Atomistic Binding Free Energy Estimations for Biological Host–Guest Systems

Dissertation zur Erlangung des Doktorgrades der Naturwissenschaften (Dr. rer. nat.) am Fachbereich für Mathematik und Informatik der Freien Universität Berlin

vorgelegt von

Vedat Durmaz

aus Zweibrücken

Berlin, Frühjahr 2016

Die vorliegende Arbeit wurde unter Anleitung von Priv.-Doz. Dr. Marcus Weber in der Arbeitsgruppe Mathematischer Molekülentwurf des Zuse-Instituts Berlin durchgeführt.

1. Gutachter: Priv.-Doz. Dr. Marcus Weber,
Mathematik für Lebens- und Materialwissenschaften,
Mathematischer Molekülentwurf,
Zuse-Institut Berlin

2. Gutachter: Prof. Dr. Paul Wrede
Institut für Molekularbiologie und Bioinformatik,
Charité Universitätsmedizin Berlin

Tag der Disputation: 14.12.2016

Abstract

Accurate quantifications of protein–ligand binding affinities by means of *in silico* methods increasingly gain importance in various scientific branches including toxicology and pharmacology. *In silico* techniques not only are generally less demanding than laboratory experiments regarding time as well as cost, in particular, if binding assays or synthesis protocols need to be developed in advance. At times, they also provide the only access to risk assessments on novel chemical compounds arising from biotic or abiotic degradation of anthropogenic substances. However, despite the continuous technological and algorithmic progress over the past decades, binding free energy estimations through molecular dynamics simulations still pose an enormous computational challenge owed to the mathematical complexity of solvated macromolecular systems often consisting of hundreds of thousands of atoms. The goals of this thesis can roughly be divided into two categories dealing with different aspects of host–guest binding quantification. On the one side algorithmic strategies for a comprehensive exploration and decomposition of conformational space in conjunction with an automated selection of representative molecular geometries and binding poses have been elaborated providing initial structures for free energy calculations. In light of the dreaded trapping problem typically associated with molecular dynamics simulations, the focus was laid on a particularly systematic generation of representatives covering a broad range of physically accessible molecular conformations and interaction modes. On the other side and ensuing from these input geometries, binding affinity models based on the linear interaction energy (LIE) method have been developed for a couple of (bio)molecular systems. The applications included a successful prediction of the liquid-chromatographic elution order as well as retention times of highly similar hexabromocyclododecane (HBCD) stereoisomers, a novel empirical LIE-QSAR hybrid binding affinity model related to the human estrogen receptor α (ERα), and, finally, the (eco)toxicological prioritization of transformation products originating from the antibiotic sulfamethoxazole with respect to their binding affinities to the bacterial enzyme dihydropteroate synthase. Altogether, a fully automated approach to binding mode and affinity estimation has been presented that is content with an arbitrary geometry of a small molecule under observation and a spatial vector specifying the binding site of a potential target molecule. According to our studies, it is superior to conventional docking and thermodynamic average methods and primarily suggesting binding free energy calculation on the basis of several heavily distinct complex geometries. Both chromatographic retention times of HBCD and binding affinities to ERα yielded squared coefficients of correlation with experimental results significantly higher than 0.8. Approximately 85 % (100 %) of predicted receptor–ligand binding modes deviated less than 1.53 Å (2.05 Å) from available crystallographic structures.

"In a time of universal deceit – telling the truth is a revolutionary act."

George Orwell

Contents

To the memory of my brother Ibrahim

1 Introduction

One of the most prominent as well as challenging tasks tackled by molecular simulations is the investigation of molecular interactions playing a crucial role in fairly all chemical processes. Scientists from various areas of applied and life sciences including medicine, pharmacy, chemistry, biology, and material sciences are eagerly interested in a reasonable quantification of interactions between molecules.[1,2] In fact, it is due to molecular interactions that life in the form we know it exists. The entire metabolism and morphogenesis of living cells from the very first contact of two gametes through to a matured organism and beyond is quintessentially determined by interatomic and intermolecular forces. Comprising way more than 10^5 sorts of small molecules and macromolecules, the extremely sophisticated dynamic and kinetic arrangement of human compounds is primarily triggered by binding affinities between them. In vertebrates and other highly developed organisms, these reversible association–dissociation processes particularly include signal transduction through hormones and neurotransmitters triggering some biochemical reaction, chemical modifications of small compounds or proteins by enzymes, gating small molecules or ions through cell membranes, and the elimination of exogenous substances/antigens by antibodies.[3,4] What all those processes have in common is the involvement of a *host* or *target molecule* that is typically a protein and some small compound such as a hormone, neurotransmitter, substrate, or other ligand which we will term *guest molecule*. Especially during drug discovery, one is frequently interested in the determination of binding affinities for such *host-guest complexes* arising from non-covalent association of the two components. That is because pharmaceutical drugs are often intended for high affinity binding to particular target proteins either in order to activate their biological function (acting as an *agonist*) or to suppress it (*antagonistic* effect) where the latter type of drugs are often categorized as *inhibitors*. It's in particular the inhibiting effect on the binding of endogenous (natural) agonists that affect human metabolism and lead to undesirable adverse effects.[5] Therefore and in the light of development costs in the order of one billion US-$ for new drugs[6,7] one is, prior to chemical synthesis, heavily interested in as much foreknowledge about potential molecular interactions with human target structures as possible. At this point, computational methods come into play and the whole story this thesis is about indeed

concerns the prediction of binding affinities for host–guest systems with the help of computational tools also referred to as *in silico* methods. At a very early stage of target-based drug discovery denoted as *hit identification*, possible *lead compounds* are filtered from large digital libraries of small chemicals using rapid and, therefore, approximative high-throughput *virtual screening* techniques. Upon subsequent *lead optimization*, the resulting set of hits is further optimized and narrowed down to one or few candidates for clinical development. This phase certainly requires much more accurate computational methods that produce more reliable results at the cost of time. Although computational methods considerably save time and money needed for laboratory experiments, the entire drug development process usually lasts a couple of years.[8] An as accurate as possible estimation of host–guest binding affinities is not only strived after by the pharmaceutical sector. For human and ecotoxicologists, *in silico* techniques often provide a convenient approach to toxicity assessment in addition to *in-vitro* and *in-vivo* methods which are generally more expensive regarding time and money.[9] This particularly concerns the vast number of *transformation products* arising from biotic or abiotic degradation of anthropogenic substances. Though many of them are detected by experimental methods, they are often not available for intensive risk assessment by means of laboratory experiments due to lacking synthesis protocols.[10]

In terms of thermodynamics the host–guest binding affinity is related to the difference ΔG in *Gibbs free energies* associated with two more or less distinct states of the system under investigation, namely the bound and the unbound one (see Figure 1.1). Later, we will see that ΔG is composed of an enthalpic and entropic term. The former contribution represents changes in the system's inner energy that is energies due to atomic motions and interactions. Temperature-dependent entropic contributions, in contrast, quantify the loss of conformational flexibility upon binding.[11] As already indicated, the pallet of common methods for binding affinity estimation that we will discuss in detail in the next chapter range from very fast but less accurate similarity-based regression models and scoring functions up to as accurate as computationally demanding physical methods requiring a large number of *molecular dynamics* (MD) or *Monte Carlo* (MC) simulations.[12,13] Thus, for the choice of a proper algorithm one generally needs to

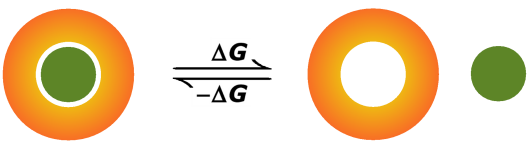


Figure 1.1: Schematic illustration of a reversible and non-covalent association/dissociation reaction of a ligand (green) and target (red) molecule associated with binding energy ΔG.

trade off computational costs against accuracy. Most strategies including particularly the accurate ones try to calculate an estimate for those thermodynamic quantities which is, for several fundamental reasons, a highly nontrivial task. On the one hand, more or less accurate physical models referred to as *force fields* are required for the quantification of atomic interactions due to repulsive or attractive forces. These models can be arbitrarily fine-grained starting form quantum mechanical (QM) *ab-initio* methods that directly solve Schrödinger's equation by taking into account the influence of every single electron. However, even with several semi-empirical approximations including the *density functional theory*, a quantum mechanical representation of solvated biological systems consisting of thousands to millions of atoms such as a typical membrane protein depicted in Figure 1.2 is absolutely impractical. [14,15] Not until an entirely classical treatment of biological systems it was possible to mimic molecular dynamics

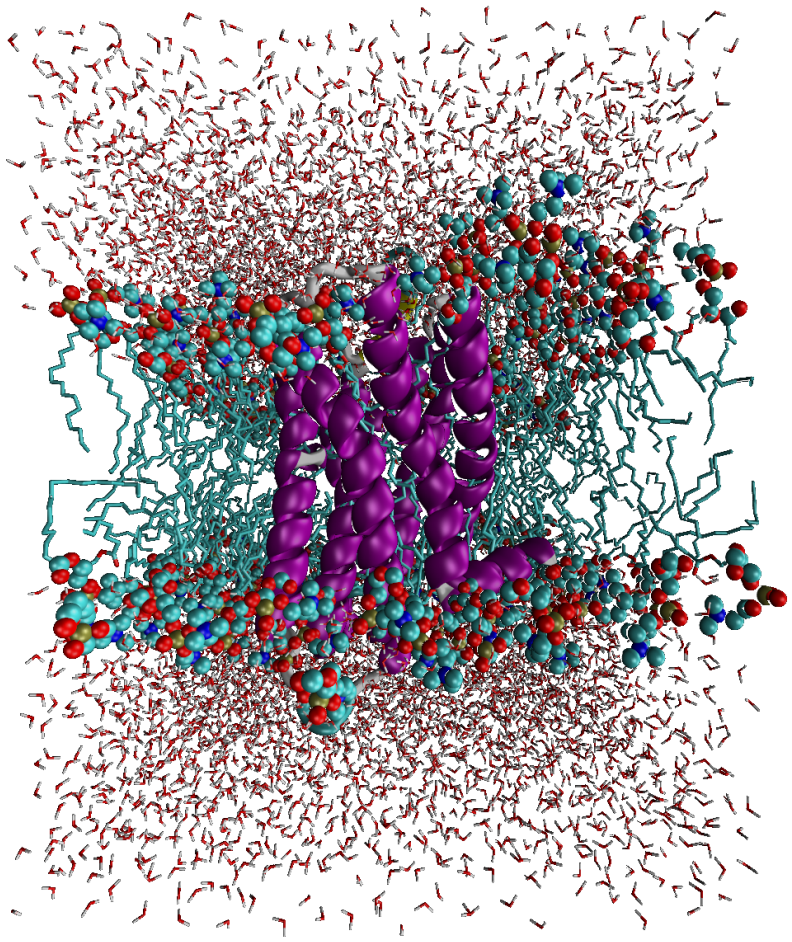


Figure 1.2: VMD snapshot of a rectangular explicit solvent simulation box containing approximately 10^5 atoms: the μ-opioid receptor (represented by violet α-helical secondary structure elements) embedded in a horizontal cell membrane (blue fatty acid carbons with red oxygen atoms).

and calculate thermodynamic energies within reasonable time. The term *classical* is attributed to Newton's classical mechanics which, contrary to QM, builds the fundament for classical molecular mechanics force fields and MD simulations (discussed in detail in Chapter 3). In the classical approximation, electrons are not considered explicitly anymore but represented implicitly in the form of covalent bonds and partial charges that are assigned to atomic nuclei.[15] Analogously, any type of interaction between particles (nuclei) is modelled using preferably simple additive potential functions comprising predetermined parameters provided by an empirical force field.

The first attempts of MD simulations by the end of the 1950s are attributed to Alder and Wainwright on the repulsive interaction of hard spheres modelled by discrete functions.[16,17] Only few years before, Metropolis and co-workers had already developed and published the first Monte-Carlo (MC) approach to the simulation of hard spheres which, in contrast to the deterministic MD method, performs some random walk through conformational space.[18,19] These calculations were carried out on the MANIAC I computer hosted at Los Alamos National Laboratory. Over decades, the molecular systems under consideration grew, potential functions became continuous (and more time-consuming)[20] and charge potentials[21] as well as condensed phase systems including macromolecules[22,23] came into play. Until today, classical MD simulations of complex macromolecular systems surrounded by explicit water molecules have

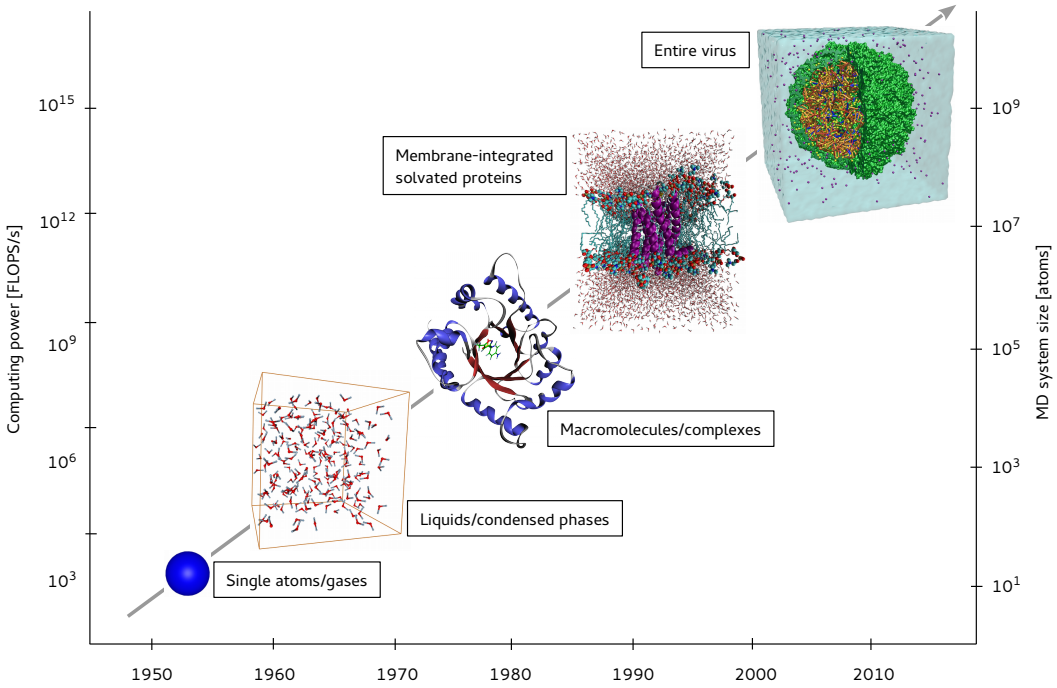


Figure 1.3: Size of MD systems and high performance computing power vs. time. Virus image taken from *www.ks.uiuc.edu*

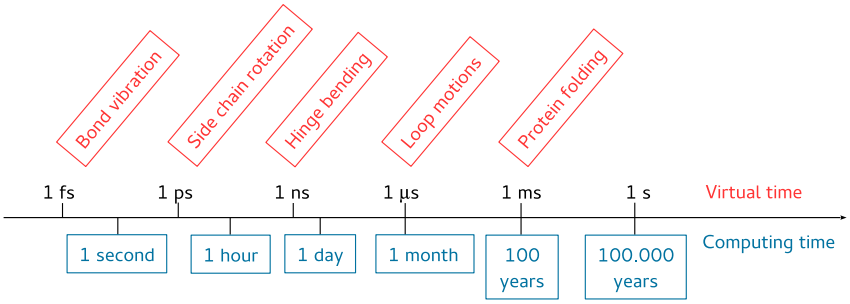


Figure 1.4: Time scale problem of MD simulations and periods of typical molecular oscillations on the basis of biological systems.

evolved into ordinary day-to-day routines. The dynamics of protein–protein, protein–DNA complexes, and protein-protein complexes embedded in cell membranes is easily investigated in particular due to the progress and interplay of parallel computing strategies, intelligent algorithms and fast processors.[23,24] Figure 1.3 shows the performance development of supercomputers measured in floating point operations per second (FLOPS) and the size of typical MD systems over the past six decades. Unfortunately, there is a technical limit to the degree of parallelization of MD simulations since corresponding software inherently does not scale linearly. If the calculation of a molecular system is distributed over too many computing cores, additional CPU time due to increased communication between the cores referred to as *overhead* will, from some number of cores on, outweigh the gain of time through parallelization. As a consequence and despite of all the sophisticated enhancing techniques available today, the real computational time required for the classical MD simulation of a virtual second of typical biological systems amounts to thousands of years! The time-scale problem associated with deterministic MD simulations is illustrated by Figure 1.4 on the basis of a macromolecular 200 k atoms system (to a great extent consisting of explicit space-filling solvent molecules) MD run using 24 modern Intel Xeon Haswell computing cores. The computational effort is mainly due to a tiny time step size in the order of 10^{-15} s $=$ 1 fs used by the numerical MD integrator. This is a substantial requirement for a sufficient sampling of the fastest molecular oscillations as are related to chemical bonds. Consequently, 10^{10} to 10^{15} are necessary in order to observe considerable conformational changes in proteins (Figure 1.4). Applied to distances, that difference in time scales is comparable to recirculating our solar system by making steps of 1 m size. However, our simulations used for binding free energy estimations in upcoming chapters will reflect virtual time in the order of 0.1 to 1 ns which is sufficient for reasonable energies averaged over side chain rotations. And indeed, such torsional processes are mainly responsible for the conformational fit of a host–guest complex. This brings us to the first of two related reasons why a rigorous calculation of free energies

actually implies a spacious sampling of the entire conformational space.[25–27] Bound and unbound states are neither static nor clearly distinct but rather statistical entities connected through a conversion process termed *induced fit*. The underlying molecular recognition mechanism is characterized by mutually induced conformational changes of the involved molecules upon binding.[28,29] Further, thermodynamic free energy differences as determined by biological assays are related to distributions of a vast number (typically in the order of 10^{23}) differing host–guest molecular geometries (conformations) which are present simultaneously.[15,24] It seems therefore reasonable to estimate binding free energies statistically on the basis of properly distributed *ensemble* averages of *microstates* instead of relying on a single geometry. Certainly, the question arises of how many samples would be necessary in order to have the *entire* conformational (some would say *configurational*) space covered. The answer brings us immediately to the next, by means of computer simulations, challenging property of biological systems which is related to the huge number of atoms. Not only does a system consisting of N atoms require – for every single time step – a time-consuming computation of $3N$ partial derivatives for conformational sampling purposes and about N^2 force field evaluations for potential energy calculation if the interaction between any pair of two atoms is considered.[24] Apart from the complexity of the mathematical space it is in particular the combinatorial number of possible conformers that drastically increases. For the sake of illustration, consider the linear molecule *n*-butane consisting of four serial carbon (indexed with 1-4) and ten hydrogen atoms depicted in Figure 1.5. As with any other single bond that is not strictly embedded in a cyclic structure, this molecule is able to rotate around (among others) the central C_2–C_3 bond resulting in various conformers due to differing dihedral angles which are associated with different potential energies U. The dihedral angle ϕ also referred to as torsion angle formed by four consecutively connected atoms $\{C_1,\ldots,C_4\}$ serves as a measure for the rotational degree. In the context of molecules, it corresponds to the angle between two planes which are defined by $\{C_1, C_2, C_3\}$ and $\{C_2, C_3, C_4\}$, respectively. The absolutely lowest energy conformer denoted as *anti* with $\phi_{anti} = 180^\circ$ (sketched in colors at the top of Figure 1.5) is attributed to a maximum distance between the two terminal methyl groups accompanied by minimum mutual steric hindrance. In contrast to the highest energy geometry indicated by the Newman projection in the upper right corner with $\phi \approx 0^\circ = 360^\circ$), the anti conformation would correspond to the staggered case with the methyl group in the foreground pointing downward. Since molecules tend to occupy states with low energy, *anti* happens to be the most probable conformation. Due to two further staggered conformations both denoted as *gauche* ($\phi_{gauche} = 180^\circ \pm 120^\circ$) and associated with slightly higher potential energies than the *anti* rotamer itself, the potential energy

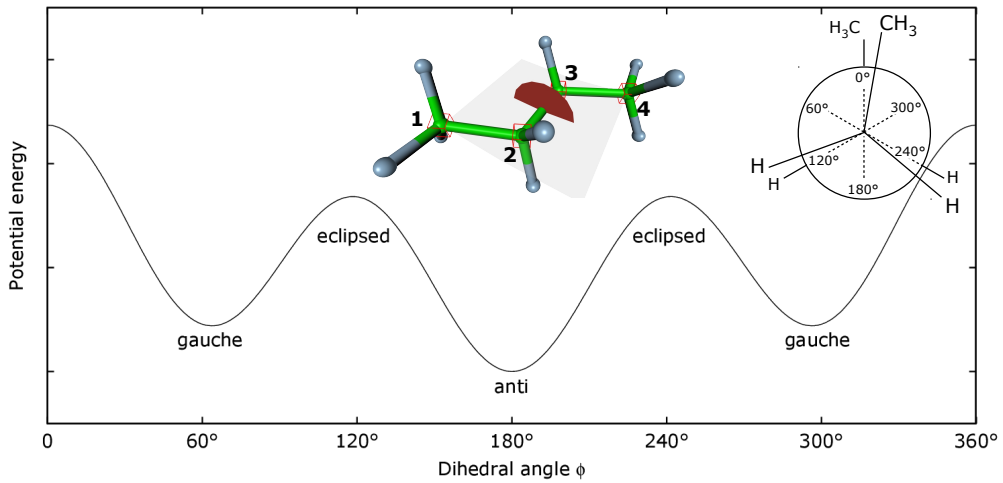


Figure 1.5: Potential of butane depending on the all-carbon dihedral angle.

surface (PES) projected on the central butane dihedral angle exhibits three energetically favorable conformations within 2π which are seperated by relatively high energy barriers.[30] Free energy calculations of small molecules such as butane are feasible nowadays. Instead, imagine a molecule consisting of thousands of rotatable bonds. Already the backbone atoms alone of a medium-sized protein consisting of 500 amino acid (AA) residues form 1000 relevant dihedral angles. Theoretically, this yields a combinatoric diversity of 3^{1000} or circa 10^{477} local minima on the thousand-dimensional PES defined on the protein backbone's set of ϕ and ψ angles. This mathematical problem also known from other fields than molecular simulation is often referred to as *curse of dimensionality*. Figure 1.6 illustrates the concept of these two dihedral angles using a tripeptide consisting of AA alanine by way of example. The overall secondary and tertiary structure of proteins is sufficiently described by its ϕ and ψ angles. Besides, in our complexity calculations we have neglected rotatable dihedrals of AA side chains and the fact, that both ϕ and ψ angles exhibit even more than three minima upon a complete cycle.

