, in case where the motions and its evolution over time are of primary interest, only MD can provide this information [7]. In MD simulations, which have been for more than 40 years used to study proteins in silico [34], Newton’s equations of motion are integrated over time to generate possible conformations of the system of interest. For that purpose, the forces \( \vec{f}_i \) acting on the degrees of freedom \( i \) (with mass \( m_i \)) have to be evaluated. These degrees of freedom may be chosen at the atomic level (as in the current work) but can also be treated at a higher (QM) or lower (coarse–grained) level of detail. Per time step, the \( \vec{f}_i \)'s can be calculated from the gradient of the potential energy \( V \), in order to update the velocities and the positions \( \vec{r}_i \) of the degrees of freedom using integrators such as Leap–Frog [35] or Verlet [36], as follows:

\[
m_i \frac{d^2 \vec{r}_i}{dt^2} = \vec{f}_i(\vec{r}) = -\frac{\partial V(\vec{r})}{\partial \vec{r}_i}
\]

Integrators are embedded in MD software packages such as GROMACS [37], AMBER [38], GROMOS [39], NAMD [40], CHARMM [41], Desmond [42], Q [43], and OpenMM [44]. The success of the integration of motion to describe the dynamics depends directly on the quality of the force field used. Atomistic force fields describe the potential energy
of the simulated system with a set of atomic coordinates \( r \), by means of a collection of covalent and noncovalent interaction terms:

\[
V(r) = V_{\text{bond}}(r) + V_{\text{angle}}(r) + V_{\text{dihedral}}(r) + V_{\text{improper}}(r) + V_{\text{vdw}}(r) + V_{\text{ele}}(r)
\]  

(1.2)

with \( V_{\text{bond}}(r) \), \( V_{\text{angle}}(r) \), \( V_{\text{dihedral}}(r) \), and \( V_{\text{improper}}(r) \) describing the bonded interactions and \( V_{\text{vdw}}(r) \) and \( V_{\text{ele}}(r) \) describing the nonbonded interactions. Covalent bonds, angles and impropers are typically treated as harmonic springs and dihedral angles via a goniometric potential–energy function, while the van der Waals and electrostatic interaction energies are usually evaluated with a Lennard–Jones and Coulomb potential, respectively. Using Equation 1.2, the internal energy of the system and its gradient is evaluated by penalizing deviations from optimal configurations for the different types of interactions considered. Popular biomolecular force fields are AMBER [45], GROMOS [46], CHARMM [47] and OPLS [48].

By obtaining atomic positions and velocities over time, the dynamic effects on the atomic and molecular interactions can be observed. In MD and MC simulations, the solvent can be included explicitly by using periodic boundary conditions [49] (to circumvent any solvent–vacuum interface, which would again lead to possible surface tension effects) or implicitly by using a continuum electrostatic model [50]. Solvent mobility has been demonstrated to be an important factor in determining the atomic fluctuations above 180 K, although intrinsic protein effects become dominant at lower temperatures, confirming that including solvent effects in studying biomolecular system helps to understand functionally important dynamics of the protein [51]. Not only the structure but also the dynamics of proteins can in many cases play a key role in their functionality, considering e.g. that they may well demonstrate conformational changes upon activation and/or inhibition [7]. Consequently, biomolecular properties and processes in such cases can only be understood when dynamics are explicitly taken into consideration [52]. Obviously, this is not only restricted to inclusion of induced fit effects, protein malleability and/or promiscuity in case of binding free energy calculations, but also to the study of e.g. allosteric effects, signaling or protein (un)folding. In such cases MD simulation has played important roles in studying for example folding or drug binding, as well as many other aspects of life sciences governed by the forces and interactions at the atomic level of detail.

1.3. Protein–ligand binding free energy calculations

Mutual molecular recognition is the starting point for a wide variety of biological processes [53]. Binding affinity governs ligand binding to target proteins, and being able to quantitatively understand and predict affinity can greatly support lead finding and/or optimization [54]. Binding affinity can be quantified as the change in the Gibbs free energy (\( \Delta G_{\text{bind}} \)) of a system when a protein and ligand form a complex [55]. \( \Delta G_{\text{bind}} \) is composed of an enthalpic and entropic contribution and can be directly related to the dissociation constant \( K_D \) via:

\[
\Delta G_{\text{bind}} = RT \ln K_D
\]

(1.3)

For reversible competitive inhibitions, \( K_D \) is often set equal to the inhibition constant \( (K_i) \) [56], such that Equation 1.3 is also of direct use to derive \( \Delta G_{\text{bind}} \) from experimental...
1. Introduction

Data in case $K_i$’s but no $K_D$’s are available.