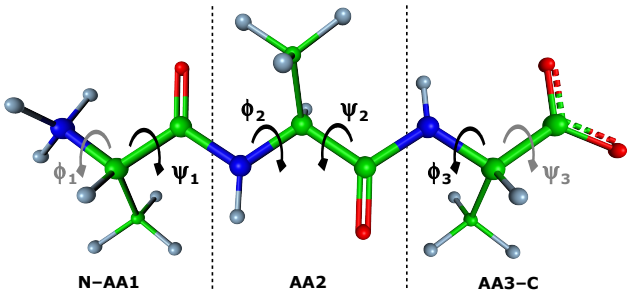


Figure 1.6: ϕ/ψ dihedral angles of protein and peptide backbones illustrated using trialanine peptide.

Despite the issues discussed above, several more or less accurate strategies (discussed in Chapter 2) allowing for the estimation of free energies and biochemical binding affinities have been developed.[26,27] What they have in common is that only a tiny portion of space representing the most favorable host–guest complex is taken into account. On that note, it is a matter of vital importance to have the native macromolecular 3D-structure in the main available. Various methods developed for template-based or *de novo* prediction of secondary and tertiary protein structure[31–35] certainly suffer from the mathematical complexity as well and do not often yield sufficiently accurate results.[36,37] As a general rule, calculated binding affinities are the more reliable the less approximations and predictions have been incorporated. Fortunately, more than 10^5 3D protein structure files resolved experimentally by X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy[38–40] are available online at the Protein Data Bank[41] (PDB) organized by the Research Collaboratory for Structural Bioinformatics. Many entries of the PDB include structural information about bound ligands, revealing host–guest binding modes and target binding sites. Without these specifications, binding affinity estimations for protein–ligand systems would basically become a funny guessing game without any award. Structural data of PDB does not only substantially increasing the reliability of free energy estimations but is in turn useful for the development of force field parameters and gives insight into structural properties and molecular interactions. Figure 1.7 exemplarily shows the procaryotic enzyme dihydropteroate synthase (DHPS) in complex with a biphosphorylated pteridine derivate on the left (PDB entry 1AJ2) and with its exzymatic product dihydropteroate on the right (PDB entry 1TX0) which resulting from the replacement of bisphosphate by *para*-aminobenzoate

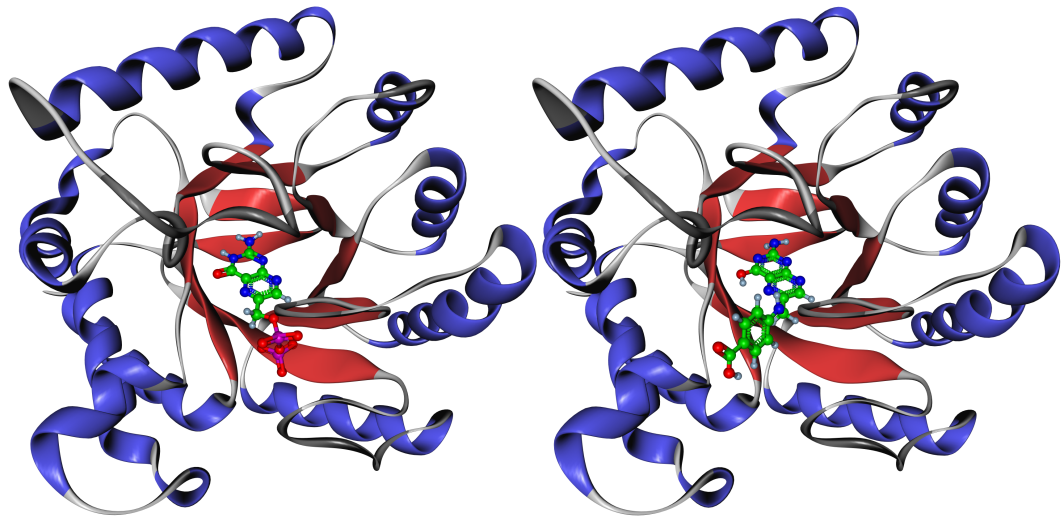


Figure 1.7: Secondary structure of the bacterial enzyme dihydopteroate synthase (PDB entry 1AJ2) complexed with a substrate (left) and product (right, taken from PDB entry 1TX0) molecule.

(PABA).[42] Since dihydropteroate serves as a precursor of the essential molecule folic acid (vitamin B_9), the antibiotic sulfamethoxazole (SMZ) can be exerted in case of certain bacterial infections as a competitive antagonist of PABA in order to suppress its synthesis.[43] We will further devote ourselves to this enzyme and further potential agonists in Chapter 7.

Motivation and outline

This thesis primarily deals with the development of empirical models for the prediction of binding affinities for various host–guest systems and purposes. Particular emphasis is put on physical relevance and balanced trade-off between model accuracy and speed. However, in order to achieve somewhat reliable results, a considerably larger value is attached to accuracy. For these reasons, a method meeting these criteria and referred to as *linear interaction energy* (LIE) provides the basis for all predictive models presented in the following. Prior to specific applications, Chapter 2 summarizes the entire theoretical background in terms of phenomenological as well as statistical thermodynamics and gives an overview over the state-of-the-art technology of average-based thermodynamic binding affinity prediction. Afterwards, the theoretical basis of molecular modeling algorithms used for molecular mechanics, MC, and MD simulations as well as for the adaptation of physical and geometric constraints are elaborated in Chapter 3. In a first application described in Chapter 5, we will investigate whether the LIE method originally designed for protein–ligand systems in combination with classical force field MD simulations is practicable for chromatographic problems as well. Using the challenging example of a brominated cyclic flame retardant we will try to derive the elution order and retention times of its six highly similar stereoisomers associated with high-performance liquid chromatography. In Chapter 6 and using by way of example the human estrogen receptor α, an extension of the originally two-parameter LIE model is developed and evaluated. Among other modifications, the influence of various MD settings and a recently published Monte Carlo approach to conformational entropies is investigated. The final model will be compared to other state-of-the-art free energy estimators. In a final application (Chapter 7) we will investigate whether the empirical LIE model which originally needs a training set of molecules with known binding affinities for parameter estimation is applicable to a toxicological prioritization of transformation products (TP) *without* having been parameterized before. This is an important aspect of this work since an accurate risk assessment of TPs emerging from abiotic or biotic degradation of anthropogenic substances is more and more coming into view after having been neglected for many decades. One of the very crucial steps of the entire process

of binding affinity estimation is related to the selection of one or more representative binding modes. According to the time scale problem, no significant changes in the (relative) host–guest conformation can be expected during MD simulations of reasonable duration. In Chapter 4 we will therefore develop and evaluate strategies for a systematic decomposition and exploration of the conformational space followed by a proper choice of one or more representatives for free energy calculations. All in all, we want to develop a fully automatized work flow that returns the desired affinity given a target molecule along with the specification (spatial vector) of its binding site and a ligand molecule under observation.

2 Theory and state of the art of binding free energy calculations

In the following we will engage with existing methods for the purpose of binding affinity (or, synonymously, binding free energy) estimations. This is an as central as challenging task in computational biology and medicinal chemistry since molecular interactions of proteins and ligands play a decisive role in biological functions and reactions including enzyme catalysis and intracellular signal transduction. Hence, it is of great interest for drug designers and toxicologists to be able to accurately predict host–guest binding affinities. Computer-aided drug development substantially reduces the need for time and money in drug discovery.[9] Though the structure of novel chemicals such as biotic or abiotic transformation products (discussed in more detail in Chapter 7) may be determined by spectroscopic methods, amounts sufficient for toxicity tests often cannot be synthesized due to lacking protocols. In such situations, *in silico* methods come into play as the only access to an critical assessment on novel substances allowing a preliminary prioritization for experimental investigations. The lack of an appropriate reference state prohibits the calculation of *absolute* free energies. And even if we could in theory, there are still practical difficulties since MD and MC simulations are not able to sample the entire phase/conformational space (in particular high energy regions) of a molecular system in reasonable time. However, relative free energies (free energy differences) with respect to two slightly different systems or system states can be estimated due to a variety of methods.[44]

The methodological pallet for the task of affinity estimations ranges from exhaustive thermodynamic perturbation methods demanding large sets of expensive MD trajectories up to substantially faster but generally less accurate quantitative structure-activity relationship (QSAR) strategies and simple scoring functions as implemented in docking programs. Obviously, the computational expense corresponds with the extent to which the model copes with physical principles since accurate MD-based calculations require the evaluation of a huge number of (pairwise) interatomic forces related to target as well as ligand atoms. In contrast, QSAR methods are generally ligand-based and do require neither MD simulations nor the consideration of the target molecule. The choice of

method depends on at which stage of the drug discovery process it will be employed and there is typically a trade-off between accuracy and efficiency.[45] However, before elaborating common approaches to binding affinity calculation, we will, in what follows, first outline its physical basis including classical and statistical thermodynamics. Molecular mechanical/dynamical aspects as well as algorithms applied to the modelling and sampling of host-guest systems investigated throughout this thesis are discussed in the subsequent Chapter 3.

2.1 Thermodynamic background of binding affinities

Thermodynamics deals with energetic processes in a macroscopic, i. e., experimentally measurable sense. In particular, these processes include the role of heat and temperature regarding energy, work and the transformation of one form of energy into another. Its origin is tightly interwoven with the development and progress of steam engines during the first decades of the 19th century which transformed heat into mechanical work[46] at a very low efficiency. For the investigation of system properties it is necessary to define its boundary permeability regarding energy, heat, and matter. Commonly used thermodynamic systems are:

Isolated No interaction with surroundings at all.

Closed No exchange of matter, but of work and heat.

Adiabatic No exchange of heat and matter, but of work.

Open No limitation on exchange.

In terms of phenomenological thermodynamics, the current state of a physical system is sufficiently described by a set of thermodynamic parameters denoted by *state variables* or *state functions* which are *path-independent*, that is to say they do not depend on the path having yielded the current state, but only on the given state itself. These state variables are divided into *extensive* (e. g. various energy quantities, entropy, volume, number of particles) that depend linearly on the system size and *intensive* state variables (e. g. temperature, density, pressure) that do not change with the system's size. This characteristic makes intensive variables suitable for the comparison of thermodynamic systems and the investigation of equilibrium properties. Whether a thermodynamic change or chemical reaction requires energy input or evolves spontaneously releasing energy and questions regarding the location of chemical equilibria can be answered following the four principle laws of thermodynamics.[1] The zeroth law introduces the

temperature T as the state variable that is equal for two systems if they are at thermal equilibrium. The first law is derived from the law of energy conservation in isolated systems and states that energy can neither be destroyed nor created. According to the *fundamental thermodynamic relation*

$$\mathrm{d}U = \delta Q + \delta W = T\,\mathrm{d}S - p\,\mathrm{d}V + \sum_i \mu_i\,\mathrm{d}N_i \ , \tag{2.1}$$

the system's internal energy $U\,(S, V, N)$ as a function of the entropy S, the volume V, and the number $N = \{N_i\}$ of particles of type i can only change due to an exchange of heat Q or work W across system boundaries. W might, for instance, refer to pressure-volume work $p\mathrm{d}V$, a change in volume caused by a pressure p, or chemical reactions changing the amount N_i of some particle type i associated with the chemical potential μ_i.[47] In contrast to state variables, heat Q and work W fall in the category of *path-dependent* variables since their values depend on the path having leaded to a given state. In a classical manner, U consists of the kinetic energy $K = f\,(p)$ related to the motion of bodies or, to be precise, on their momenta p including undirected and temperature-dependent *Brownian* motion and the potential energy $V = f\,(r)$ depending on the particles' spatial arrangement (atomic coordinates r) and describing external forces acting on them due to interactions with each other or with external *force fields*. Considering the right hand side of Equation 2.1, it appears that each product the change of the thermodynamic potential $U\,(S, V, N)$ is composed of incorporates one extensive state variable multiplied with its *conjugated* intensive state variable such that the change of the internal energy U can be expressed by differential changes in its (extensive) variables. Biological association reactions, however, are generally not entirely isolated from external influences (e. g. heat). Thus, further thermodynamic potentials with different sets of natural variables are derived from Equation 2.1 by a coordinate transformation approach denoted as *Legendre transformation* using first derivatives of U with respect to its extensive variables. These thermodynamic potentials completely describe a physical system at its thermodynamic equilibrium where – once it is reached – all state variables become constant. Depending on the underlying physical boundary condition, this *stationary state* corresponds to the minimum of the respective thermodynamic potential and a maximum entropy.[1] A common Legendre transformation (along with their differential form) derived from U and useful as theoretical basis for MD simulations of biological systems is the *Helmholtz free energy*

$$A = U - TS; \quad \mathrm{d}A\,(T, V, N) = -S\,\mathrm{d}T - p\,\mathrm{d}V + \sum_i \mu_i\,\mathrm{d}N_i\,.$$

It describes equilibrium conditions of closed systems with constant temperature T in addition to constant particle numbers N_i and volume V. It follows from the second law

of thermodynamics that this potential's change $\mathrm{d}A < 0$ unless arriving at equilibrium where $\mathrm{d}A = 0$ and A reaches its minimum. If, in addition to the temperature, pressure p as a second intensive state variable becomes a characteristic variable, we arrive at the *Gibbs free energy* (*free enthalpy*)

$$G = U + pV - TS; \quad \mathrm{d}G\,(T, p, N) = -S\,\mathrm{d}T + V\,\mathrm{d}p + \sum_i \mu_i\,\mathrm{d}N_i \tag{2.2}$$

as a further thermodynamic potential specified by constant (T, p, N) and its minimum at equilibrium. A deviates from G by the pressure-volume work pV solely and they become equal if no pressure-volume work is done, $pV = 0$. This property is as well of relevance for host–guest complexes solvated in incompressible fluids such as water. There exist further thermodynamic potentials including the enthalpy $H\,(S, p, N)$ and particularly the entropy. A system tends to, according to the second law of thermodynamics, maximize its entropy and more entropy emerges at increasing temperatures, since the system is able to occupy more energy levels and thus, more microstates. Accordingly, the temperature-dependence of the entropy is formulated by the third law of thermodynamics stating that upon approximating absolute zero, the entropy approaches zero, $\lim_{T\to 0} S = 0$. Usually, fundamental state functions cannot be determined directly by experiment. However, combining their first and second partial derivatives with results of experimentally measurable quantities like pressure, temperature, volume, etc. permits the determination of their values.[47]

From a thermodynamic point of view, the *binding affinity* of a molecular host–guest system under atmospheric conditions (constant pressure and volume) is defined by a macroscopic quantity denoted as *Gibbs free energy of binding* ΔG. As sketched in Figure 1.1, the Δ symbol is related to the system's energy change upon association (dissociation) of the molecular complex under consideration

$$\Delta G = G^{\text{bound}} - G^{\text{unbound}}. \tag{2.3}$$

Basically, the binding free energy is composed of enthalpic (ΔH) and temperature T dependent entropic $(T\Delta S)$ contributions

$$\Delta G = \Delta H - T\Delta S \tag{2.4}$$

and becomes minimal at thermal equilibrium.[1] Hence, if the binding of some ligand L to a protein P is considered as a chemical reaction

$$\mathrm{P} + \mathrm{L} \underset{k_{\text{off}}}{\overset{k_{\text{on}}}{\rightleftharpoons}} \mathrm{PL}$$

with the on-rate k_{on} and off-rate k_{off} for forward and, respectively, backward reaction, its *binding constant* as a special case of the *equilibrium constant* derived from the law of mass action and valid for molecular systems at chemical equilibrium is defined as

$$K_a = \frac{1}{K_d} = \frac{k}{k'}.$$

Equilibrium constants associated with the subscripts "a" and "d" are called *association* and *dissociation constant*, respectively, and quantify, in terms of concentration (indicated by squared brackets) ratios the extent to which the system at equilibrium ("eq") is dominated by the bound or the unbound state[48]

$$K_a = \frac{[PL]_{eq}}{[P]_{eq}\ [L]_{eq}} \tag{2.5}$$

Binding constants, consequently referred to as binding affinities, are directly related to the Gibbs energy difference according to

$$\Delta G = \Delta G^\circ - RT\ \ln \frac{[PL]}{[P]\ [L]} \tag{2.6}$$

where R stands for the gas constant and ΔG° denotes the standard change of the reaction in Gibbs free energy that is the energy associated with the reaction under standard conditions (defined temperature, pressure, and educt concentrations). At equilibrium, i. e., when there is no net reaction flow ($\Delta G = 0$), Equation 2.6 becomes

$$\Delta G^\circ = -RT\ \ln K_a = RT\ \ln K_d \tag{2.7}$$

after little rearrangement and the quotient of concentrations (now at equilibrium) can be substituted by an equilibrium constant such as K_d. In the special case of reactions where exactly *one* molecule of each of the two components form one complex (such as the association/dissociation of a protein-ligand pair), K_d equals that ligand concentration at which half of the total number of protein molecules is complexed. It should be noted that the K_d value must be dimensionless in order to take its logarithm. Indeed, the unit of K_d vanishes due to division of concentrations by a reference concentration of usually one mol per liter.[48] Besides, uncertainties in the experimental binding free energies are typically at least 2.0 kJ/mol and K_d values are often derived from IC_{50} values using the Cheng-Prussoff relation.[49] In common literature dealing with binding affinity estimations, one often encounters the equation

$$\Delta G = RT\ \ln K_d \tag{2.8}$$

incorporating ΔG instead of ΔG° that is used synonymously to Equation 2.7 although no standard conditions were given during the *in silico* experiment. We will follow the notation depicted in Equation 2.8 throughout this thesis.

2.2 Statistical thermodynamic background

Statistical thermodynamics also referred to as statistical mechanics can be considered as a bridge between experimental observations on a *macroscopic* level described by phenomenological thermodynamics on the one side and the underlying molecular processes on a *microscopic* level on the other. It is used to derive macroscopic quantities of large populations (typically in the order of $\mathcal{O}\left(10^{23}\right)$ states) of a molecular system from a set of individual *microstates* that are sufficiently described by Cartesian coordinates and velocities of all system particles. With the aid of statistical mechanics, thermodynamic *state variables* (*state functions*) are predicted in a statistical manner by investigating the molecular mechanics and probability distribution of an ensemble of microstates.[1,50]

Conformational space and potential energy surface

Consider a many-body system consisting of N discrete atoms. Any spatial geometry (conformation or configuration) of the entire system can be described by a vector $r_i = (x_i, y_i, z_i)$ of Cartesian coordinates per atom i. Putting all Cartesian coordinates of the entire system into one variable yields a single vector $r \in \mathbb{R}^{3N}$ specifying the given geometry as a single point

$$r = (r_1, \ldots, r_{3N})^\top$$

in the *conformational space*[1] which becomes exceedingly high-dimensional when considering macromolecular systems along with explicit solvent molecules where $N > 10^4$. For reasons of practicability and simplification regarding the solution of equations of motion, one got into the habit of using a mathematical transformation of r yielding *generalized coordinates* q and possibly a substantial reduction of *dimensionality* $d = 3N$.[47] q might, for instance, include distances between atoms or angles spanned by them also known as *internal coordinates* that are independent from an external coordinate system and its origin.[24] Given a system consisting of two atoms, $N = 2$, of which we are interested in the distance only, the dimensionality can be reduced from $d = 3N = 6$ degrees of freedom to $d = 1$ by neglecting three translational and two rotational degrees of freedom without having influenced the systems internal state or energy. In general, a system's conformation/configuration as well as its potential energy are invariant under translation and rotation such that six external degrees of freedom associated with translations in three dimensions and rotations in three (or two in the special case of two

[1] In statistical mechanics, one would use the term *configurational space* instead which can be misleading in a macromolecular sense with covalently bound atoms where, in contrast to the *conformation*, the molecular *configuration* cannot change during a dynamic process.

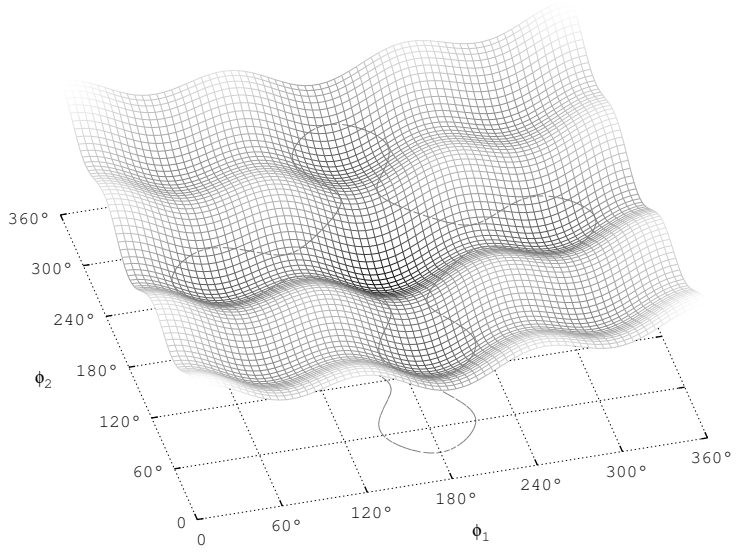


Figure 2.1: Potential energy surface of pentane reduced to its two all-carbon torsional angles ϕ_1 and ϕ_2 and a contour plot at a particular energy level.

atoms, $N = 2$) dimensions can be neglected for many problems. However, generalized coordinates may also comprise Cartesian coordinates of each particle, $q = r$, in particular, if no reduction is expected or possible. Assuming this,

$$q = (q_1, \ldots, q_{3N})^\top$$

constitutes a single point in the Cartesian configurational space representing a single conformation of the molecular system under consideration. Upon conformational changes of the dynamic system over time t, the point q moves through configurational space accordingly forming a time-dependent curve in $\mathbb{R}^{3N}$ referred to as *trajectory* or *time series* $q(t)$. Each geometry q of a molecular system is associated with a potential energy $V(q)$ which is related to forces acting on the particles. The underlying interactions originate from external force fields (gravitation, electro-magnetism) or internal repulsive and attractive forces between atoms. In the classical limit of statistical thermodynamics, $V(q)$ of an N-atomic system in $3N - 6$-dimensional configurational space gives a continuous *potential energy surface* (PES) of dimension $3N - 6$ if external forces are neglected. Since an additional dimension is required for the assignment of a potential energy, the hyperplane is located in a $3N - 5$-dimensional space. It can be considered as an energy landscape on top of the space spanned by the system's generalized coordinates. Figure 2.1 shows a two-dimensional PES defined neither on the $3N = 51$ external nor on the $3N-6 = 45$ internal but only on the two all-carbon dihedral angles of pentane which consists of $N = 17$ atoms. Neglecting all bonds and bond angles (keeping them constrained) is an accepted approach to space reduction in the

field of molecular modelling since the conformation and physico-chemical properties of pentane and in particular of organic molecules with a significantly higher complexity are mainly attributed to their torsion angles. The energy landscape reveals several local minima and saddle points that are helpful for the identification of favorable (stable) conformations and, respectively, of transition states, reaction paths, and reaction rates. The lowest energy and, thus, most likely conformation of pentane located in the middle of the PES is associated with twice *anti* (180°) regarding both torsion angles.

Phase space and ergodicity

A trajectory $q(t)$ in configurational space may reveal intersections indicating identical molecular geometries that emerged from different paths. In order to be able to differentiate between these geometries and obtain unique states, it is necessary and sufficient to additionally consider the coordinates' *conjugate momenta* $p \in \mathbb{R}^{3N}$

$$p = (p_1, \ldots, p_{3N})^\top$$

associated with the N particles of an instant geometry q. In a classical mechanical sense, a thermodynamic microstate $x \in \mathbb{R}^{6N}$

$$x = (q, p) = (q_1, \ldots, q_{3N}, p_1, \ldots, p_{3N})^\top$$

is now completely and uniquely specified by a point (q, p) in the $6N$-dimensional *phase space*.[51] Considering the time-evolution

$$x(t) = (q(t), p(t))$$

of *deterministic* phase space trajectories, one should be aware of the fact that two of them with distinct initial conditions $x(0) = (q(0), p(0)) \in \mathbb{R}^{6N}$ will *never* intersect in phase space. A phase space trajectory can be constructed using a computer simulations on the basis of a convenient mechanistic framework (discussed in Chapter 3) starting from an initial set $x = (q, p)$ of atomic coordinates and conjugate momenta. It is therefore, theoretically, possible to predict an isolated system's deterministic evolution (particle motion) in future as well as back in time. In practice, though, the correlation between time steps will decrease quickly in time due to technical (machine precision) and numerical (condition and stability) reasons besides many other approximations regarding molecular modeling. Slightly varying initial conditions $x(0)$ are likely to yield substantially diverging trajectories. As a consequence, we cannot expect long term *molecular dynamics* (MD) simulations to provide us with useful informations for the investigation of slow molecular processes such as receptor–ligand binding events

or considerable conformational changes in a sense of in which direction the system might evolve. All the more they are suitable for the statistical estimation of observable averages. In contrast to an ordinary MD simulation, a typical macroscopic measurement of a physical quantity O in the laboratory comprises a number of system copies in the order of the *Avogadro* constant $\mathcal{O}\left(10^{23}\right)$ yielding an average result for the underlying statistical ensemble. Indeed, the basic idea behind MD simulations is that an experimental *ensemble average* $\langle O \rangle$ can be approximated by some *time average* $\overline{O}\left(x\left(t\right)\right)$ of a long time series considered as a *statistical ensembles*,[52]

$$\lim_{t \to \infty} \overline{O}\left(x\left(t\right)\right) = \langle O \rangle. \tag{2.9}$$

This assumption strongly relies on the *first postulate of statistical thermodynamics*, which is related to the *ergodic hypotheses* and states for an isolated many-body system with constant total energy E that sooner or later every microstate in phase space will be occupied if the system is observed for a sufficiently long time.[47]

Microcanonical ensemble

Statistical ensembles of microstates are always associated with a characteristic set of physical boundary conditions. A statistical ensemble of an isolated system specified by a constant number of particles N, constant volume V, and in particular, constant total energy E is denoted as *microcanonical* or NVE ensemble. Generally, a vast number of different microstates or arrangements of a many-body system correspond to the same internal energy E at a given set of (N, V), a property that is referred to as *degeneracy* of energy in quantum mechanics where energy is considered as a quantized quantity. For the degeneracy of an NVE ensemble, i. e., for the number of possible microstates with energy E we will use the quantity Ω. Since each microstate i of an NVE ensemble is associated with the same magnitude of energy, the *second postulate of statistical thermodynamics* stating that all microstates are equally likely to be occupied

$$P_i = \frac{1}{\Omega} \tag{2.10}$$

is fulfilled. As a consequence, the system under consideration will most likely adopt the energy distribution with the largest number of microstates if more than one possible distribution exists which holds for many-body systems in general. In analogy to the second law of thermodynamics, this physical property is also referred to as *maximization of entropy*. From a microscopic perspective, the *statistical entropy*

$$S = -k_{\mathrm{B}} \sum_{i=1}^{\Omega} P_i \ln P_i = k_{\mathrm{B}} \ln \Omega \tag{2.11}$$

can be expressed on the basis of the number Ω of microstates i and their probabilities P_i [47] where the *Boltzmann constant* $k_\mathrm{B} = 1.38066 \times 10^{-23}\mathrm{J/K}$ serves as a proportionality factor named after the Austrian physicist Ludwig Boltzmann, the (co-)founder of statistical mechanics and in particular of the microscopic entropy formulation. Besides, except for the logarithm's basis, the formulation is equivalent to the *Shannon entropy* known from information theory. Basically, it is the relation between the number of available states and the entropy stated by Equation 2.11 that justifies the approach to conformational entropies described in upcoming chapters for the sake of binding affinity estimation.

In contrast to quantum mechanics, classical mechanics treats microstates and energies as continuous quantities such that the definition of a microstate is associated with an energy range $[E, \Delta E]$. Then, the degeneracy Ω is replaced by the classical *microcanonical partition function*

$$Q_{NVE} = \Omega = \frac{1}{\mathrm{h}^{3N}\xi} \int\limits_{\mathbb{R}^{6N}} \mathrm{d}q \; \mathrm{d}p \, \delta\left(\mathcal{H}\left(q,p\right) - E\right). \tag{2.12}$$

measuring the volume of the integral spanned by the system's $3N$ coordinates and $3N$ momenta in the $6N$-dimensional phase space and bounded by the energy hyperplane defined on the subset of feasible states with constant energy E. The *delta function* $\delta\left(\mathcal{H}\left(q,p\right) - E\right)$

$$\delta\left(x\right) = \begin{cases} 1 & \text{if } x = 0, \\ 0 & \text{else} \end{cases} \tag{2.13}$$

makes sure that only microstates with a total energy of some particular value (or range), $\mathcal{H}\left(q,p\right) = E$, are considered by the integration. The total energy $\mathcal{H}\left(q,p\right)$ also referred to as system *Hamiltonian*

$$\mathcal{H} = \mathcal{H}\left(q,p\right) = U\left(q\right) + K\left(p\right). \tag{2.14}$$

is composed of the potential energy $U\left(q\right)$ as a function of atomic coordinates and the kinetic energy

$$K\left(p\right) = \frac{1}{2} p^\top M^{-1} p \tag{2.15}$$

depending on momenta only. [24] We will discuss some special properties of the Hamiltonian particularly useful for MD calculations in Section 3.1. According to quantum mechanics, the expression h^{3N} in Equation 2.12 incorporating the Planck constant h defines a minimal volume element of the phase space consisting of one single microstate insofar as permitted by Heisenberg's uncertainty principle related to the product of

Table 2.1: Physical and statistical properties of common thermodynamic ensembles. Legend: number of particles N, volume V, temperature T, energy E, chemical potential μ, and pressure p.

Ensemble	Variables	Isolation	Thermodyn. potential	Distribution of microstates
Microcanonical	N, V, E	Isolated	Entropy S	Uniformly
Canonical	N, V, T	Closed	Free energy A	$\exp(-\beta E)$ (Boltzmann factor)
Isothermal-isobaric	N, p, T	Closed	Free enthalpy G	$\exp(-\beta\,(E + pV))$
Grandcanonical	μ, V, T	Open	Grandcan. potent.	$\exp(-\beta\,(E - \mu N))$ (Gibbs factor)

coordinates and momenta. It was introduced mainly due to the requirement of expressionlessness by certain mathematical transformations. Likewise in agreement with quantum mechanical requirements is the additional normalization factor ξ

$$\xi = \begin{cases} N! & \text{if consisting of one particle type only,} \\ \prod_{i=1}^{k} N_i! & \text{if consisting of } k \text{ types of particles} \end{cases} \tag{2.16}$$

related to the number of permutations N_i per particle type i and considering the indistinguishability of particles of the same type. The partition function of an ensemble is the central quantity in statistical mechanics from which all thermodynamic quantities such as the entropy in Equation 2.11 as well as the average of any observable O

$$\langle O \rangle_{NVE} = \frac{1}{Q_{NVE}} \sum_{i=1}^{N} O_i$$

can be derived.[47] Nevertheless it should be noted that an as accurate as possible computation of Q for complex many-body systems is a challenging task not to mention impossible even with modern supercomputers since it requires an exhaustive sampling of the entire conformational space. Although the microcanonical ensemble appears idealized in a sense, it serves as a basis for further ensembles that are more convenient for real-world systems. Table 2.1 shows the most popular statistical ensembles along with their degree of isolation, characteristic variables and the thermodynamic potential they are related to. Due to the permeability properties of their physical boundaries regarding heat (and mechanical pressure-volume work), most physical and, in particular, all biochemical systems are associated with constant temperature (and constant pressure) instead of constant energy which is specific only for isolated systems. Closed systems are characterized by a new type of partition function and an uneven energy distribution. In the following, we will engage with such statistical ensembles that are more convenient for free energy calculations of (biological) host–guest systems under atmospheric conditions.

Canonical ensemble

We first consider a thermodynamic constant-volume system that is not able to perform mechanical work due to changes of its volume V but to exchange energy with its environment in the form of heat. If the surrounding reservoir is much larger than the system under observation, it will act as a *thermostat* that keeps the system's temperature T constant either by absorbing its kinetic energy or transferring kinetic energy to it, respectively, depending on which compartment is hotter. Such an exchange process will inevitably result in maximum entropy and thermal equilibrium around which the system Hamiltonian fluctuates with variance

$$\sigma^2 = k_\mathrm{B} T^2 C_\mathrm{V} \tag{2.17}$$

proportional to the isochoric heat capacity C_V.[1] According to the *equipartition theorem* which postulates that the kinetic energy K (Equation 2.15) is approximately equally distributed over the system's degrees of freedom N_f, the temperature is related to the average kinetic energy per degree of freedom by[24]

$$\left\langle \frac{1}{2} m_i v_i^2 \right\rangle = \frac{1}{2} k_\mathrm{B} T \tag{2.18}$$

and the system's internal velocities obey a Maxwell-Boltzmann distribution[53]

$$P\left(v_i\right) = \sqrt{\frac{m_i}{2\pi k_\mathrm{B} T}} \exp\left(-\frac{m_i v_i^2}{2 k_\mathrm{B} T}\right). \tag{2.19}$$

Equations 2.17 and 2.19 are relevant for temperature coupling and, respectively, the generation of initial velocities regarding MD simulations as carried out in Chapters 5-7. The energy distribution of the *canonical* or *Gibbs ensemble*, a thermodynamic system with a given set of the characteristic variables (N, V, T) at thermal equilibrium, is well described by the *Boltzmann distribution* of probabilities

$$\pi\left(q, p\right) = \frac{1}{Q_{NVT}} \exp\left(-\beta \mathcal{H}\left(q, p\right)\right) \tag{2.20}$$

of the classical Hamiltonian $\mathcal{H}\left(q, p\right)$.[54] The unnormalized probability $\exp\left(-\beta \mathcal{H}\left(q, p\right)\right)$ of a state (q, p) is also referred to as *Boltzmann factor*. In analogy to the microcanonical ensemble, the classical canonical partition function

$$Q_{NVT} = \frac{1}{\mathrm{h}^{3N} \xi} \int\limits_{\mathbb{R}^{6N}} \mathrm{d}q \ \mathrm{d}p \exp\left(-\beta \mathcal{H}\left(q, p\right)\right)$$

is defined as an integral (or sum) over (Boltzmann) weights of each microstate. Thus, it serves as a normalization factor for the probability $\pi\left(q, p\right)$ of a microstate $x = (q, p)$,

$$\int\limits_{\mathbb{R}^{6N}} \mathrm{d}q \ \mathrm{d}p \, \pi\left(q, p\right) = 1.$$

ξ is defined according to Equation 2.16 and the inverse temperature at thermal equilibrium is represented by β

$$\beta = \frac{1}{-kT}$$

comprising either the Boltzmann constant, $k = k_{\mathrm{B}} = 1.38066 \times 10^{-23}\mathrm{J/K}$, or the *gas constant*, $k = R = 8.31446\,\mathrm{J\,(mol/K)}$ depending on whether for energies the unit J or, respectively J/mol is preferred. By the way, both natural constants are related to each other by the fundamental *Avogadro constant* $N_{\mathrm{A}} = R/k_{\mathrm{B}}$. The thermodynamic potential related to the NVT ensemble is referred to as *Helmholtz* or *free energy*

$$A = -\frac{1}{\beta} \ln Q_{NVT} \tag{2.21}$$

which becomes minimal at thermodynamic equilibrium. In a similar manner, all macroscopic thermodynamic quantities of the canonical ensemble can be computed through the canonical partition function,[1] e. g., the internal energy

$$U = -kT^2 \left(\frac{\partial \ln Q_{NVT}}{\partial T} \right)_{NV},$$

the entropy $S = k \ln Q_{NVT} + U/T$, and intensive state variables as well such as the pressure

$$p = kT \left(\frac{\partial \ln Q_{NVT}}{\partial V} \right)_{NT}.$$

In analogy to Equation 2.21, one often finds in common literature the notation

$$A = -\frac{1}{\beta} \ln Z_{NVT}$$

where A is related to the *configurational partition function* Z_{NVT} which is, in contrast to the phase space partition function Q_{NVT}, associated only with coordinates q and corresponding potential energies $V(q)$ but neither with momenta nor kinetic energies[44]

$$Z_{NVT} = \int_{\mathbb{R}^{3N}} \mathrm{d}q \exp\left(-\beta V(q)\right).$$

Due to the exceeding phase space complexity of (biological) many-body systems, the partition function cannot be calculated analytically. However, even with numerical methods the phase space can only be sampled roughly in reasonable time. As a consequence, the theoretical canonical ensemble average

$$\langle O \rangle = \int_{\mathbb{R}^{6N}} \mathrm{d}q \; \mathrm{d}p \, O(q,p) \, \pi(q,p)$$

of an interesting observable O can only be roughly approximated from the time average

$$\langle O \rangle = \frac{1}{n} \sum_{i=1}^{n} O\left(q_i, p_i\right)$$

of n states sampled using an MD or MC simulation provided that the sampled microstates follow the Boltzmann distribution of energies. A couple of thermostat algorithms meeting that condition are available[53,55] and discussed in Chapter 3. The *Boltzmann ratio*

$$\frac{\pi\left(q_{\mathrm{A}}, p_{\mathrm{A}}\right)}{\pi\left(q_{\mathrm{B}}, p_{\mathrm{B}}\right)} = \frac{\exp\left(-\beta \mathcal{H}\left(q_{\mathrm{A}}, p_{\mathrm{A}}\right)\right)}{\exp\left(-\beta \mathcal{H}\left(q_{\mathrm{B}}, p_{\mathrm{B}}\right)\right)} = \exp\left[-\beta\left(\mathcal{H}\left(q_{\mathrm{A}}, p_{\mathrm{A}}\right) - \mathcal{H}\left(q_{\mathrm{B}}, p_{\mathrm{B}}\right)\right)\right]$$

of probabilities $\pi\left(q, p\right)$ associated with two states A and B is easily derived from respective Boltzmann factors without knowledge about integral Q_{NVT} in the denominator of Equation 2.20 which cancels out. One should bear in mind that apparently small energy differences can yield significant differences in the probabilities of occurrence. Given an energy difference of $\Delta_{\mathrm{AB}}\mathcal{H} = \mathcal{H}\left(q_{\mathrm{A}}, p_{\mathrm{A}}\right) - \mathcal{H}\left(q_{\mathrm{B}}, p_{\mathrm{B}}\right) = -20\,\mathrm{kJ/mol}$ at room temperature (298 K) yields a $\exp\left(-\beta\Delta_{\mathrm{AB}}\mathcal{H}\right) \approx 3200$ fold likeliness of state A over state B. In general, the Boltzmann ratio in favor of a subspace against another can be used for the calculation of the Helmholtz free energy difference

$$\Delta_{\mathrm{AB}}A \approx -\frac{1}{\beta} \ln\left[\frac{N_{\mathrm{A}}}{N_{\mathrm{B}}}\right] \tag{2.22}$$

between two subsets $A, B \in \mathbb{R}^{6N}$ of the system's phase space. If the states sampled during a molecular simulation process are distributed according to the Boltzmann distribution in Equation 2.20, $\Delta_{\mathrm{AB}}A$ can be approximated by the ratio of the numbers N_{A} and N_{B} of states of the respective subspaces as stated by Equation 2.22. As a consequence of the negligible *statistical weights* (probabilities of occurrence) associated with high energy states it is often sufficient to consider low energy states or areas of the phase space for the purpose of estimating free energy differences. The basic idea was already pointed out in Chapter 1 in the context of crystallographic structure files of proteins as representing their most preferential conformations.