When molecular simulation can be used to obtain similar accuracy in $\Delta G_{\text{bind}}$ as experiment (which is typically on the order of 4–5 kJ mol$^{-1}$ [57, 58] or within an order of magnitude in binding constants, Equation 1.3), $\Delta G_{\text{bind}}$ computation can be not only of help in identifying binders from databases of compounds and/or in proposing potential new lead compounds, but it can also provide insight at the atomistic level for the key determinants in the binding process. Obviously such insights can be of great value to any kind of prediction and rationalization within the drug design process as well [59]. However, although computing $\Delta G_{\text{bind}}$ via simulation has been finding its way into drug design settings in recent years [60], it is still a challenging task to efficiently account for the enthalpic and entropic contributions in an appropriate way, e.g., due to the potential role played by protein flexibility and/or solvent effects, and the possible difficulties to accurately obtain a detailed description of interatomic interactions within the biomolecular system [61]. For these reasons, $\Delta G_{\text{bind}}$ computation can still be a time-consuming task that requires expertise, for which several strategies within the second and third tiers can be distinguished. These are shortly reviewed in the next subsections and summarized in Figure 1.2.

![Figure 1.2: Classification of $\Delta G_{\text{bind}}$ calculation methods based on speed and accuracy.](image)

1.3.1. Docking and scoring

For the purpose of quantitatively predicting $\Delta G_{\text{bind}}$, docking and subsequent scoring have only been applied with limited success [53]. An important reason is that docking and scoring functions attempt to approximate $\Delta G_{\text{bind}}$ typically by using a single protein structure [62], and the exclusion of conformational sampling, (bulk) solvent effects and entropy makes the calculation less accurate. Even affinity ranking can be difficult to be accomplished by this category of method [63]. As a result, docking and scoring are
1.3. Protein–ligand binding free energy calculations

1.3.2. Alchemical methods
Alchemical free energy approaches rigorously evaluate (differences in) $\Delta G_{\text{bind}}$ by following statistical mechanics and by taking conformational sampling and solvent effects into account. For these purposes, intermediate states representing the condition where atoms are generated, annihilated, or changed in terms of their identity (e.g. atom type, charge, interaction with surrounding) are performed to obtain free energy differences between compounds of interests\(^{[59, 65, 66]}\). Examples of methods belonging to this category are free energy perturbation (FEP)\(^{[67]}\), thermodynamic integration (TI)\(^{[68]}\), one–step perturbation (OSP)\(^{[69]}\), umbrella sampling\(^{[70]}\), replica exchange (RE)\(^{[71]}\), enveloping distribution sampling (EDS)\(^{[72]}\), Bennett acceptance ratio (BAR)\(^{[73]}\), and $\lambda$–dynamics\(^{[74]}\). The compute–demanding characteristics of these methods caused by the calculation of intermediate states still limits their typical application to congeneric series of ligands and makes them impractical for high–throughput scenarios\(^{[75, 76]}\).

In Chapter 5 of this thesis, FEP and (multi–state) BAR are employed in order to combine alchemical solvation free energy calculations with LIE and in this way efficiently improve binding affinity calculation based on end–point computation, which is introduced in the following subsection.

1.3.3. End–point methods
End–point methods aim to cover a balance between accuracy and efficiency required for $\Delta G_{\text{bind}}$ computation. For that purpose, they combine conformational sampling (typically in MD) with a relatively fast scoring approach. As in alchemical approaches, the thermodynamically relevant phenomena for $\Delta G_{\text{bind}}$ calculation (conformational ensembles, solvent effects) can be explicitly included but, analogous to docking, scoring has an experimental and/or approximate basis. By definition, end–point methods only require initial and final states to be simulated, i.e. the ligand free in solution and bound to the target protein, respectively, and/or interactions being either turned off and on. Hence, they are significantly more efficient than alchemical methods that also require several intermediate nonphysical states. The entropic contribution can be optionally approximated via e.g. additional empirical parameters. It also allows users to treat structurally–diverse ligands that are common in large datasets, and are difficult to be handled with rigorous methods. With this advantage, end–point methods seem to be promising for (semi–)high throughput application in drug design, in order to identify hits that can be compared and ranked according to their affinity to the target protein, either in terms of relative or absolute $\Delta G_{\text{bind}}$ values.