Isothermal–isobaric ensemble

A thermodynamic system even more convenient for the description of natural processes than the canonical ensemble is characterized by constant pressure in addition to constant temperature. It describes nearly all biological systems since the earth's atmosphere itself provides these two constraints. The extensive natural variable V associated with

the (micro)canonical ensemble is replaced by its conjugate extensive variable p resulting in energy fluctuation with variance

$$\sigma^2 = V k_\mathrm{B} T \beta_\mathrm{T}. \tag{2.23}$$

where the fluctuation is, in analogy to the canonical ensemble, proportional to the isobaric heat capacity c_p [53]

$$\sigma^2 = k_\mathrm{B} T^2 c_\mathrm{p}. \tag{2.24}$$

In conjunction with a constant number of particles we have specified the *isothermal–isobaric* NpT ensemble also referred to as *Gibbs ensemble*. It is related to the thermodynamic potential $G\left(N, p, T\right)$ denoted as *Gibbs free energy* of a closed system capable of heat exchange as well as pressure-volume work as expressed by Equation 2.2.[1] As pointed out by Hünenberger, the fixed values (p and T) as well do fluctuate around their corresponding macroscopic valus, though, the corresponding magnitudes will vanish in the limit of a macroscopic system.[53] Since $G\left(N, p, T\right)$ differs from the Helmholtz energy only by an amount representing pressure-volume work

$$G\left(N, p, T\right) = A\left(N, V, T\right) + pV,$$

its probability distribution of states at thermodynamic equilibrium includes an additional energy term pV addressing the mechanical work

$$\pi\left(q, p\right) = \frac{1}{Q_{NpT}} \exp\left(-\beta\left(\mathcal{H}\left(q, p\right) + pV\right)\right)$$

As a consequence, its partition function

$$Q_{NpT} = \frac{1}{\mathrm{h}^{3N}\xi} \int_0^\infty \mathrm{d}V \int_{\mathbb{R}^{6N}} \mathrm{d}q \; \mathrm{d}p \exp\left(-\beta\mathcal{H}\left(q, p\right)\right)$$

comprises an additional integral representing the variable volume. The Gibbs energy can be derived from the partition function in analogy to Equation 2.21 for the canonical ensemble using[1]

$$G = -\frac{1}{\beta} \ln Q_{N,p,T}. \tag{2.25}$$

Of course, there are further statistical ensembles coping with other boundary conditions. Among these, the most popular one is the *macrocanonical* or (μ, V, T)-*ensemble* representing systems with variable numbers N_i of particles due to chemical reactions. Here, the extensive state variable N has been replaced by its conjugate intensive variable μ, the chemical potential quantifying the potential of chemical reactions. In general, the

set of natural variables for a thermodynamic potential must include at least one extensive variable in order to yield a complete thermodynamic description of the underlying physical system.[47] However, the canonical as well as the isobaric-isothermal ensemble are the most suitable theoretical frameworks for the investigation of free energies and binding affinities regarding host–guest systems. The reader might get the impression in the following that both Helmholtz and Gibbs energies are interchangeable since, depending on the reference literature used to describe free energy estimation methods, the one or the other notation is used here. Although they are indeed not in theory, it has been shown that, in practice, the difference in energies calculated with different statistical ensembles (canonical or isobaric-isothermal) vanish quickly with increasing system size.[56] Furthermore, the difference between the two quantities amounting to the volume work $p\Delta V$ is negligible in incompressible fluids such as water. As a consequence, the difference of enthalpy ΔH reduces to a change in internal energy[49] which equals the ensemble/time-average of the potential energy ΔU. Accordingly, the theoretical fundamentals of the methods developed for the purpose of free energy estimation as discussed in the following either make use of the Gibbs or Helmholtz formulation.

2.3 Thermodynamic path methods

Due to a strong reliance on thermodynamic principles and statistical mechanics, this class of methods constitutes the most rigorous approach to the estimation of binding free energies. Basically, they determine the thermodynamic work needed to transfer system A to system B.[57] These two systems might, for instance, represent two well-defined microscopic states or some differing macroscopic parameters such as the temperature associated with the *same* Hamiltonian (same molecules involved in both systems). Furthermore, A and B might include the same protein slightly differing due to a point mutation (one differing amino acid position) or different ligands resulting in two distinguishable Hamiltonians. In terms of a *thermodynamic cycle*,

$$\begin{array}{ccc} \mathrm{P} + \mathrm{L} & \xrightarrow{\Delta G_1} & \mathrm{PL} \\ \Delta G_3 \downarrow & & \downarrow \Delta G_4 \\ \mathrm{P} + \mathrm{L}' & \xrightarrow{\Delta G_2} & \mathrm{PL}' \end{array} \tag{2.26}$$

the former type of chemical change associated with unmodified Hamiltonians would correspond to vertical arrows representing typical association/dissociation reactions and yielding a binding free energy difference such as ΔG_1 or ΔG_2 following Equation 2.3.

In contrast, the latter type of reaction associated with an alteration in the system Hamiltonian (system topology) is represented by vertical arrows in "Equation" 2.26. Therefore, these changes are commonly referred to as "*alchemical transformations*" that are, in general, not related to a real chemical event. As a consequence, such a calculation scheme does not provide simple differences in free energies but *differences in differences* in free energies between the two host–guest systems which are related to *relative binding affinities* (RBA)[44]

$$\Delta\Delta G_{12} = \Delta G_2 - \Delta G_1.$$

Hence, according to the thermodynamic cycle above, ΔG_2 can be calculated on the basis of data sampled from the two systems forming the bottom line (direct route) or, alternatively, following the vertical reactions and given that ΔG_1 is known (or calculated directly) since

$$\Delta G_2 = \Delta G_1 + \Delta G_4 - \Delta G_3. \tag{2.27}$$

Equation 2.27 associated with a thermodynamic cycle is valid because the Gibbs energy is a state function which means that its value does not depend on the path in phase space the system has taken. Using a thermodynamic path method, the chemical transformation corresponding to any of the arrows in 2.26 is split into several intermediate states with overlapping distributions along a *reaction path*. These states are often specified by a set $\{\lambda_i\}$ of order (coupling) parameters along the path for which the thermodynamic work is calculated.[57] What this class of free energy estimation methods essentially has in common is the construction of a *Potential of Mean Force* (PMF) profile along that coordinate.[58] The *PMF profile* notation is synonymously used for a free energy profile due to the equality

$$\frac{\partial}{\partial \xi}\Delta G_{A\to B} = -\left\langle F_\xi \right\rangle_\xi \tag{2.28}$$

which relates the average (*mean*) force acting on some coordinate ξ to the partial derivative of the free energy with respect to ξ. Equation 2.28 also quantifies the constraint force required to fix the system's reaction coordinate at a particular value.

Free energy perturbation

A popular thermodynamic approach to the estimation of free energy differences is known as *free energy perturbation* published by Zwanzig in the middle of the last century.[59] Basically, it was designed to compute the physical work needed for changing a reference system characterized by the Hamiltonian $\mathcal{H}_0(q,p)$ to a target system $\mathcal{H}_1(q,p)$

$$\mathcal{H}_1(q,p) = \mathcal{H}_0(q,p) + \Delta\mathcal{H}(q,p)$$

where $\Delta\mathcal{H}(q,p)$ consists of all discriminating energy terms. Then, the Helmholtz free energy difference between the reference and target system (or state) is given as

$$\Delta A = -\frac{1}{\beta} \ln \frac{Q_1}{Q_0} \tag{2.29}$$

which is derived from the free energy definition according to Equation 2.21 and incorporating the canonical partition function Q. As an alternative to the partition functions used in Equation 2.29, one can use configurational integrals Z

$$\Delta A = -\frac{1}{\beta} \ln \frac{Z_1}{Z_0} \tag{2.30}$$

as defined on potential energies only. Both are considered equivalent if the particle masses of both systems are identical (and the kinetic term of the Hamiltonian that can be determined analytically cancels out) or if we are interested in the *excess* Helmholtz free energy. According to Chipot et al.,[56] this is usually the case and we will focus on the potential energy in the following. It was shown[59] that ΔA can be calculated using only a sampling of reference equilibrium configurations as depicted through the fundamental and formally exact FEP formula

$$\Delta A = -\frac{1}{\beta} \ln \left\langle \exp\left(\Delta U(q)\right)\right\rangle_0 \tag{2.31}$$

However, this direct strategy is only applicable if the probability distribution of microstates

$$P(q,p) = \frac{\exp\left(\Delta\mathcal{H}(q,p)\right)}{\int\int \exp\left(\Delta\mathcal{H}(q,p)\right) \mathrm{d}q\, \mathrm{d}p}$$

of both systems overlap sufficiently or, which qualitatively amounts to the same, if the Gaussian-like probability distribution function $P(\Delta\mathcal{H})$ or $P(\Delta U)$ has low variance.[56] Since, in practice, this will most likely not be the case, one needs to apply a *stratification* (staging) strategy by defining, in addition to the two end states, $N-2$ proper intermediate states with sufficiently narrow distributions $P(\Delta U_{i,i+1})$ for any pair of two consecutive states i and $i+1$. These are not necessarily physically meaningful. Using a coupling parameter λ with $\lambda_i \in [0,1]$ and $i \in \{1,\ldots,N-1\}$ where $\lambda_1 = 0$ and $\lambda_{N-1} = 1$ represent the reference and target state, respectively, corresponding Hamiltonians take the form

$$\mathcal{H}(\lambda_i) = \lambda_i \mathcal{H}_1 + (1-\lambda_i)\mathcal{H}_0 = \mathcal{H}_0 + \lambda_i \Delta\mathcal{H}.$$

The final free energy difference is then given as a sum

$$\Delta A = \sum_{i=1}^{N-1} \Delta A_{i,i+1} = -\frac{1}{\beta} \sum_{i=1}^{N-1} \ln \left\langle \exp\left(-\beta \Delta\lambda_i \Delta\mathcal{H}\right)\right\rangle_{\lambda_i} \tag{2.32}$$

of free energy differences between any two consecutive states where $\Delta\lambda_i = \lambda_{i+1} - \lambda_i$. Clear statements about the choice of N and λ_i are not possible. It is assumed that, in general, a high number of intermediate states increases the accuracy at the expense of efficiency.

For the explicit purpose of binding free energy calculations, one might define an order (coupling) parameter on the basis of the distance between the ligand and the binding center and determine the change in free energy along the upper horizontal arrow of the thermodynamic cycle addressed by Equation 2.26. In practice, though, this often turns out to be complicated due to large conformational changes required between the bound and unbound state. Thus, for the direct route, other thermodynamic methods described below seem more convenient.[56] Nevertheless, as represented by the horizontal and vertical arrows, FEP techniques have been successfully applied to the determination of absolute and, respectively, relative host–guest binding free energies using classical MD simulations.[60–62] The vertical routes yielding relative binding affinites are often called *alchemical transformations*. In spite of their high computational costs due to usually tens of MD trajectories along the coupling parameter, FEP binding free energies have meanwhile been determined on the basis of QM/MM time series associated with a fivefold computational demand.[63] QM/MM methods are characterized by the combination of a classical molecular mechanics force field ("MM" part) with a quantum chemical representation of the active site and the guest molecule ("QM" part) allowing for the modelling of chemical reactions (bond breaking/forming).

Thermodynamic integration

Another popular thermodynamic work-based approach to the rigorous estimation of binding free energies referred to as *thermodynamic integration* (TI) was developed already in the 1930's by Kirkwood.[64] Just as the FEP method, it rests on a phase space decomposition (stratification) through an order parameter λ which, in contrast to FEP, couples partial derivatives of the Hamiltonian with respect to λ_i. Thus, both FEP and TI provide a free energy profile along a (generally unphysical) reaction coordinate. However, by considering $\Delta A\,(\lambda)$ as a function of λ, TI evaluates it as the area

$$\Delta A = \int_0^1 \left\langle \frac{\partial U\,(q,\lambda)}{\partial \lambda} \right\rangle_\lambda \mathrm{d}\lambda \tag{2.33}$$

under the curve of averaged partial derivatives of the potential energy with respect to the reaction coordinate λ, $\partial U/\partial\lambda$. We remember its relationship

$$\langle F_\lambda \rangle_\lambda = -\left\langle \frac{\partial U\left(q,\lambda\right)}{\partial\lambda} \right\rangle_\lambda$$

to the mechanical force acting on the reaction coordinate λ and averaged over all other (generalized) coordinates q_i at a fixed (addressed by the subscript notation of the closing angular bracket) λ value. For that reason, this procedure has a particularly obvious relation to PMF. One should be aware of the fact that, in general, rather than dealing with a real particle coordinate, λ may represent an arbitrary function of atomic coordinates.[65] Concretely, the integral in Equation 2.33 would be approximated by numerical methods such as the commonly used simple trapezoidal rule

$$\Delta A = \int_0^1 \left\langle \frac{\partial U}{\partial\lambda} \right\rangle_\lambda \mathrm{d}\lambda \approx \frac{1}{2}\sum_i^{N-1} \left(\lambda_{i+1}-\lambda_i\right)\left[\left\langle \frac{\partial U}{\partial\lambda_i} \right\rangle_{\lambda_i} + \left\langle \frac{\partial U}{\partial\lambda_{i+1}} \right\rangle_{\lambda_{i+1}}\right]$$

or other methods probably comprising more than two out of N (intermediate) states at a time.[66]

Bennett acceptance ratio

FEP and TI quickly yield large statistical errors in the free energy difference if the perturbations ΔU is not close enough to zero. Hence, on his way to a minimal statistical error, Bennett presented an alternative strategy denoted as *Bennett Acceptance Ratio* (BAR) which, in contrast to the former methods, involves samplings from *both* states equally.[67] He showed that the free energy difference between two (intermediate) states i and j can be determined through

$$\Delta A = \beta^{-1}\left[\ln\frac{\langle f\left(\mathcal{H}_i-\mathcal{H}_j+C\right)\rangle_\mathrm{j}}{\langle f\left(\mathcal{H}_j-\mathcal{H}_i+C\right)\rangle_\mathrm{i}}\right]+C$$

on the basis of the two Hamiltonians $\mathcal{H}$ (or, alternatively, potential energies U) and the Fermi function

$$f\left(x\right)=\left[1+\exp\left(-\beta x\right)\right]$$

C is determined iteratively such that

$$\langle f\left(\mathcal{H}_i-\mathcal{H}_j+C\right)\rangle_\mathrm{j} = \langle f\left(\mathcal{H}_j-\mathcal{H}_i+C\right)\rangle_\mathrm{i}$$

and the free energy difference between two overlapping states, i. e. $j=i+1$, is obtained as[66]

$$\Delta A_{i+1,i} = \beta^{-1}\ln\frac{n_{i+1}}{n_i}+C.$$

In particular if the number of frames n_i and n_{i+1} of the two corresponding states are equal, it becomes obvious that C approaches the free energy difference between these two states. The overall energy difference ΔA between the two terminal states associated with the ends of the reaction path can then be expressed in terms of overlapping intermediate states

$$\Delta A = \sum_{i=1}^{N-1} \Delta A_{i+1,i}$$

again. In the light of an obvious similarity with Equation 2.32, the BAR method can be considered as a modification of the original FEP method. Comparative studies revealed a higher robustness of the BAR method since it is both less demanding regarding the number of intermediate states and less dependent on their distribution than TI. However, the latter algorithm is much more straightforward to implement and effectively faster.[66]

Weighted histogram analysis method

So far, we have only elaborated free energy methods based on stratification techniques that consider pairs $(i, i+1)$ of states upon calculation of a free energy difference. A popular multiple histogram-based strategy analysing all states at one go that is commonly referred to as *Weighted Histogram Analysis Method* (WHAM) was proposed by Ferrenberg et al. as early as 1989.[68] Kumar et al. transferred it to alchemical constant temperature simulations only few years later. That is the original equations were extended to molecular mechanics force potentials characterizing biomolecules and useful for free energy profiles along coupling parameters and/or as a function of the temperature.[58] The reaction path will most likely include conformational transition regions with unfavourable energies that are hardly sampled during an MD or Monte-Carlo simulation resulting in a slow convergence of the probability distribution $P(q)$ of coordinates q. To remedy this problem one may start several simulations at different positions i along the reaction path $\xi = f(q)$ and bias the underlying potential U

$$U'(\xi) = U(\xi) + w_i(\xi)$$

in order to sufficiently sample unfavourable regions as well. The weighting function $w(\xi)$ often taking a quadratic ("umbrella") form

$$w_i(\xi) = \frac{k_w}{2}(\xi - \xi_i)^2$$

penalizes deviations from the desired conformation $\xi_i = \xi(q_i)$ of any intermediate state i along the chosen reaction path.[15] Such biasing techniques are denoted as *Umbrella Sampling* (US) and generate non-Boltzmann (non-equilibrium) distributions.[69]

Afterwards, a reweighting strategy is required so as to derive Boltzmann averages from that data. This is where, among alternative methods, WHAM comes into play. It combines sets of simulations of N_w discrete states with different biasing potentials $w_i(\xi)$ and constructs histograms on the basis of bins[15] along ξ yielding *biased* individual distributions $\langle \rho(\xi) \rangle_i^{\text{biased}}$ from which *unbiased* individual distributions

$$\langle \rho(\xi) \rangle_i = \mathrm{e}^{\beta w_i(\xi)} \ \langle \rho(\xi) \rangle_i \ \mathrm{e}^{-\beta F_i}$$

are derived. The unbiased combined probability distribution is given by

$$\langle \rho(\xi) \rangle = \frac{\sum_{i=1}^{N_w} n_i \langle \rho(\xi) \rangle_i}{\sum_{j=1}^{N_w} n_j \mathrm{e}^{-\beta[w_j(\xi) - F_j]}} \tag{2.34}$$

as a ξ-dependent weighted sum of N_w individual distributions. n_i denotes the number of independent data points used to construct the biased distribution function of state i. Respective free energies F_i are determined using the optimal estimate for the distribution function

$$\mathrm{e}^{-\beta F_i} = \int \mathrm{d}\xi \ \mathrm{e}^{-\beta w_i(\xi)} \ \langle \rho(\xi) \rangle \,. \tag{2.35}$$

Since the quantities of interest, $\rho(\xi)$ and F_i, depend on each other, Equation 2.34 and 2.35 must be solved in a self-consistent manner that is, in practice, iteratively.[70]

As we have seen, the free energy calculation methods discussed in this section require stratification strategies likely in combination with enhanced sampling techniques due to the quasi-nonergodicity of macromolecular biological systems. Often, multiple samplings/trajectories of usually tens of initial states are carried out. As a consequence, the implementation becomes difficult and the computations highly complex with regard to time. Furthermore, the applicability of thermodynamic paths-based methods is limited to sufficiently similar substances.[25,26]

2.4 Thermodynamic end point methods

In contrast to the previous class of approaches to the estimation of free energy differences, the methods discussed in the following consider (at most) two end states only, mostly one bound and one unbound state of an association/dissociation process. In contrast to many of the methods described in the previous section performing alchemical transformations, these calculations are associated with horizontal arcs according

to Figure 2.26. Since no intermediate states along some thermodynamic work path are considered resulting in far less computation time, these methods are much more appropriate for virtual screening and structure-based drug design. Nevertheless, these techniques as well cope with thermodynamic principles as they are based on molecular mechanical force fields describing physical interactions between atoms.

Molecular mechanics Poisson-Boltzmann surface area

A method referred to as *Molecular Mechanics Poisson-Boltzmann Surface Area* (MM/PBSA) combines force field-based molecular mechanics ("MM") simulations with a solvation term according to the Poisson-Boltzmann ("PB") equation which describes electrostatic interactions of particles and molecules in solution. Using the PB equation for the calculation of solvation free energies implies that the solvent is represented implicitly as a continuum rather than explicitly through specific atomic coordinates. However, the original MM/PBSA protocol comprises explicit water molecular dynamics (MD) simulations of each system/component $X \in \{\mathrm{PL}, \mathrm{P}, \mathrm{L}\}$ (representing the protein–ligand complex, protein, and ligand, respectively) from which the water molecules are removed prior to the computation of free energies.[71] The corresponding functional form of the free energy

$$\langle G_X \rangle_Y = \langle E_{\mathrm{MM}} \rangle_Y + \langle E_{\mathrm{PBSA}} \rangle_Y + T S_{\mathrm{MM}} \tag{2.36}$$

associated with a particular system X is composed of a couple of energy contributions derived from MD trajectory (time series) Y: the molecular mechanical energy E_{MM} equating the sum of all bonded and non-bonded force field terms (discussed in more detail in Chapter 3), is determined as a time-average (indicated by $\langle \cdot \rangle$) of component X. The third term on the right of Equation 2.36 is related to the solute's conformational entropy and typically determined by applying[45] the *quasi-harmonic approximation*[72] or a *normal mode analysis*.[73] Finally, the solvation energy E_{PBSA} (averaged as well) is obtained

$$E_{\mathrm{PBSA}} = E_{\mathrm{polar}} + E_{\mathrm{nonpolar}}$$

as the sum of a polar and nonpolar contribution which are responsible for hydrophilic and, respectively, hydrophobic interactions of the solute (complex or single reactants) with each other or with the solvent possibly including ions. E_{polar} is usually calculated numerically for a set of Cartesian grid points through some finite-difference solver of the Poisson-Boltzmann equation resulting in an electrostatic potential at grid points i which are simply summed up

$$E_{\mathrm{polar}} = \frac{1}{2} \sum_i z_i \left(\phi_i^{\mathrm{so}} - \phi_i^{\mathrm{I}} \right).$$

z_i and ϕ_i represent the charge and, respectively, the calculated electrostatic potential at grid point i for the transfer from vacuum associated with permittivity $\varepsilon_0 = 1$ to water with the dielectricity constant (relative permittivity) $\varepsilon = 80$.[74] In contrast, the nonpolar contribution

$$E_{\text{nonpolar}} = \gamma\,\text{SA} + b$$

is simply derived linearly from the *solvent accessible surface area* (SASA; we will use SA instead) using a surface tension amounting to $\gamma = 20.9 \pm 2.9\,\text{J}/\left[\text{mol}\,\text{Å}^2\right]$ and an offset at $b = 3.6 \pm 0.4\,\text{J/mol}$ fitted to small alkanes using least-squares.[74] Over time, various other values for these two constants have been proposed.[45] According to

$$\Delta G = \begin{cases} \langle G_{\text{PL}}\rangle_{\text{PL}} - \langle G_{\text{P}}\rangle_{\text{P}} - \langle G_{\text{L}}\rangle_{\text{L}} & \text{using 3 simulations } \{\text{PL}, \text{P}, \text{L}\}, \text{ or} \\ \langle G_{\text{PL}} - G_{\text{P}}\rangle_{\text{PL}} - \langle G_{\text{L}}\rangle_{\text{L}} & \text{using 2 simulations } \{\text{PL}, \text{L}\}, \text{ or} \\ \langle G_{\text{PL}} - G_{\text{P}} - G_{\text{L}}\rangle_{\text{PL}} & \text{using 1 simulation } \{\text{PL}\}. \end{cases} \tag{2.37}$$

the final binding free energy difference is calculated as the difference of energies between the bound system (PL) and the two single reactants' samplings (P and L). The first line of Equation 2.37 characterized by $X = Y$ corresponds to the original approach proposed by Kollman where each system (PL, P, L) was simulated and averaged independently. However, one might as well extract all averages from a proper subset of the complex (PL) sampling only as represented by the last line of Equation 2.37. Indeed, it has been shown that this simplification often yields more accurate results (associated with a substantially lower standard error) from a single than from three distinct simulations.[45] Considering, in addition, the ligands' reorganization energy by performing a second MD sampling consisting of ligand snapshots only can further improve the result.[75]

Molecular mechanics generalized Born surface area

The *Molecular Mechanics Generalized Born Surface Area* (MM/GBSA) based on an approximation to the exact Poisson-Boltzmann equation is another popular continuum-solvation method for the estimation of binding affinities is called. As a consequence, free energies

$$\langle G_X\rangle_Y = \langle E_{\text{MM}}\rangle_Y + \langle E_{\text{GBSA}}\rangle_Y + TS_{\text{MM}} \tag{2.38}$$

are calculated in analogy to Equation 2.36 used for the MM/PBSA model. However, another strategy for solvation energies

$$E_{\text{GBSA}} = E_{\text{cav}} + E_{\text{vdW}} + E_{\text{pol}}$$

was proposed.[76] Cavity formation and hydrophobic solute–solvent interactions are quantified by

$$E_{\text{cav}} + E_{\text{vdW}} = \sum_k \sigma_k \text{SA}_k$$

using the total SASA of atoms of type k and an empirical atomic solvation factor σ_k that was preliminarily set to $\gamma = 30.1\,\text{J}/\left[\text{mol}\,\text{Å}^2\right]$. The solute-solvent electrostatic polarization term E_{polar} is obtained through the generalized Born equation[76]

$$E_{\text{pol}} = \frac{1}{8\pi}\left(\frac{1}{\epsilon_o} - \frac{1}{\epsilon}\right)\sum_{i,j}^{N} \frac{z_i z_j}{\sqrt{r_{ij}^2 + a_i a_j e^{-D}}}$$

where r_{ij} is the distance between particles i and j, a_i denotes the effective Born radius defining a particles burial inside the solute, and[76]

$$D = \left(\frac{r_{ij}}{2\sqrt{a_i a_j}}\right)^2.$$

Finally, binding free energies are determined in analogy to the MM/PBSA model according to Equation 2.37.

Both elaborated continuum-solvation models MM/PBSA and MM/GBSA rigorously decompose the binding free energy into contributions originating from different interactions[77] and were shown to successfully reproduce and rationalize experimental observations[45]. Moreover, they do not require any training set to fit parameter coefficients for different energy contributions. However, they are sensitive to the choice of the solute dielectric constant which is, therefore, recommended to reflect characteristics of the protein–ligand binding interface. According to Hou et al., though, and depending on the target system, MM/PBSA tends to perform better than MM/GBSA in estimating absolute but not necessarily regarding relative binding free energies. As a consequence and due to its efficiency, they suggest the MM/GBSA model for the ranking of inhibitors in drug design.[77] Genheden et al. have pointed out that, due to its implicit representation, certain effects of water such a bridging hydrogen bonds at the binding site are neglected.[45] From this point of view, these two methods seem particularly inappropriate for systems comprising critical water molecules as we will see in Chapter 6.

Linear interaction energy method

The *Linear Interaction Energy* (LIE) method as the last of this class of algorithms was introduced by Åqvist et al. in the 1990s and differs in several aspects from the two

previous ones.[78] It originates from the *linear-response approximation* which relates free energy differences of binding (or solvation)

$$\Delta G \simeq \frac{1}{2}\left[\left\langle E^{\text{elec}}\right\rangle_{\text{PL}} - \left\langle E^{\text{elec}}\right\rangle_{\text{L}}\right] \tag{2.39}$$

to electrostatic molecular mechanical interactions between the ligand and its environment. Among other derivations, the approximation depicted by Equation 2.39 can be obtained by expanding the exponent of Zwanzig's FEP formula (Equation 2.31) and truncating after the linear term (or after the second, assuming equal mean square fluctuations of both potentials that would cancel out). Equation 2.39 states that the electrostatic contribution to the free energy difference equals half of the corresponding solute–solvent interaction energy.[78] In practice, interaction energies E^{elec} are averaged over explicit water simulations of both the free ligand only, $\langle\cdot\rangle_{\text{L}}$, and the ligand in complex with the target molecule, $\langle\cdot\rangle_{\text{PL}}$. The value of 0.5 for the factor in Equation 2.39 originating from the assumption of harmonic (parabolic) free energy curves was confirmed by several computational studies including FEP that had been applied to the estimation of free energies of ionic solvation.[78] Obviously, hydrophobic effects due to non-polar van-der-Waals (VDW) forces are not considered. It was, however, shown that the solvation free energy of alkanes depends approximately linearly on their length.[79] For that reason, the LRA equation was simply extended

$$\Delta G = \frac{1}{2}\left[\left\langle E^{\text{elec}}\right\rangle_{\text{PL}} - \left\langle E^{\text{elec}}\right\rangle_{\text{L}}\right] + \alpha\left[\left\langle E^{\text{vdw}}\right\rangle_{\text{PL}} - \left\langle E^{\text{vdw}}\right\rangle_{\text{L}}\right] \tag{2.40}$$

by a second term addressing, in analogy to the electrostatic contribution, hydrophobic (VDW) interactions of the ligand molecule with its environment. Due to difficulties with the theoretical derivation of α, the developers of the LIE method decided to calibrate its value empirically using an aspartic proteinase called endothiapepsin (EP) as a test system. Several crystal structures of native EP as well as in complex with five different inhibitors and experimental binding data had been available at that time. Having fitted free energies using the entire set or various combinations of four inhibitors yielded $0.158 \leq \alpha \leq 0.165$. Later, it proved necessary to recalibrate α for each target system under investigation and an additional empirical constant γ was introduced to the LIE model

$$\Delta G = \alpha\left\langle \Delta E^{\text{vdw}}\right\rangle + \beta\left\langle \Delta E^{\text{elec}}\right\rangle + \gamma \tag{2.41}$$

as well, though, it was often set to zero.[49] Aside from that, the coefficient for electrostatic interactions denoted as β here was, in degoration from its original value, set to particular values $\beta \leq 0.5$ depending on the ligand's charge and the number of hydroxyls.[80] Moreover, in recent applications, β was treated as an empirical parameter to be calibrated just like α.[81,82] This progress illustrates the difficulty in developing a

generalized LIE model with coefficients transferable to any target system. Obviously, no consensus values for α, β, and γ can be determined.[49] instead, it seems necessary to recalibrate them independently for every target system using a trainig set of ligands with known binding affinities.

Further extension of the LIE model through an entropic or non-physical (structural) descriptors has been investigated as well,[49,83] surely having inspired our research regarding the development of predictive models in subsequent chapters of this thesis. Due to large fluctuations during an MD run, the calculation of conformational entropies requires a large number of snapshots. On the other hand, longer MD simulations often achieve worse results and the simulation length should be chosen carefully.[77] All in all, results gained from (extended) LIE applications turned out to correlate very well with experimental affinities.[83–85] Typical average errors of the LIE model between predicted and experimental free energies amount to approximately 4.0 kJ/mol which is close to experimental errors of affinity measurements (about 2.0 kJ/mol).[49] In contrast to the two elaborated implicit solvation models (MM/PBSA, MM/GBSA), the LIE method requires a training set with known affinities for the estimation of α, β, and perhaps γ. However, taking into account the effect of water explicitly must be considered as an advantage (regarding hydrogen bonds, etc.). Furthermore the LIE method is considerably easier to implement since the former two require in addition a numerical Poisson-Boltzmann solver. Finally, some investigations reveal a better reliability of the LIE method,[77] while others give the same statement in evidence about continuum-solvation models.[71]

2.5 Molecular docking and scoring functions

Molecular docking algorithms are structure-based tools particularly designed for a quick identification of physically reasonable binding poses (hence the term *docking*) followed by a subsequent estimation of binding affinities on the basis of so-called *scoring functions*. Their rise and popularity is inevitably accompanied by the increasing number of protein structure determinations over the last decades which led to the development of three-dimensional structure data bases such as the PDB. Using a given target structure, it became possible to quickly screen a large set of chemical compound. Aside from protein–ligand complexes they have as well been applied to the investigation of protein–protein or protein-nucleic acid interactions.[86–88] Due to an outstanding performance compared to MD-based techniques described in previous sections, these methods are particularly suitable for virtual screening of large compound libraries and hit identifi-

cation (and sometimes lead optimization) during drug discovery processes within reasonable time.[86,89] In particular, they have been attested to quickly and successfully generate a series of useful binding poses to work with. However, neither in prioritizing the natural pose nor in accurately calculating binding affinities, several critical studies on docking methods revealed substantial reliability.[49,90,91] Therefore, and due to the vast number of available docking tools, we will only give a short overview over basic ideas rather than discussing the underlying algorithms in detail. The list of docking techniques sketched in the following is, thus, far away from being complete.

To a large extent, the accuracy and computational expense depends on the degree of representation of the involved molecules' flexibility. In early days of molecular docking, that is in the beginning of the 1980s, all binding partners were considered rigid resulting in the term *rigid-body docking*. Later, as denoted by *semi-flexible docking*, it became more and more popular to take into account the flexibility of small molecules (usually ligands). This strategy constitutes the most common nowadays and is particularly suitable for lead optimization whereas rigid docking is usually carried out during early stage of hit identification.[86] In more advanced and, consequently, much costlier applications, target conformations are considered flexible as well (*flexible docking*). Protein flexibility is, however, often restricted to amino acids defining the active site or to side-chain rotations on the basis of rotamer libraries with experimentally determined preferential side-chain conformations. Alternatively, it is possible to map ensembles of protein conformations onto a spatial grid which may as in case of the DOCK[92,93] algorithm represent potential energies and subsequently score ligand conformations against sets of grid values.[87] Since, in general, all docking approaches perform a combination of two major steps, we will discuss the most common of these strategies in the following.

Binding pose identification

In a first step, a set of binding poses/conformations are sampled at the target's active site for which various strategies and combinations of them have been developed. Due to several internal (conformational) as well as external (in relation to the target) degrees of freedom even of simple ligand molecules, this stage is considered very challenging. Kitchen et al. roughly divided those strategies into the categories systematic, random/stochastic, and molecular mechanical.[87]

Shape matching As its name implies, this class of methods mainly considers shape complementarity of two molecules, thereby, avoiding geometrical overlap between their molecular surfaces and matching complementary pharmacophore properties. Therefore, it is particularly suited for rigid-docking and limited in accuracy. A popular example for this type of systematic methods is the rigid-docking tool ZDOCK.[88] However, flexibility can be emulated by using a set of predetermined conformations that are docked in a rigid fashion. Basically, these methods are useful for binding site identification as well. The flexible-docking tool DOCK for instance employs this technique in a preliminary step in order to detect regions for the placement of ligand atoms.[86]

Incremental construction Ligands are decomposed to molecular fragments and reconstructed at the binding site. Usually, the algorithm starts with placing rigid parts serving as anchors that are incrementally connected by more flexible fragments possibly containing rotatable bonds. By this means, molecular flexibility is considered and the ligand is constructed in a *de-novo* manner at the active site. Popular tools of this class with slightly differing incremental strategies are DOCK and FLEXX.[86] DOCK, for instance, poses anchors according to steric complementarity and lets flexible side chains grow bond by bond where each bond's orientation space is explored systematically. In contrast, the tactic implemented in FLEXX[94] is based on interaction geometries between fragments and receptor groups including hydrogen bond donator/acceptor or hydrophobic interactions. Another tool termed Hammerhead[95] scores fragment poses before connecting those associated with high scores where each linkage is followed by an energy minimization procedure.[87]

Molecular mechanics This class of algorithms is grounded on molecular mechanics and dynamics dealing with the motion of particles due to atomic interactions. Their theoretical foundation are extensively discussed in Chapter 3. As a consequence, all such methods generally fall into the category of flexible docking. At best, they perform an local energy minimization (EM) of an initially constructed pose as described above in case of Hammerhead. DOCK as well carries out an EM routine after each construction step and before evaluating the final pose by means of a scoring function. Substantially more conformational motility is achieved by generating new poses using MD simulations. Since, however, molecular flexibility during an MD run is hindered by high energy barriers of the potential energy surface, one may elevate the system temperature in order to flatten those barriers and reach a wider range of conformations and poses.[87] Based on the same idea, the globally energy minimization method *simulated annealing* (SA) systematically elevates and reduces the temperature in multiple cycles

during an MD run. Such an exhaustive exploration of the conformational space is certainly accompanied by an exceeding demand on computer resources. Nevertheless, the popular AutoDock[96] tool, for instance, routinely utilizes SA in combination with other random or stochastic algorithms that are discussed below.[86]

Random search New binding modes can be generated by randomly changing the ligand geometry and/or its relative orientation with respect to the active site. Two major strategies are known for that purpose: *genetic algorithms* and Monte-Carlo methods.[87] Methods of the former class are inspired by natural evolution where new phenotypes/species are created through genotypic modifications, that is mutations or crossover of existing genes, and selected according to their fitness. Translated to molecular docking, new poses are generated by re-combining structural properties of existing geometries such as torsion angles or functional groups and evaluated through some scoring function. For instance, DOCK and GOLD[97] belong to the class of GAs which are principally designed to find the global energy minimum. A variant of GA denoted as Lamarckian GA is implemented in AutoDock as well as its derivative AutoDock Vina.[98] It performs an energy minimization after each genotypic change and maps back the phenotypic result to the genotypic level.[86] In contrast, Monte-Carlo (MC) methods discussed in Chapter 3 follow a slightly different idea. They explore the conformational space by directly and randomly changing a ligand's external (atomic) or internal (torsional) degrees of freedom. The new state's fitness in relation to the previous one is assessed in accordance with the Boltzmann factor. AutoDock combines SA with an MC approach to the generation of random poses during each SA cycle.[86]

Scoring functions

Having constructed binding modes by means of one or a combination of the described strategies, an assessment is required in order to predict corresponding binding free energies, rank poses according to their probability of occurrence (binding free energies), and, if desired, optimize them. As already stated, molecular docking tools usually employ scoring functions for that purpose. The actual task of these functions is to pick those candidates, that are most likely to occur in reality. Their quality can be evaluated by applying them to protein–ligand complexes of which crystallographic structures are available as reference – a procedure that is referred to as *redocking*. To be more precise, after a structural alignment (maximum atomic superimposition) of the crystallographic target and the one used for redocking, the two corresponding ligand poses A and B defined by N atoms are usually compared by means of a root mean square deviation

(RMSD)

$$\text{rmsd}\,(q) = \sqrt{\frac{1}{3N}\sum_{i=1}^{3N}\left(q_{\text{A},i} - q_{\text{B},i}\right)^2}$$

of their $3N$ atomic coordinates q_A and q_B from each other. The best results of high-quality algorithms are expected to achieve RMSD values below 1.5Å.[86] Many scoring functions have been developed during the past decades. Some of them come standalone whereas most are implemented in docking programs. Basically, three categories of scoring functions are known most of which have in common an additive decomposition of bindinding free energies

$$\Delta G = \sum_i w_i\,\Delta G_i$$

to some number of physical and/or structural contributions representing, for instance, changes in solvation, interaction energies, and conformations weighted through coefficients w_i. In case of the two continuum-solvation methods (MM/PBSA and MM/GBSA) described above, these weights are set to zero. In this respect, they resemble thermodynamic end state methods discussed previously.

Force field methods This class of scoring functions incorporate functional terms describing atomic interactions according to molecular mechanics force fields as used for MD simulations. A prominent example for this category is the widely used tool AutoDock. Its scoring function

$$\begin{aligned}\Delta G = {} & \Delta G_\text{vdw}\sum_{i<j}\left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}\right)\\ & + \Delta G_\text{hbond}\sum_{i<j}E\,(t)\left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}}\right)\\ & + \Delta G_\text{elec}\sum_{i<j}\frac{z_i z_j}{\varepsilon\,(r_{ij})\;r_{ij}}\\ & + \Delta G_\text{tor}\,N_\text{tor} \;+\; \Delta G_\text{sol}\sum_{i<j}\left(S_i V_j + S_j V_i\right)\mathrm{e}^{-r_{ij}^2/2\sigma^2}\end{aligned} \tag{2.42}$$

combines three molecular mechanical terms related to intermolecular interactions between atoms i and j with distance r_{ij} due to van der Waals forces ("vdw"), electrostatic ("elec"), and hydrogen bonds ("hbonds") with two terms quantifying conformational (torsional) diversity ("tor") by means of the number N_tor of rotatable bonds and, respectively, desolvation ("sol"). Each free energy term is associated with a coefficient ΔG_x determined empirically through linear regression analysis of a set of protein–ligand complexes with known binding constants.[96] We will not discuss Equation 2.42 more thor-

oughly since the concept of force field is detailed in Section 3.2. Further examples for this class of scoring functions are, D-Score, G-Score, GoldScore, and DOCK.[87]

Empirical scoring functions Empirical scoring function reveal a functional form similar to force field scoring functions. They, as well, calculate free energy differences additively from weighted contributions. In contrast to the first category, however, these contributions include *counts* of geometrical and/or structural properties rather than force field potentials. In a pioneering work,[99] Böhm developed the first empirical scoring function for the approximation of binding energies. This function was implemented in the docking tools LUDI and FlexX[94]

$$\begin{aligned}\Delta G = \Delta G_0 &+ \Delta G_{\text{h-bonds}} \sum_{\text{h-bonds}} f\left(\Delta R, \Delta\alpha\right) \\ &+ \Delta G_{\text{ionic-int}} \sum_{\text{ionic-int}} f\left(\Delta R, \Delta\alpha\right) \\ &+ \Delta G_{\text{aro-int}} \sum_{\text{aro-int}} f\left(\Delta R, \Delta\alpha\right) \\ &+ \Delta G_{\text{aro}} \left|A_{\text{lipo}}\right| + \Delta G_{\text{tor}} N_{\text{tor}}\end{aligned}$$

using an additional term which accounts for aromatic interactions. Here, the coefficients ΔG_x are used for weighting counts of protein–ligand interactions (caused by hydrogen bonds, electrostatic forces, and aromatic groups) as well as the hydrophobic surface area A_{lipo} and the number of flexible bonds N_{tor}. Deviations from ideal chemical geometries (distances and angles) are quantified by a penalty function f. Other popular empirical scoring functions are F-Score, ChemScore, and the standalone tool X-Score[100]

Knowledge-based potentials Knowledge-based scoring function constitute a rather different approach to the approximation of binding constants. They rely on statistical distributions of intermolecular atomic distances determined on the basis of large sets of three-dimensional protein–ligand complex structures mainly retrieved from PDB. The idea behind this approach is that particular atomic interactions with a higher occurrence than expected through a random distribution are assumed to be energetically favorable and contribute stronger to the binding energy. The statistical data is used to derive potentials of mean force (PMF) using *radial distribution functions*. According to the PMF scoring function,[101] the score

$$\text{PMF\_score} = \sum_{ij} A_{ij} = \sum_{ij} \left[-k_{\text{B}} T \ln \left(f^{j}_{\text{Vol-corr}}\left(r\right) \frac{\rho^{ij}_{\text{seg}}\left(r\right)}{\rho^{ij}_{\text{bulk}}} \right) \right]$$

is calculated as the sum of interaction free energies between any pair of a protein atom type i and ligand atom type j with distance r. T and k_{B} denote the absolute temperature and Boltzmann constant, respectively. $f^{j}_{\text{Vol-corr}}(r)$ is the ligand volume correction factor, $\rho^{ij}_{\text{seg}}(r)$ represents the number density of pairs of type ij in a structural database in a particular range of radius indicated by "seg", and ρ^{ij}_{bulk} is the reference distribution of i and j if they are not interacting. The quotient of the latter two designates the pair correlation or radial distribution function of the corresponding two atoms in the structural database.[101] Further scoring functions falling into the category of knowledge-based potentials are the standalone tool DrogScore and SMoG.[87]

According to a great many studies, predestination of a scoring category that outweighs the others in most cases seems very unlikely. However, it has been shown that the incorporation of conformational (rotational) entropy and solvation terms (even if quick and inaccurate) significantly elevates success rate. Nevertheless, it appears impossible "to develop scoring functions that perform equally well across many different protein families, regardless of their complexity and sophistication".[87] In recent years, attempts have been made to combine results from two or more scoring functions in order to balance individual errors. This strategy is known as *consensus scoring* and often further improves prediction accuracy.[87] Due to the involvement of molecular mechanics terms, force field-based scoring functions obviously resemble thermodynamic ensemble methods described in previous sections. Since, furthermore, many critical evaluations show that thermodynamic average-based methods mostly yield significantly better results than docking tools,[49,83–85,90,91] it seems, for the purpose of accurate binding free energy predictions possibly upon lead optimization or toxicity estimations, more reasonable to consider statistical ensembles of conformations/poses rather than relying on the score of a single binding mode only. Many aspects including solvation effects, induced-fit changes, and explicit water molecules are still not (sufficiently) treated by docking programs although developers strive to.[63,102] Thus, for accurate toxicological risk assessments or very late drug discovery stages, quick scoring functions are not sufficient. For virtual screening of data bases consisting of thousands and more compounds and hit identification, in contrast, they are definitely more practical.