The very basic and earlier end–point method is the linear response approximation (LRA)\(^{[77]}\). LRA directly relates $\Delta G_{\text{bind}}$ to the electrostatic interaction energy involving the ligand and its surrounding. In this method, two end–states are required, represented by a charged and uncharged ligand (electrostatic interactions turned off) in complex with the protein. Assuming linear response theory, these interactions are in LRA scaled by a factor of 0.5 to obtain the corresponding change in free energy\(^{[77]}\), which can be repeated for the unbound ligand in water to obtain the electrostatic contribution to
ΔG_{bind}. In a similar spirit, the change in ligand–environment van der Waals interaction upon going from the unbound to the protein–bound state can be linearly scaled (using e.g. an empirically calibrated value), which is on the basis of LIE theory as well (Section 1.3.4, Chapter 2). Together with LIE, LRA is applied in Chapter 5.

More end–points methods and extensions emerged afterwards, most notably including, besides LIE [78], the molecular mechanics Poisson–Boltzmann surface area method (MM/PBSA) [79] and its alternative MM/GBSA (in which use of Poisson–Boltzmann solvation free energy evaluation is replaced by the Generalized Born approach) [79]. MM/PBSA is described and compared to LIE in Chapter 4 and is together with LIE the most widespread used end–point method. This is illustrated by Figure 1.3, which compares the number of citations of LIE and MM/PBSA with LRA and other end–point methods (i.e. mining minima [80] and solvated interaction energy (SIE) [81]).

Figure 1.3: Number of citations of several end–point methods principle papers, as recorded per 16th of June 2019 in Web of Science.

1.3.4. Linear interaction energy (LIE)
Linear interaction energy (LIE) is an empirical end–point method which was coined in 1994 by Åqvist and coworkers [78]. It is derived from LRA to compute the electrostatic contributions to the binding affinity, while the nonpolar contribution to the ΔG_{bind} calculation is represented by calculating the difference in the average van der Waals interaction energies between the ligand and its environment in either the bound and the unbound state. The approximation of LIE is inherited from LRA, since LRA itself is an approximate method based on an expansion of the Zwanzig expression for free–energy
perturbation theory [82].

In its original version, LIE assumes that intramolecular energies, entropic terms, desolvation effects, or other factors contributing to $\Delta G_{\text{bind}}$ can be handled and cancelled out by fitting and scaling empirical parameters, as it is assumed to correlate linearly with the intermolecular interactions [83]. This fitting allows for the calculation of "absolute" (direct) values for $\Delta G_{\text{bind}}$, which in turn makes it straightforward to use a Boltzmann–like weighting scheme to include multiple binding poses of ligands combined into a single prediction of $\Delta G_{\text{bind}}$ [30], Section 2.4. This is relevant for flexible proteins such as Cytochrome P450s that may bind their ligands in different orientations or that may adopt multiple (partial) conformations upon complexation [30]. LIE can also handle diverse ligands in the dataset that may involve too large perturbations to be simulated (which may become impractical for alchemical free energy calculations), while simultaneously accounting for the unbound state of the ligand that is not considered by most empirical scoring functions [62].

To compute $\Delta G_{\text{bind}}$ from the simulations of the ligand either bound to the protein of free in solvent, the obtained average van der Waals ($vdw$) and electrostatic ($ele$) interaction energies of the ligand with its environment are scaled by LIE parameters $\alpha$ and $\beta$ [78]:

$$\Delta G_{\text{bind}} = \alpha (\langle V_{vdw}^{\text{lig-sur}} \rangle_{\text{bound},i} - \langle V_{vdw}^{\text{lig-sur}} \rangle_{\text{free}}) + \beta (\langle V_{ele}^{\text{lig-sur}} \rangle_{\text{bound},i} - \langle V_{ele}^{\text{lig-sur}} \rangle_{\text{free}}) \quad (1.4)$$

Originally LRA was followed and $\beta$ was set to 0.5 but later it was found to be system–dependent and it became an adjustable parameter afterwards, just like $\alpha$ [84, 85]. After $\alpha$ and $\beta$ are calibrated based on a training set of compounds with experimentally known affinities, Equation 1.4 can be used to predict binding affinities of ligands with unknown experimental data. For a more detailed overview on LIE and its methodological details, possible limitations and proposed extensions for further improvements, we refer to Chapter 2.

1.4. Aims and scope

The general aims of this thesis are to evaluate the practical applicability of LIE in real–life drug design scenarios and to improve its performance in calculating protein–ligand $\Delta G_{\text{bind}}$ values. Individual chapters of this thesis focus on different aspects of $\Delta G_{\text{bind}}$ calculation by using several strategies described above, with an emphasis on the evaluation and development of LIE models.