2.6 Ligand-based QSAR methods

Quantitative structure–activity relationship (QSAR) sometimes as well referred to as *quantitative structure–property relationship* (QSPR) methods follow an entirely different philosophy. In the context of binding affinity prediction for host–guest systems, the target

molecule is not taken into account. Instead, variations in biological activity BA are related to changes in some N suitable molecular properties p_i of the ligand molecule itself. That is why QSAR techniques first presented in the 1960s[103,104] are categorized as *ligand-based* methods. The underlying assumption is that similar molecules should have similar effects and that the properties under investigation determined by means of computational or experimental methods are obtained more efficiently than their biological activity using *in-vivo* or *in-vitro* experiments, probably after chemical synthesis.[105] After having determined the required properties of a molecule under observation, the relationship to its activity is generally described by a linear equation

$$BA = c_0 + \sum_{i=1}^{N} c_i \, p_i$$

where the coefficients c_i are calculated through a fitting procedure using statistical techniques. Due to an outstanding speed, they are, just as docking tools, particularly suitable for hit identification and possibly other early stages of drug discovery. Depending on the choice of descriptors, a QSAR model is not only useful for predictive purposes but as well for identifying and understanding molecular processes, suggesting new design strategies, narrowing the dose range for a planned assay, and revealing chemicals that deviate from the QSAR model.[105] A frequently used and one of the very first physicochemical properties is the water octanol/water partition coefficient ($\log P$) which is related to the compound's hydrophobicity. Basically, all descriptors in those days were determined through laboratory experiments or derived from the molecular topology (2D structure) leading to the term 2D descriptors. Later, when it became technically feasible and molecular modelling more and more popular, various 3D descriptors calculated on the basis of atomic coordinates were taken into account as well.[105] Examples out of a huge number of descriptors available in data bases and coming into question are either of a physicochemical, constitutional, functional, topological, or quantum mechanical type such as the number of atoms, hydrogen bond donors/acceptors, molecular volume, SASA (polar and nonpolar), atomic partial charges, electronegativities, ionization potentials, etc.[106]

In a preliminary step, a diverse set of training data with known activities defining the chemical space and covering a wide range of values is collected. After having carefully compiled that training set of compounds ideally associated small errors, QSAR model development performs two major steps: During *feature selection*, a set of descriptors significantly correlating with the activity is either collected manually or often extracted from large commercial data bases consisting of thousands of parameters by using machine learning or genetic algorithms. Genetic algorithms as used in this con-

text is related to recombining descriptors from two parent sets of descriptors, possibly followed by a point mutation, and comparing the new activity correlation with that of the two parents. Afterwards, the model is developed through linear (partial least squares, multiple linear regression) or non-linear methods (neuronal networks, support vector machines) used for mapping properties onto activities. Models (feature sets) that correlate well with biological activity and have been attested high stability using for instance leave-one-out cross-validation (LOOCV) are then chosen for the evaluation of new compounds.[105]

Fragment-based 2D-QSAR

In *fragment-based QSAR* as a prominent member of the class of 2D-QSAR methods, compounds are fragmented to a set of molecular substructures from which the activity is derived. These methods rely on the assumption that the activity is linearly related to fragment contributions to some molecular property.[105] Fragment-based QSAR methods show an obvious analogy to molecular fingerprints used for similarity searching in compound data bases. In an early approach, for instance, the overall $\log P$ value

$$\log P = \sum_i n_i \, a_i$$

was determined from individual atomic $\log P$ contributions a where n_i specifies the number of atoms of type i.[107] In a modern approach denoted as *hologram QSAR* (HQSAR), diverse structural fragments (linear, branched, cyclic, overlapping) of each molecule are split into bins of an fixed-length array forming a so-called molecular hologram. Corresponding occupancies of each of the N bins serve as compositional and topological descriptors p_{ij} of any compound j for a QSAR model

$$BA_j = c_0 + \sum_i^N c_i \, p_{ij}$$

with coefficients c_i developed using the *partial least squares* (PLS) method. The principle components-based PLS algorithm is particularly useful for model development whenever the number of available independent parameters (descriptors) is larger than the data set (number of molecules in the trainings set) which is as critical as typical for QSAR models. However, due to its speed and reproducibility, HQSAR is well suited for quickly prioritizing large sets of chemicals.[108]

3D-QSAR

The class of 3D QSAR methods mainly differs from 2D techniques due to a 3D representation of chemicals. They require atomic coordinates of reasonable conformers either determined by experimental or molecular mechanics methods. Usually, a preliminary knowledge-guided alignment of the compounds is required since the corresponding descriptors encode location-dependent structural characteristics.[105] 3D QSAR techniques are for these reasons significantly more expensive than 2D methods. One of the (if not the) most popular algorithms of this class is the *Comparative molecular field analysis* (CoMFA) first published by Cramer et al. by the end of the 1980s. It belongs to the *lattice-type* of 3D QSAR models since structurally aligned ligands of the training set are embedded in a 3D grid such that for each compound and grid point a steric (van der Waals) and electrostatic (Coulombic) field is calculated using an sp3 carbon atom probe with one positive charge adequately placed at the grid points. In fact, the data set alignment is performed in order to maximize the superposition of these fields. Afterwards, changes in these fields are related to changes in the activity using the PLS method. CoMFA was successfully applied to many biological systems and proved to be robust regarding partial charges and conformations of the training set. Rather, the methods used to determine charges and geometries should be applied consistently to the entire set of compounds. In contrast, the grid spacing as well as a proper data set alignment have a significant influence on the robustness of lattice-based methods yielding worse results if the grid was chosen too coarse (2Å spacing instead of 1Å).[105,109] A recent and promising extension of the CoMFA model denoted as *template CoMFA* comes up with an improved and automated alignment mode substantially reducing the requirement for manual work.[110] Other 3D QSAR models such as the *comparative molecular similarity indices analysis* (CoMSIA) basically differ in the type of descriptors used to describe chemical structures. However, corresponding regression models are commonly derived using PLS algorithms.[109]

QSAR technologies are continuously developed and applied to biological systems yielding excellent as well as elusive results.[13,109,111] Regarding binding affinity estimation of various host–guest systems, the best predictive results in terms of squared coefficients q^2 of LOOCV range between 0.6 and 0.7 in a study of Tosco et al.[13] which is quite impressive in the light of being a ligand-based method. However, it seems somewhat surprising that one of the squared coefficients presented in this work is negative indicating either a lapse or the usage of non-squared correlation coefficients. Various QSAR strategies (HQSR, CoMFA, CODESSA) applied to the estrogen receptor yielded q^2 values ranging between 0.5 and 0.7 depending on both the underlying data set

and QSAR model.[108,109] It should be noted, however, that the data set alignment was guided by binding mode information available from several PDB crystal structures.[109] Interestingly, the two correlation coefficients associated with model fitting on the one hand (r^2) and cross validation on the other (q^2) diverge substantially in case of QSAR models. In most cases, q^2 is more than 30% less than r^2 indicating an overfitted model. Overfitting generally poses a great danger in QSAR studies since, usually, the number of descriptors exceeds the number of compounds in the training set.

2.7 Summary

Apart from the physical basis of binding free energies by means of classical thermodynamics and statistical mechanics we have elaborated a classification of popular methods developed for their calculation. The common methodology ranges from very fast and approximative up to accurate as well as costly techniques. In this respect, there is a fundamental trade-off between computational costs versus accuracy and different classes of techniques are particularly convenient for different purposes.

In early drug discovery stages including hit identification and lead discovery, one would typically screen large databases often consisting of thousand of chemicals with respect to their binding affinity to some interesting target protein. Most convenient techniques for this task are target-based molecular docking scoring functions and ligand-based QSAR methods due to an outstanding speed, though they are known for less accuracy. In contrast, promising lead candidates for further development are then better examined by methods incorporating time averages from MD or MC simulations in order to achieve more accurate results. From this class of methods we have discussed the empirical LIE model as well as two continuum-solvent strategies. Just as QSAR models, LIE requires a set of training data used for weighting one or both interaction potentials and possibly further terms whereas MM/PBSA and MM/GBSA do not. However, both strategies are based on MD samplings from the binding reaction's end states only making them impractical for virtual screening. Many publications have attested them substantial reliability in terms of lead optimization as well as toxicity estimation. Even more accuracy is commonly accepted for thermodynamic work/path methods. They have in common extensive samplings of many additional intermediate states lining a chemical reaction path. According to a thermodynamic cycle, these non-parametric methods are able to directly provide relative binding free energies associated with alchemical transformations. Their suitability for processing large libraries, in contrast, is (still) extremely limited due to computational and time complexity.

Insofar, end state sampling methods yield an optimal trade-off between accuracy and speed. In the framework of this thesis, particular emphasis is set on the development and application of high-quality LIE models for a couple of host–guest systems and purposes. Chapter 5 describes the prediction of the elution order regarding HPLC using a simple LIE approach. A similar approach to the prediction of protein–ligand binding affinities will be presented in Chapter 6. This LIE model was extended by an Monte-Carlo entropy estimator published recently and further constitutional descriptors as known from QSAR technologies. In a final application (Chapter 7) the LIE model descriptors are used for a toxicological prioritization of transformation products despite (and due to) the lack of training data.

All these calculations presume the presence of one or more reasonable binding poses. And all binding affinity calculation is at most as performant as the exploration of the space of binding poses. Significant changes in the (relative) conformation of a host–guest complex are rare events during an MD simulation such that predetermined initial binding poses ideally cover a wide range of space. Currently, docking algorithms provide (apart from few complex experimental methods) the only computational access to binding pose prediction. Since mainly either a fragment-based growth mechanism or some random pose generator is utilized, a simple systematic binding mode generator and other approaches to both exploration and clustering of the conformational space will be elaborated in Chapter 4.

3 Methodological background of atomistic force field simulations

During the second half of the 20'th century, *in silico* experiments and, in particular, molecular mechanics simulations emerged as popular standard tools for the investigation and understanding of (bio)molecular structure, function and dynamics on a microscopic level.[112] In general, the classical atomistic simulation of many-body systems is based on classical equations of motion describing interatomic forces also referred to as *molecular mechanics* (MM) and, consequently, on the time-evolution of the constituting atoms using either deterministic methods like MD or stochastic techniques such as Monte Carlo (MC) methods. From the collection of sampled geometries and respective potential plus kinetic energies one can derive many thermodynamic quantities as well as structural observations. Since in macromolecular, biochemical systems structure determines functionality and due to the mechanism of molecular recognition,[113] many biological tasks including protein (un)folding, ligand binding, signalling pathways, gene regulation and catalytic processes can be investigated on the basis of theoretical results. In particular, the possibility to quantitatively estimate binding affinities of host–guest systems allows to significantly reduce the complexity of laboratory experiments needed for virtual screening, drug design, and toxicological studies.[100,114,115] Data obtained from MD simulations is useful for kinetic investigations of molecular processes, too. After having determined metastable[1] sets, i. e., almost invariant subsets of the conformational space,[116] it is possible to compute transition rates between these conformers.[117]

Classical simulations require a *parameterization* of the molecular system, i. e. the assignment of predetermined parameters quantifying the strength of various types of physical interactions between the particles under consideration. 3D coordinate files of the involved molecules making the *molecular topology* evident are usually given as input. A consistent set of force field parameters along with a functional form for the

[1] In contrast to other definitions, metastability throughout this thesis is associated with a stable (set of) molecular conformation represented by considerable energetic minimum for which the probability to switch to another metastable set is very small whereas the probability of remaining inside is high

potential energy $U(q)$ of a geometry specified by its atomic coordinates q is denoted as *empirical force field* (FF). Low energy regions of the potential energy function are expected to correspond to states populated preferentially at thermal equilibrium. Furthermore, the energy gradient, that is partial derivatives of the energy function U with respect to atomic coordinates, yields forces acting on the particles.[112] The combination with a convenient concept of mechanics enables to predict and investigate atomic motion (molecular dynamics). Having introduced the mathematical basis of (molecular) mechanics as well as empirical molecular mechanics FFs using by way of example the popular Amber force field, this chapter will describe deterministic and random models for sampling molecular geometries from the Boltzmann distribution. In addition, common algorithms used for energy minimization, molecular dynamics simulation, FF parameterization including partial charge assignment, and any other purpose relevant for this thesis will be thoroughly discussed. Insofar, this chapter must be considered as the central methodological part of the entire thesis as most of the techniques are employed in the one or other upcoming chapter.

3.1 Classical mechanics

Classical mechanics is a physical field concerning with the motion of bodies under the influence of forces. Sir Isaac Newton was the first scientist to develop a rigorous mathematical framework for the investigation of such systems. Newton mechanics and, in addition, some reformulations with practical relevance constitute the physical basis for classical MD simulations. For this reason, the theory of classical mechanics is briefly sketched in the following sections.

Newton mechanics

In the seventeenth century, Sir Isaac Newton carried out intensive studies of the motion of bodies resulting in the *laws of motion* that are still fundamental for classical as well as quantum mechanics. Using his theoretical framework, it was possible to predict the spatial position $r(t) = (x(t), y(t), z(t))$ and motion of bodies (planets, cannon balls, atoms, etc.) more or less accurately as a function of time t.[51] Furthermore, the conservation of energy and (angular) momentum can be derived from Newton's laws. In principle, the acceleration a, equaling the second-order time-derivative $\ddot{r}$ of the position of a body/particle with mass m, is considered proportional to the sum F (net force) of

all conservative force vectors acting on it

$$F = ma = m\ddot{r} = m\frac{\mathrm{d}^2 r}{\mathrm{d}t^2}. \tag{3.1}$$

Equation 3.1 belongs to the class of *equations of motion* that completely describe the development of mechanistic systems in time and space.[51] Since r is most likely a vector associated with, say, $3N$ dimensions, Equation 3.1 is better expressed in terms of the potential's gradient

$$ma = -\nabla_q U(r) = \left[-\frac{\partial U(r)}{\partial r_1}, \cdots, -\frac{\partial U(r)}{\partial r_{3N}}\right] \tag{3.2}$$

that is partial derivatives with respect to coordinates r_i. A *uniform motion* $v(t)$ as addressed by Newton's first law is characterized by zero acceleration, $a = 0$, due to zero net forces, $F = 0$, resulting in a constant velocity v or time-derivative $\dot{r}$

$$v(t) = \dot{r}(t) = \frac{\mathrm{d}r}{\mathrm{d}t} = \text{const}$$

of the body's position. An instant position would then be computed as the sum of a term representing the uniform motion during time t and the initial position $r(t_0)$ at the beginning t_0 of the observation,[51]

$$r(t) = r(t_0) + t\dot{r}(t). \tag{3.3}$$

In contrast, if the forces do not sum up to zero, the body under consideration expierences an acceleration $\ddot{r} \neq 0$ proportional to F as stated by the second law. One would observe a constantly *accelerated motion* where the position at time t is obtained in analogy to Equation 3.3 but with an additional quadratic term accounting for the acceleration

$$r(t) = r(t_0) + t\dot{r}(t_0) + \frac{t^2}{2}\ddot{r}(t_0). \tag{3.4}$$

Equations of motion are used in the field of molecular mechanics as well where the evolution of a system's microstate is determined by the change in atomic coordinates and velocities or, equivalently, momenta $p = mv$. Equation 3.1 is a second-order time-differential equation where time is considered in terms of the second derivative $\frac{t^2}{2}\frac{\mathrm{d}^2 r}{\mathrm{d}t^2}$ only. As a consequence, velocities are not considered directly, but need to be derived from the microstate's coordinates by integration. Consequently, the solutions for t and $-t$ are always equal satisfying the property of *time-reversibility*. However, this property does not hold in general since some processes of an isolated system are irreversible as stated by the second law of thermodynamics. For this and other reasons some reformulations and generalizations of Newtons equation of motion shaped up as more convenient when it comes to an MD sampling of microstates.

Lagrangian mechanics

From a physical point of view, Newtonian mechanics is valid for all classical mechanistic systems and particularly suitable in combination with Cartesian coordinates. However, solving Newton's equation of motion for other coordinate systems quickly becomes cumbersome. Due to its independence from the underlying coordinate system, a reformulation denoted as *Lagrangian mechanics* shapes up as much more universally usable such as, for instance, regarding polar coordinates (angles). Apart from an easier determinability of its equations of motion, it allows to include further physical/mathematical constraints and is useful for other physical fields than classical mechanics including electromagnetism and relativity.[24,51] The central Lagrangian function

$$\mathcal{L}\left(q,\dot{q}\right) = K\left(\dot{q}\right) - U\left(q\right) \tag{3.5}$$

is expressed as the difference between kinetic $K\left(\dot{q}\right)$ and potential energy $U\left(q\right)$ that are depending only on velocities $\dot{q}$ and, respectively, generalized coordinates q. The Lagrange equations

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial\mathcal{L}}{\partial\dot{q}_i} - \frac{\partial\mathcal{L}}{\partial q_i} = 0 \tag{3.6}$$

are derived from the *principle of least action*. Plugging the Lagrangian (Equation 3.5) for some particular mechanistic system into Equations 3.6 and solving the latter yields the desired equations of motion.[51] The SHAKE algorithm briefly sketched in Section 3.7 uses a Lagrangian for constraining covalent bonds.

Hamiltonian mechanics

The *Hamilton* formalism of mechanics developed by the Irish scientist William Rowan Hamilton provides another elegant instrument particularly useful for the description of a molecular system's time evolution and mostly used for MD calculations. It is derived from the Lagrangian reformulation of classical mechanics through a Legendre transformation. Therefore, the Hamiltonian formulation as well is compatible with any type of coordinate system.[24] Time is taken into account, in contrast to Newton's formulation, already with the first derivative of the equations of motion. The central quantity of the classical time-independent Hamiltonian formulation for a system consisting of N particles is the total energy as the sum of kinetic and potential energy contributions

$$\mathcal{H}\left(q,p\right) = U\left(q\right) + K\left(p\right). \tag{3.7}$$

The Hamiltonian of any system investigated throughout this thesis is of a time-independent form. As usual, the potential energy depends on generalized coordinates q

only whereas the kinetic energy is as a function of conjugated momenta p only. We will discuss the construction of $U(q)$ for the purpose of MD simulations in detail in Section 3.2 when it comes to the concept of classical force fields. The kinetic energy is defined through particle masses and momenta as sketched by Equation 2.15 or, equivalently, using velocities v

$$K(v) = \frac{1}{2}Mv^2 \tag{3.8}$$

with the diagonal matrix $M \in \mathbb{R}^{3N\times 3N}$ of masses m_i of each atom i out of N atoms[118]

$$M = \begin{pmatrix} m_1 & & & & & & \\ & m_1 & & & & \mathsf{O} & \\ & & m_1 & & & & \\ & & & \ddots & & & \\ & & & & m_N & & \\ & \mathsf{O} & & & & m_N & \\ & & & & & & m_N \end{pmatrix}. \tag{3.9}$$

Partial derivatives of Equation 2.14 with respect to p and q yield the equations of motion, that is two first-order differential equations describing the change of q and, respectively, p over time

$$\dot{q} = \frac{\mathrm{d}q}{\mathrm{d}t} = \frac{\partial \mathcal{H}}{\partial p}; \quad \dot{p} = \frac{\mathrm{d}p}{\mathrm{d}t} = -\frac{\partial \mathcal{H}}{\partial q} \tag{3.10}$$

according to the Hamilton formulation.[24] Theoretically, a molecular system's evolution in phase space could be calculated analytically using Equations 3.10. Unfortunately, no general analytical solution is known for the integration of equations of motion for systems consisting of three or more coupled particles, an issue that is referred to as *three-body problem*.[119] Hence, the integration of q and p over time for the construction of trajectories requires numerical integrators such as *leap-frog*[120] or other algorithms of the *verlet* class[121] that are introduced in Section 3.5. From the equations of motion (Equation 3.10) of the time-independent form follows the conservation of the Hamiltonian over time[24]

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{H}(q,p) = \frac{\partial \mathcal{H}}{\partial q}\dot{q} + \frac{\partial \mathcal{H}}{\partial p}\dot{p} = \frac{\partial \mathcal{H}}{\partial q}\frac{\partial \mathcal{H}}{\partial p} - \frac{\partial \mathcal{H}}{\partial p}\frac{\partial \mathcal{H}}{\partial q} = 0. \tag{3.11}$$

The property of energy conservation copes with the first law of thermodynamics holding for isolated systems. In order to provide for that property, the numerical integrator chosen for the MD sampling needs to be *symplectic*, that is to be able to preserve the phase space volume (*Liouville's Theorem*).[122] As a consequence of this property, all microstates sampled during one MD simulation based on Hamilton dynamics are associated with the same total energy on average. This is depicted in Figure 2.1 by the use of an isosurface, a constant energy contour plot integrated in the PES of pentane.