Chapter 2 serves as a review of state–of–the–art in LIE at the start of the PhD research project presented in this thesis. It includes a discussion of the central LIE equation and approximations taken, selected model–parameterization and application studies, and a discussion of extensions to the method that have been proposed to improve its predictive performance.

LIE allows computing direct ("absolute") values for $\Delta G_{\text{bind}}$, which in turn facilitates inclusion of multiple ligand and/or protein binding conformations in LIE predictions. However, as for any empirical QSAR model, the need for training and the use of fitted parameters ($\alpha$ and $\beta$) may raise the question of how reliable the predicted value will be for an arbitrary query compound. To answer that question one would like to define the
domain of applicability of a given LIE model, which is a set of knowledge or information on which the training set of the model has been developed, in order to get a measure for the confidence in a given prediction for a query compound. Various strategies are available to quantify applicability domains (ADs) of ligand–based predictive models in terms of the descriptors used [86], but methods to analyze the AD of protein–structure (and –dynamic) based models such as LIE were until recently lacking. Only a few years ago, a novel approach was introduced to allow AD assignment of LIE models, based on simulation data only [87, 88]. This approach is described in Chapter 3, which critically evaluates its application in the context of the Drug Design Data Resource (D3R) blind prediction Grand Challenge 2 (GC2) [89], that serves as an unbiased test for evaluating the performance of $\Delta G_{\text{bind}}$ calculation methods. The chapter aims to evaluate the utility of LIE in the real–life scenario when dealing with a large dataset obtained from industry. The applicability of the model was also investigated by combining our applicability domain (AD) metrics with protein–ligand interaction profiling.

D3R GC2 focused on binding affinity prediction of a structural variety of ligands to Farnesoid X receptor (FXR), which is a nuclear receptor (NR) with a key role in regulating cholesterol and bile acid homeostasis [90]. As other NRs, FXR can adopt different conformations when being ligand–bound, in which helices adjacent to the binding site adopt different orientations (Figure 1.4). Considering the structural flexibility of FXR and its ability to bind a variety of ligands that can adopt different orientations, Chapter 3 thus also addresses a general challenge in binding affinity computation, i.e., the prediction of $\Delta G_{\text{bind}}$ for flexible proteins that may bind ligands in multiple orientations. In such cases, obtaining sufficient conformational sampling for free energy calculation can be especially difficult [61]. Similarly, Chapter 5 focuses on binding affinity computation for flexible and promiscuous proteins as well (i.e. Cytochrome P450 isoforms 2A6 and 2E1 that are related to nicotine and alcohol metabolism, respectively), while Chapter 4 presents $\Delta G_{\text{bind}}$ studies on the Sirtuin 1 receptor (SIRT1), which is responsible for regulating and/or deacetylating modulators for stress resistance and metabolism [91] and which can bind its ligands in different parts of the binding pocket (Figure 1.4). To address the challenges in protein flexibility and possible different ligand binding poses in the context of $\Delta G_{\text{bind}}$ prediction, we employ a Boltzmann–like weighting scheme to combine results from multiple simulations (Section 2.4) [30]. A prerequisite of a direct employment of this weighting scheme is the availability of "absolute" $\Delta G_{\text{bind}}$, which is featured in LIE. This in turn makes weighting more straightforward in LIE when compared to other methods such as MM/PBSA or most alchemic approaches.

In Chapter 4, the performance of (Boltzmann–weighted) LIE in calculating SIRT1–ligand $\Delta G_{\text{bind}}$ is compared to use of MM/PBSA. In addition to evaluating their accuracy, we particularly investigate the basis for differences in required computational costs and possibilities to (further) increase efficiency. As in Chapter 3, we also evaluate the potential of using Boltzmann–weighted LIE in an automated fashion, using our in–house automated eTOX ALLIES pipeline [92]. Chapter 5 investigates the potential to relatively and efficiently improve the performance of LIE by rigorously including solvation thermodynamics into our models via alchemical free energy perturbation of the ligands in the unbound state. We also use these results to inspect possible difficulties in traditional LIE modeling. The thesis ends with conclusions and perspectives in Chapter 6.
Figure 1.4: Target proteins used as test cases in this thesis: Farnesoid X receptor (A), Sirtuin 1 (B), Cytochrome P450 2A6 (C), Cytochrome P450 2E1 (D).

References


1. Introduction


References


[56] C. Yung–Chi and W. H. Prusoff, *Relationship between the inhibition constant (K_i) and the concentration of inhibitor which causes 50 per cent inhibition (k_50) of an enzymatic reaction*, Biochem. Pharmacol. 22, 3099 (1973).


1. Introduction


1. Introduction