Though, the contour plot is for the sake of illustration only, because Figure 2.1 represents the configurational space instead of the phase space including momenta. Since the resulting trajectory remains on one isosurface in phase space preventing an extensive sampling of the conformational space, modifications (discussed in Section 3.6) of the equations of motion are required if one desires to obtain molecular geometries with varying energies distributed in accordance with a particular constant temperature. Corresponding statistical ensembles associated with differing physical boundary conditions such as constant temperature have been sketched in Section 2.2.

3.2 Classical molecular mechanics force fields

Several force fields have been developed during the past decades that are, among others, characterized by different parameter sets, mathematical functions, or the type of systems they are convenient for. All-atom FFs (AMBER,[123,124] CHARMM,[125] OPLS-AA[126], MMFF[127]) consider each atom explicitly whereas united atom FFs such as OPLS-UA[128] and GROMOS[129] combine, for computational efficiency reasons, methyl(ene) hydrogens with the carbon atoms they are bonded to resulting in one effective particle. Even more efficiency-driven abstraction of groups of atoms is provided by so-called *coarse-grained* models such as the MARTINI[130] force field combining more atoms to one effective particle. Some FFs take into account the polarizibility of atoms (X-Pol[131], PIPF[132]) others do not. Some (MMFF[127]) are designed for the simulation of small drug-like organic molecules in vacuum only[133] others for macromolecular systems containing biopolymers (DNA, proteins), probably embedded in cell membranes, and explicit water (AMBER,[123] GROMOS,[129] OPLS,[126] CHARMM[125]). In principle, all empirical FFs are designed in a similar fashion largely representing the same types of atomic interactions by similar mathematical functions.[112] Thanks to its functional form,

$$U = f(q),$$

a FF can be considered as a definition of a potential energy surface in configurational space as it provides the potential energy U as a function of coordinates q. Throughout this thesis, the *AMBER99SB* force field[134] of a family of FFs denoted as *Assisted Model Building with Energy Refinement* (AMBER) and the *Merck Molecular Force Field* (MMFF) were used for all classical MD and, respectivly, MC simulations. This decision was encouraged by AMBER's particular capability of biological condensed phase systems including protein–ligand complexes[134] on the one hand and the suitability of MMFF regarding small organic molecules such as typical ligands[127] on the other. On

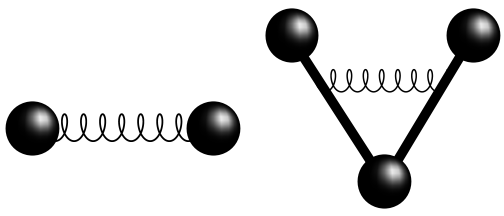
that account, the concept of classical FFs will be depicted based on AMBER. First of all, chemical elements are distinguished with respect to their chemical context resulting in a set of carefully defined *atom types*. In particular, these types are characterized by their orbital hybridization as well as the elements/atom types they are bound to. A double bond between two sp2-hybridized carbon atoms, for example, is shorter in average than one associated with two sp3-hybridized carbons, but longer than the double bond of a keto group. Finally, the FF contains a predefined atomic *van der Waals radius* for each element. Prior to the simulation process, partial atomic charges need to be calculated. We will deal with common charge estimation methods in Section 3.3. Ensuing from these basic specifications, a typical FF comprises a couple of simple mathematical functions quantifying the interaction (potential energy) between two or more atoms that may be covalently bound (bonded terms) or not (nonbonded terms). Parameters for these functions (also referred to as *potentials*), mainly originate from X-ray crystallographic and spectroscopic experiments as well as high-level quantum mechanical calculations and were probably fitted in order to reproduce macroscopic physicochemical quantities like density and heat capacity.[15,112]

Harmonic representation of bonds and angles

The most common (as in case of the AMBER FF) physical model for the representation of a covalent bond between two atoms of type i and j oscillating around a reference bond length r_{ij} is a *harmonic oscillator* best represented by a spring as depicted in Figure 3.1.[15] According to *Hooke's law*

$$F = Kx, \tag{3.12}$$

the force F extending/compressing the bond (spring) is linearly correlated with the deviation $x = r - r_{ij}$ from the reference distance. The spring constant K serves as a material-specific proportionality constant characterizing the spring's stiffness. Integrat-



(a) bond (b) angle

Figure 3.1: Bond stretching and angle bending modelled using a harmonic oscillator following Hooke's law.

ing Equation 3.12 over x yields the harmonic energy function

$$U^{\text{bond}}(r) = \frac{K_{ij}}{2}(r - r_{ij})^2 \tag{3.13}$$

also denoted as *bond potential* in terms of molecular mechanics. $U^{\text{bond}}(r)$ quantifies the potential energy taken up by the bond (in addition to the bond-dissociation energy that is not considered by classical FFs). The force constant K_{ij} associated with the bond stiffness and r_{ij} specifying the preferential interatomic distance are typical examples for molecular mechanics FF parameters.[123] The angular oscillation frequency of the bond length is defined as

$$\omega = \sqrt{\frac{K_{ij}}{\mu}}, \tag{3.14}$$

with the *reduced mass*

$$\mu = \frac{m_i\, m_j}{m_i + m_j}$$

and the masses m_i and m_j of the atoms forming the bond.[135] Figure 3.2 shows the influence of the two force field parameters on the parabolic energy function. Large force constants K_{ij} yielding tighter parabolic shapes corresponding to stiffer/stronger bonds ($\mathrm{C} - \mathrm{C}$ single bond is weaker than $\mathrm{C} = \mathrm{C}$ or $\mathrm{C} \equiv \mathrm{C}$ bonds[15]). The concept of reduced masses ensures that the oscillation frequency is particularly dependent on the lighter atom and, thus, remains low unless the masses of *both* involved atoms are large. That is to say, even if the light hydrogen atom which is associated with the highest frequencies is covalently bound to a heavy atom (heavier than hydrogen itself in terms of structural bioinformatics), the oscillation frequency of the resulting bond will amount to a high value as expected for hydrogens. Besides, it is due to the relationship depicted in Equation 3.14 that force constants K_{ij} can be estimated on the basis of experimental methods such as infrared or Raman spectroscopy by measuring molecular vibrational frequencies.[135] The energy minimum is associated with the reference length r_{ij} which decreases from a single bond towards a triple bond (with respect to the same pair of elements). Bond breaking and forming are not supported by the harmonic oscillator[15]. Due to its symmetry properties, stretching a bond yields the same energy penalty as compressing it by the same extent. However, this is not realistic since compressing the distance of two atoms down to extremely low values close to zero would unavoidably lead to nuclear fusion requiring significantly more energy than stretching it which would result in a less energy-demanding bond breaking. From this perspective, a more realistic model for covalent bonds is constituted by the asymmetric *Morse potential* implemented as an alternative to the harmonic oscillator in few FFs.[136–138] Nevertheless, due to both its far less computational effort and a reasonable approximation of the Morse potential in the vicinity of the minimum,[15] most FFs including

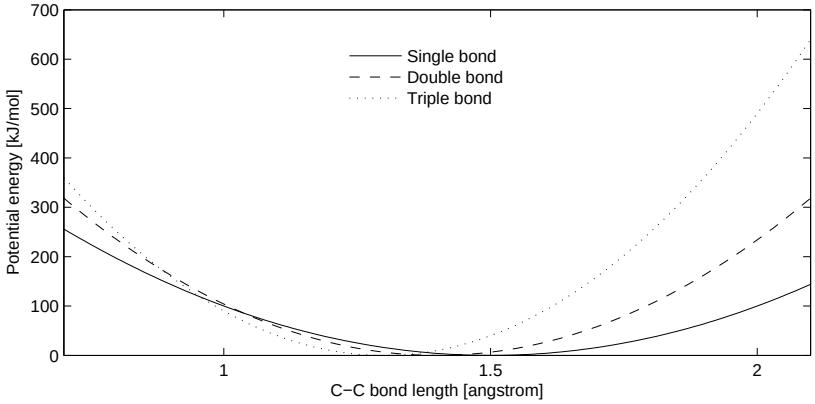


Figure 3.2: Harmonic bond potentials of carbon-carbon single, double, and triple bonds as incorporated in classical molecular mechanics force fields.

Amber and MMFF have implemented the harmonic potential. On this account, the Morse potential will not be discussed in more detail here, though, its functional form is very similar to that depicted in Figure 3.7.

As signified by Subfigure 3.1(b), harmonic potentials are suitable for modelling angle bending as well. In contrast to the harmonic bond representation, this potential

$$U^{\text{angle}}(\theta) = \frac{K_{ijk}}{2}\left(\cos\theta - \cos\theta_{ijk}\right)^2, \tag{3.15}$$

is a function of the angle θ spanned by the two bond vectors $v_{ij} = v_i - v_j$ and $v_{kj} = v_k - v_j$ defined by the three involved atoms (i, j, k) with respective position vectors (v_i, v_j, v_k). The inner product of these two bond vectors yields θ required by Equation 3.15 for energy calculation

$$\cos\theta = \frac{\langle v_{ij}, v_{kj}\rangle}{\|v_{ij}\|\,\|v_{kj}\|}.$$

Proper and improper dihedral potentials

Due to steric hindrance caused by repulsive forces between atoms that approach each other too much, the rotation about a chemical bond defined by two atoms j and k (see left structure of Figure 3.4) is usually associated with rotational barriers requiring additional energy. Regarding the central bond of butane this is illustrated by the *Newman projection* in Figure 3.3. *Eclipsed* conformations of butane (methyl group in the foreground of Figure 3.3 at 0°, 120°, or 240°) are less favorable than *staggered* conformations (methyl group at 60°, 180°, or 300°) which are characterized by larger distances from the binding partners of atom j to those of atom k as depicted in the left structure of Figure 3.4. As a consequence, a complete 2π rotation reveals a typical periodic energy

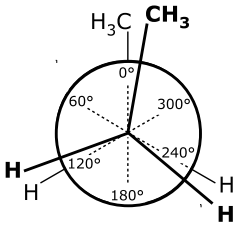


Figure 3.3: Newman projection of the central butane dihedral in the eclipsed syn-conformation.

profile depending on both the number and types of neighbors covalently bound to the atoms j and k. In order to meet these requirements, classical FFs utilize one or more cosine expressions per torsional degree of freedom. A torsion (or dihedral) angle

$$\phi = \arccos\left(\frac{\langle v_{ij} \times v_{jk}, v_{jk} \times v_{kl}\rangle}{\|v_{ij} \times v_{jk}\| \, \|v_{jk} \times v_{kl}\|}\right)$$

is defined as the angle between two planes P_{ijk} and P_{jkl} constructed on the basis of a set of four consecutive atoms (i, j, k, l) which form three consecutive bonds (bond vectors) v_{ij}, v_{jk}, and v_{kl} (see left structure of Figure 3.4 for illustration). P_{ijk} is specified by the bonds v_{ij} and v_{jk} corresponding to the first three atoms and the second plane P_{jkl} by v_{jk} and v_{kl} corresponding to the last three atoms of the quadruple. According to the *General Amber Force Field* (GAFF) for small molecules, the torsional potential

$$U^{\text{dihedral}}(\phi) = \frac{K_n}{2}\left(1 + \cos\left(n\phi - \gamma\right)\right) \tag{3.16}$$

incorporates one cosine term and the force constant K_n defining the potential's height.[123] The number of minima in the course of a complete 2π rotation and the phase shift are adjusted by multiplicity n and, respectively, by γ. By way of example, the carbon–carbon dihedral model ethane ($H_3C - CH_3$) revealing three torsional minima with

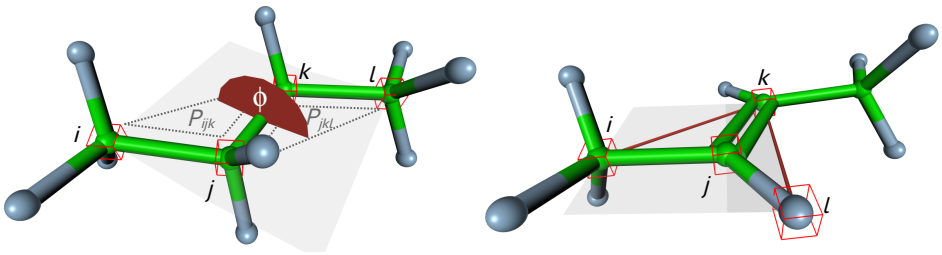


Figure 3.4: Left: Torsion angle ϕ of butane spanned by two planes $P_{i,j,k}$ and $P_{j,k,l}$ defined by atoms (i, j, k) and (j, k, l), respectively. ϕ serves as a measure for the relative position of the atoms i and l due to rotation about the j-k bond. **Right**: improper dihedral used to achieve planarity of sp2-hybridized atoms.

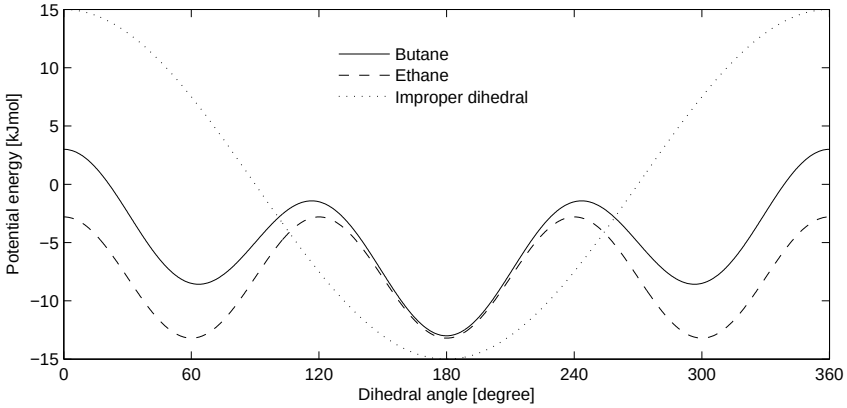


Figure 3.5: Torsion potentials with multiplicities equal to one (improper dihedral) and three (proper dihedral), respectively. All minima (staggered conformations) of ethane have the same height due to rotational symmetry.

equal energies separated by three barriers of equal height requires $n = 3$ and $\gamma = 180°$. Figure 3.5 shows the torsion potential of ethane. In contrast, the energy profile of butane is characterized by differing levels of minima and barriers due to the substitution of one hydrogen for one methyl group regarding both ethane carbons. This requirement cannot be fulfilled with one single cosine function but demands a sum of multiple terms. In AMBER99SB optimized for protein simulations three terms are utilized per dihedral[134]

$$U^{\text{dihedral}}(\phi) = \sum_{n=1}^{3} K_n \left(1 + \cos\left(n\phi - \gamma_n\right)\right),$$

whereas other FFs might have implemented the *Ryckaert-Bellemans potential*[139] proposing six cosine terms with increasing power

$$U^{\text{dihedral}}(\phi) = \sum_{n=0}^{5} C_n \left(\cos\left(\phi\right)\right)^n.$$

In contrast to proper dihedrals, *improper dihedrals* are not related to rotations about a bond but intended to force planarity regarding conjugated systems and sp2-hybridized atoms as depicted for butene on the right of Figure 3.4. Planarity of the covalent environment of an sp2-hybridized atom j is achieved using a single cosine expression as expressed by Equation 3.16 with exactly one minimum ($n = 1$) at $\gamma = 180°$. The potential of a typical improper dihedral is depicted by the finely dotted line with a high energy barrier culminating at $0° = 360°$ in Figure 3.5.

Nonbonded potentials

Due to combinatorics, the computationally most intensive type of atomic interplay is related to pair-wise nonbonded interactions associated with atoms that are either not covalently bound to each other or have at least three bonds in between. Given a molecular system consisting of N atoms the simulation software has to deal with $\mathcal{O}(N^2)$ evaluations of *pair potentials* theoretically, though, this number can be reduced significantly by using some algorithmic tricks such as neighbor lists.[24] Physical interactions commonly modelled through nonbonded potentials mainly originate from electronic or *van der Waals forces* that are basically either of a repulsive or attractive nature. Repulsive electrostatic interactions between two particles i and j associated with partial charges z_i and z_j are caused by identical charge signs whereas attractive forces are due to opposite signs. Both electrostatic cases are taken into account by the *Coulomb potential*

$$U^{\text{Coul}}(r) = \frac{z_i\, z_j}{4\pi\varepsilon\varepsilon_0 r}.$$

derived from *Coulomb's law* which states that the force F acting on two interacting particles due to (partial) charges is proportional to both the product of these charges and the inverse square of their distance. The *relative permittivity* or *dielectric constant* ε quantifies the amplifying or easing affect of the bulk medium on an electric field compared to the permittivity $\varepsilon_0 = 8.85\,\text{C}^2\,\text{J}^{-1}\text{m}^{-1}$ of vacuum. Figure 3.6 shows a monotonically decreasing plot (continuous line) of a repulsive Coulomb potential for two point charges with identical signs ($z_i\, z_j > 0$) and a monotonically increasing plot describing attractive electric forces ($z_i\ z_j < 0$) yielding (favorable) lower energies at

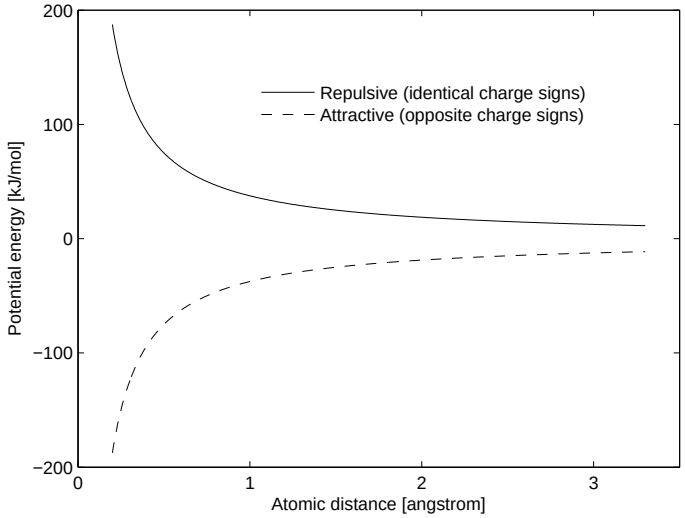


Figure 3.6: Exemplary Coulomb potential of classical molecular mechanics force fields.

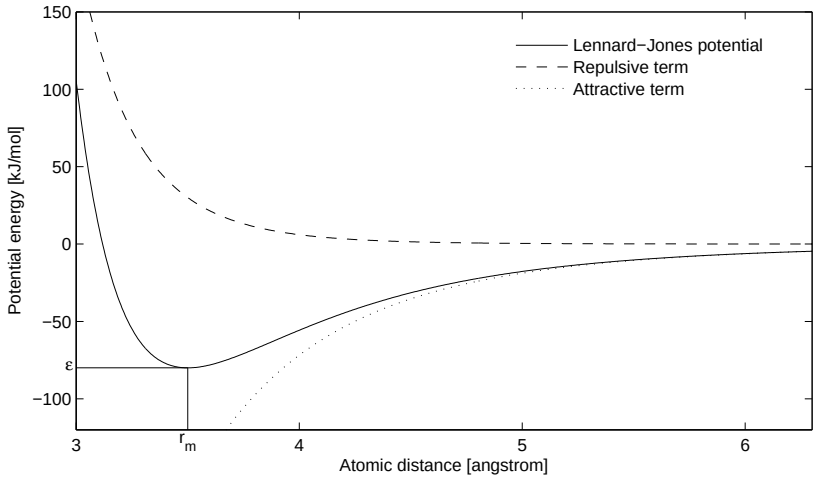


Figure 3.7: Lennard-Jones potential of classical molecular mechanics force fields.

small distances.

The Coulomb potential neither reveals an optimal distance nor does it take into account interactions due to other than electric forces. If there was, apart from the one associated with this potential, no other force acting on two nonbonded particles, the atoms would inevitably collide and share the same postion in space in case of opposite signs. Conveniently, another pair potential for nonbonded interactions commonly used in molecular mechanics simulations was proposed by John Lennard-Jones in the year 1924.[140] The *Lennard-Jones potential*

$$U^{\mathrm{LJ}}(r) = \varepsilon \left[\left(\frac{r_m}{r} \right)^{12} - 2 \left(\frac{r_m}{r} \right)^{6} \right] = \frac{A_{ij}}{r^{12}} - \frac{B_{ij}}{r^{6}} \tag{3.17}$$

is composed of two additive terms inversely proportional to some particular power of the distance r between two atoms. Repulsive forces at short ranges due to overlapping electron orbitals (Pauli exclusion principle) are usually expressed by the r^{-12} term, whereas the r^{-6} term describes long-range attractive *van der Waals forces* arising from induced dipole moments.[15] Both terms are sketched in Figure 3.7 along with their sum $U^{\mathrm{LJ}}(r)$. Slight modifications yield the expression on the right hand of Equation 3.17 where the two parameters specifying the optimal distance r_m of atom i and atom j and the potential energy ε associated with it were transformed into the AMBER force field parameters A_{ij} and B_{ij}. As in case of any other pair potential parameter, they are specific for the set of atom types (i, j) involved in the potential.

Potential energy function

For a given set of generalized coordinates q representing a particular molecular geometry, an additive force field expresses its potential energy

$$U(q) = \sum_{\text{bonds}} \frac{K_{ij}}{2}(r - r_{ij})^2 + \sum_{\text{angles}} \frac{K_{ijk}}{2}(\cos\theta - \cos\theta_{ijk})^2 + \sum_{\text{dihedrals}} \frac{K_n}{2}(1 + \cos(n\phi - \gamma)) + \sum_{\text{pairs}} \left[\frac{z_i z_j}{4\pi\varepsilon\varepsilon_0 r} + \frac{A_{ij}}{r^{12}} - \frac{B_{ij}}{r^6}\right] \quad (3.18)$$

as the sum of all bonded (bonds, angles, dihedrals) and nonbonded potentials.[124] Distances r and angles ϕ/θ used in Equation 3.18 are functions of q as well. Equation 3.18 can be considered as a formal definition of the PES in the configurational space spanned by the molecular system's generalized coordinates q as described in Section 2.2.

During a typical conformational sampling process all types of potentials comprised by U are taken into account. The molecular conformation itself, however, is almost exclusively specified by dihedral angles since other bonded potentials (bond length and angles) reveal only small variances around their reference values.[24] Already in the previous chapter we have seen that the two nonbonded potentials are additionally useful for binding affinity estimations on the basis of atomistic MD time series regarding two (or more) distinct molecules as in case of host–guest systems. Thus, we will further deal with nonbonded interaction potentials in the context of binding affinity models in Chapters 5-7.

3.3 Partial charge estimation

As we have just seen within the scope of force field potentials, the computation of pairwise electronic interactions depends on partial atomic charges z_i. The preliminary determination of reasonable partial charges is a very crucial step in classical molecular mechanics and conformational analysis since nonbonded interactions are mainly attributed to electrostatic contributions.[141] Both structural and dynamical behaviour of polar molecules are substantially affected by partial charges. In particular, this includes the tertiary/quartery structure of biological macromolecules such as proteins and nucleic acids as well as their interactions with ligands, the aquatic environment and dissolved ions.[142,143]

However, one has to bear in mind, that the partial charge of an atom is no static physical quantity that can be measured exactly but is derived from the distribution

of electrons about a molecule's atoms. Partial charges are dedicated to mimic certain properties of the continuous electron distribution of a molecule such as the electrostatic potential (ESP), the molecular charge density, or the dipole and higher electric moments.[141] The electronic distribution in turn strongly depends on the atom's or, respectively, molecule's polarizability, chemical environment and conformation. Due to a substantially higher electronegativity of oxygen compared to carbon, for example, a hydrogen atom bound to an oxygen atom reveals a significantly higher (positive) partial charge than a hydrogen bound to carbon. Nevertheless, as an approximation, partial charges for the purpose of classical force field simulations are usually determined once prior to the sampling process and kept constant during it. It is, therefore, advisable to determine partial charges on the basis of dominant conformations with outstanding statistical weights, preferably the global potential energy minimum conformation of a molecule if assessable. An according strategy for the determination of favorable geometries is therefore presented in Section 4.1.

Several computational techniques have been developed for the estimation of partial charges that, according to Heinz et al., mainly fall into two general categories: quantum mechanical *ab-initio* and semi-empirical methods.[142] The former type of methods first performs an *ab-initio* calculation of the electron density, by applying the Hückel theory, a Hartree-Fock-based method, or density functional theory, etc. followed by a partitioning in atomic basins, e. g. using Mullikan's population analysis[144] or Bader's gradient method.[145] According to Reynolds,[146] these traditional *ab-initio* charge estimation methods are either highly unreliable (Mullikan charges) or computationally too expensive. In particular, they strongly depend on the QM wave function (basis set) chosen.[142,147] The second group of partial charge estimation methods that is referred to as semi-empirical *charge equilibration* by Heinz et al.[142] is attributed to a method described by Rappé and Goddard in 1991.[148] Charge equilibration is related to electronegativity information of individual atoms which in turn depend on atomic ionization potentials, electron affinities and atomic radii. Usually, the semi-empirical charge equilibration method yields better results than *ab-initio* techniques. However, covalent bonding contributions to the cohesive energy in organic molecules are not considered.[142] In contrast, Halgren's bond topology-based approach as part of MMFF suitable for organic compounds makes use of precalculated bond charge increment (BCI) parameters.[127] As a result, no quantum chemical calculations are required at all for partial charge assignment upon parameterization of a compound[141] what makes the BCI approach exceedingly fast and suitable for large virtual data base screenings. Another purely empirical and fast approach to the assignment of partial charges that entirely avoids quantum chemical computations was developed by Gasteiger by the end of the

seventies. This method applies an iterative partial equilization of atomic electronegativities derived from ionization potentials and electron affinities regarding the atom under consideration plus its neighbors and tries to approximate Mullikan charges.[147,149]

Due to an excellent reproduction of solvation free energies of organic molecules,[150,151] charges fitted to the ESP calculated on the HF/6-31G* level for a large number of grid points encompassing the molecule of interest emerge particularly suitable for condensed phase simulations. Unfortunately, ESP-fit charge monopoles strongly depend on the molecule's conformation[146] and show numerical unstabilities related to the charge magnitude of buried atoms. On the latter issue one got a grip by restraining the magnitude of partial atomic charges in particular of nonpolar groups,[151] thereby, keeping the suitability for solvation free energy and intramolecular conformational energy calculation.[152] In order to fix the conformation dependency problem, the advocates of *ab-initio* ESP-based methods suggest using Boltzmann-weighted sums over a wide range of conformations.[146] Again, this procedure substantially increases the computational effort of charge estimation for new compounds already emerged from the *ab-initio* calculation of ESPs. A significantly faster semi-empirical approach to the estimation of ESP-based charges denoted as AM1-BCC was presented by Jakalian et al.[141] The AM1 bond charge correction (BCC) model brings atomic charges from the relatively fast semi-empirical AM1 method together with Halgren's BCI approach and the high ESP charge quality. It performs in two major stages: the first step consists of an AM1 population analysis capturing formal charges and electron delocalization. Afterwards, these charges are modified using a simple bond charge correction (BCC) term per atom that only depends on the type of this atom and its immediate neighbors. A training set of several thousand chemicals had been used in order to parameterize BCC parameters against the ESP on the HF/6-31G* level. AM1-BCC charges seem very convenient for condensed phase simulations of various polar, nonpolar, and aromatic systems[141] particularly in combination with the AMBER force field which was incorporated for intensive validation and further fitting of BCC parameters.[153] Since all condensed phase simulations of host–guest systems performed within the framework of this thesis employ AMBER force fields, the AM1-BCC method puts itself for charge assignment. In addition, partial charge estimation and topology parameter assignment to new compounds is straightforwardly done in one go by the program *Antechamber* included in the AmberTools package.

3.4 Potential energy minimization

As stated before, scientists are often interested in *critical points* of the potential energy landscape (PES) such as local/global minima and saddle points (illustrated by Figure 3.8). Since low energy states are more likely than states corresponding to higher energies, minima of the PES are related to conformations that are (locally) preferential. In non-isolated systems, these geometries and, in particular, global energy minima are associated with substantially higher statistical weights and, therefore, particularly recommended as initial structures for MD simulations. Moreover, a molecular system assembled from single building blocks (macromolecule, ligand, solvent molecules, etc.) will most likely result in an unphysical state that should to be relaxed beforehand. In all such situations, it is always advisable to perform an energy minimization procedure on the system under observation before starting the sampling process. Saddle points, in contrast, blaze preferential reaction paths flagging a conformational change from one favorable state (minimum) to another, since passing energetic barriers via saddle points requires least energy uptake. They are therefore particularly useful for investigations of molecular kinetics and possible reaction paths. From a physical point of view, molecular conformers corresponding to critical points q in configurational space are characterized by vanishing internal forces, $F = 0$, and consequently referred to as *stationary points*. As already pointed out earlier, F is calculated as the potential energy gradient

$$F = -\nabla_q U(q) \tag{3.19}$$

that is the force vector comprising all partial derivatives of U with respect to (generalized) coordinates q. At a critical point, the slope of $U(q)$ equals zero with respect to any degree of freedom of the corresponding geometry (in direction of every spatial dimension of each atom if q is a vector of Cartesian coordinates). Since both minima as well

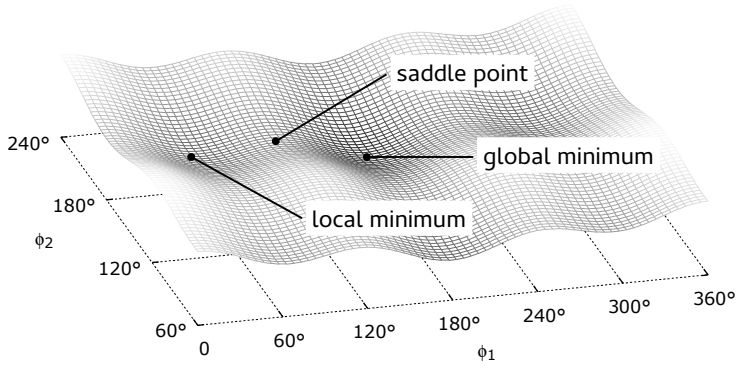


Figure 3.8: Critical points of the potential energy surface of pentane spanned by its two all-carbon dihedral angles.

as maxima are characterized by this property, a critical point in some high-dimensional space can be classified as a minimum (in each dimension), maximum (in each dimension), or saddle point (combination of minima and maxima, each associated with at least one dimension). A categorization of some critical point $q \in \mathbb{R}^{3N}$ is possible using the *Hessian matrix*[154]

$$H_U(q) = \left(\frac{\partial^2 U}{\partial q_i \, \partial q_j}(q) \right)_{i,j \in \{1,3N\}}$$

of all second-order partial derivatives of the potential energy function with respect to q. It makes a statement about a function's curvature at position q in configurational space. According to the second-order criterion, a minimum q is associated with a *positive definite* Hessian $H_U(q)$ which is characterized by positive eigenvalues only and

$$q^\top H_U(q)\, q > 0,$$

whereas maxima result in a *negative definite* matrix with all-negative eigenvalues. In contrast, the Hessian at a saddle point q characterized by positive as well as negative eigenvalues is denoted as *indefinite* and yields[154]

$$q^\top H_U(q)\, q = 0.$$

Technically, the energy minimization of a molecular geometry is equivalent with directing the set of coordinates to a (most likely local) minimum. From a mathematical point of view, the energy minimization is a matter of non-linear optimization since a great many quadratic, trigonometric, and (higher-order) hyperbolic terms of the coordinates q are involved in the potential energy function $U(q)$ as illustrated by Equation 3.18. Unfortunately, no general analytical solution is known to the determination of a minimum in particular for non-linear systems with multiple degrees of freedom. Instead, several numerical methods are established which approach a multivariate minimum iteratively as illustrated by Figure 3.9 for the one-dimensional case. During the iterative process, a given start conformer $q_0 \in \mathbb{R}^{3N}$ of a molecular system consisting of N atoms is relaxed by directing its coordinates (positions) q_k gradually towards the next local minimum. Respective methods range from those that evaluate only the potential function to those consulting first-order (gradients) or even substantially more expensive second-order derivatives (Hessians). Typical questions arising upon the choice of a suitable algorithm address *convergence* properties and computational complexity. Both the bisection and downhill simplex method[155] for instance belong to the former class of algorithms that is characterized by minimal computational effort per iteration. However, missing information about function derivatives usually protracts convergence towards a

minimum by significantly increasing the number of necessary iterations.[154,156] At the other end of the spectrum we find computationally challenging procedures like Newton's method that require the determination of expensive second-order derivatives at each iteration step.[154] Gradient-based minimization routines (GBMRs) constitute a reasonable tradeoff regarding convergence and complexity. Indeed, most well-known classical MD simulation packages (NAMD, GROMACS, AMBER, etc.) by default provide tools for the purpose of energy minimizations based on gradients due to their eminent suitability. Since, in addition, any minimization process performed throughout this thesis are of this type, we will elaborate on their functionality in little more detail. Basically, each iteration k of all GBMRs which belong to the class of *line search* methods consists of two major steps: the calculation of a search direction $d_k \in \mathbb{R}^{3N}$ in configurational space followed by some increment Δq_k of positions q_k. Hence, what all GBMRs have in common, in principle, is the calculation of first-order partial derivatives during phase one and a subsequent line search. The algorithms mainly differ in how they determine the values of d_k and Δq_k.[154]

Steepest descent

According to the *steepest descent* algorithm also referred to as *gradient descent*, the search direction

$$d_{k+1} = -g_k = -\nabla_{q_k} U\left(q_k\right)$$

of the next iteration step $k+1$ is defined as the negative gradient $-g_k$ of the energy function U with respect to the current position q_k. As a consequence, the algorithm having good sense to do so follows the direction along which U decreases most rapidly. Accordingly, the next geometry of the molecular system is calculated through

$$q_{k+1} = q_k + \gamma_k d_k. \tag{3.20}$$

where the step size γ may be determined, for instance, by a simple one-dimensional line search in direction d_k through q_k. The steepest descent algorithm performs rapidly in steep regions far away from the next local minimum. However, depending on the optimized function, it often converges very slowly due to its zig-zag course in the proximity of minima.[154] An alternative method devoid of this convergence issue is described in the following.

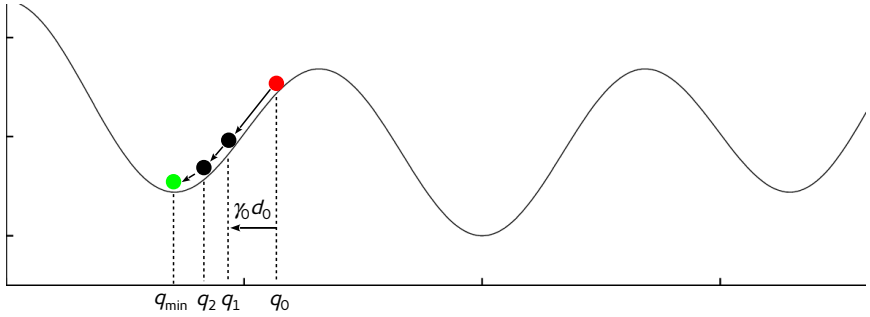


Figure 3.9: Iterative potential energy minimization routine starting with the green colored initial conformation q_0 and ending up with the red colored next local minimum q_{min}.

Conjugate gradients

Probably the most popular energy minimization algorithm in the field of MD simulations is known as the *conjugate gradients* (CG) method.[157] In contrast to steepest descent, the determination of the next search direction

$$d_{k+1} = -g_k + \beta_{k+1} d_k$$

includes not only the current gradient g_k but as well the preceding search direction d_k scaled by a factor β_{k+1}. One well-known strategy of several convenient for a determination of β_{k+1} was presented by Fletcher and Reeves in the early sixties of the past century[158] and uses the quotient of norms

$$\beta_{k+1} = \frac{d_{k+1}^\top \, d_{k+1}}{d_k^\top \, d_k}.$$

of the two previous directions. In analogy to the steepest descent method, a new step size γ_k can be calculated using a one-dimensional line search based on a golden section or Fibonacci approach.[159] Afterwards, new coordinates q_{k+1} are computed according to Equation 3.20. The CG algorithm is known to converge significantly faster to the minimum in its vicinity than steepest descent. It often seems reasonable to combine the two methods depending such that steepest descent is used at the beginning of an minimization procedure and CG for fine tuning.[154]

Resilient backpropagation

Due to nested iterations, i. e. a one-dimensional line search procedure upon each iteration step, the time complexity of the previously described GBMRs becomes very high in particular regarding macromolecular systems. In contrast to CG, an adaptive learning algorithm denoted as *resilient backpropagation* (rProp) as proposed by Riedmiller

and Braun by the end of the last century uses only the *sign* of a potential's gradient. In addition, it smoothens the potential energy surface resulting in lower minima.[160] First, the next step size $\gamma_{k+1} \in \mathbb{R}^{3n}$

$$\gamma_{k+1} = \begin{cases} \min\left(\gamma_k \cdot \eta^+, \gamma_{\max}\right) & \text{if } g_k\, g_{k-1} > 0, \\ \max\left(\gamma_k \cdot \eta^-, \gamma_{\min}\right) & \text{if } g_k\, g_{k-1} < 0, \\ \gamma_k & \text{else.} \end{cases}$$

is calculated where $\gamma_{\max}/\gamma_{\min}$ and η^+/η^- are denoted as maximal/minimal step size and, respectively, increasing/decreasing factor. Obviously and in contrast to the methods described previously, no iterative procedure is applied to the calculation of γ_{k+1} further increasing the velocity of rProp. New coordinates are determined by

$$q_{k+1} = q_k - \gamma_k \operatorname{sgn}\left(g_k\right).$$

Any iterative numerical procedure requires a convergence test in order to terminate when no further improvement of the minimization can be expected. Common convergence criteria are the difference in energies or coordinates, and the gradient's norm. If no significant change in one or several of these conditions is met, the process is terminated.[154]

Global minimization

The methods presented so far search for local minima. Since we were interested in global minimum conformations of ligand molecules as initial structures, global optimization strategies as well are briefly sketched here. As already indicated by the term *curse of dimensionality* in the introduction chapter, finding the global minimum of macromolecular systems is a highly complex and probably impossible task due to the vast amount of local minima in the exceedingly rough high-dimensional PES. Nevertheless, few strategies such as *simulated annealing* and *genetic algorithms* have been developed addressing this question. In principle, these methods perform in a similar manner as they mostly generate a certain number of new conformations that serve as starting points for independent local minimization routines. Figure 3.10 illustrates this concept. The simulated annealing algorithm, for instance, periodically increases and decreases the temperature resulting in temporarily high momenta that might be able to nudge the molecular system across energetic barriers and direct it to new local minima possibly including the global minimum.[161] Genetic algorithms, in contrast, produce new conformations by

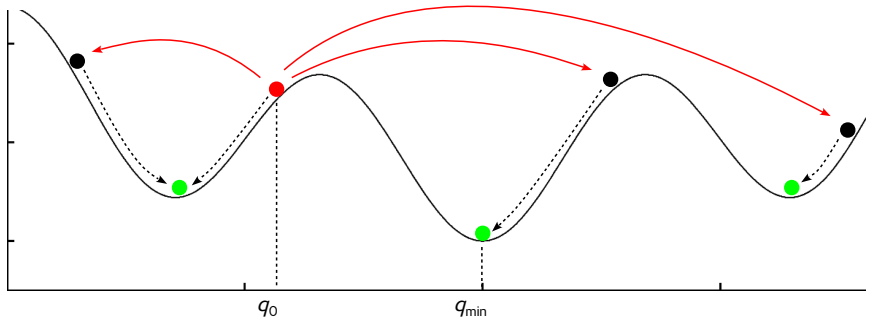


Figure 3.10: Iterative potential energy minimization routine starting from an initial conformation q_0 and ending up at several minima possibly including the global minimum q_{min}.

combining coordinates from different geometries in order to approach unexplored regions from which new local minima of the configurational space can be reached.[162] *Monte-Carlo sampling* methods (MC, described below) as well are commonly used for the localization of large numbers of minima.[163,164] However, due to its computational complexity, even with supercomputers it is rather impossible to assess every minimum of a PES associated with macromolecules within a reasonable time period. As a consequence, no reliable and efficient approach to the identification of global minima is known currently. Nevertheless, Section 4.1 describes and evaluates a simple approach to global optimization of small molecules on the basis of MC samplings at artificially high-temperatures.

3.5 Numerical integration of equations of motion

Classical MD simulations mimic atomic motions of a molecular system using Newton mechanics. In the classical approximation of quantum mechanics, these particles constitute atoms that underlie mutual interactions based on various types of forces acting on each other and, thereby, resulting in an acceleration and effective momenta associated with these atoms. As pointed out in Section 3.2, such interactions are commonly parameterized using classical force fields. The whole MD procedure ends up in sampling a deterministic sequence of time-discrete snapshots (time steps) of a molecular system along with corresponding potential and kinetic energies as well as other microscopic thermodynamic quantities. Many average structural and thermodynamic properties can be determined on the basis of these phase space trajectories which are constructed by an iterative calculation of atomic positions and momenta as illustrated through Figure 3.11. The high dimensionality of the underlying molecular system usually consisting of far more than two particles keeps off an analytical solution to the equations of mo-

tion which is why numerical methods need to be applied. In general, such *numerical integrators* include information about the equation's derivatives. Some integrators incorporate expensive second-order derivatives (Hessian) whereas others are content with the gradient only. Since, in principle, every popular MD simulation software for the solution of Hamiltonian systems incorporates the latter type, we will shortly elaborate two integration schemes.

Convenient numerical integrators are expected to satisfy a couple of important criteria. On the one hand they should be as stable as possible in terms of the accuracy of phase space trajectories. On the other they are expected to be, apart form conserving the (angular) momentum, symplectic[122] and time-reversible in order to meet the thermodynamic requirement of constant total energy in isolated systems (see Section 3.1) and Newton's laws (3.1), respectively. The symplecticity property is particularly required for the solution of Hamilton's equations of motion (Equation 3.10).[165] Starting with a Taylor expansion of the solution to Newton's equations (Equation 3.1) a couple of numerical integration schemes have been developed meeting the aforementioned criteria and incorporating atomic forces as first-order derivatives of the potential energy function (Equation 3.19). This is achieved by a truncation of the Taylor series after the quadratic term that is second-order partial derivatives of positions which are related to forces.[166,167]

Velocity Verlet integrator

A popular class of numerical integrators of equations of motion denoted as *Verlet integrators* is characterized by the sum of forward and backward propagation of the underlying

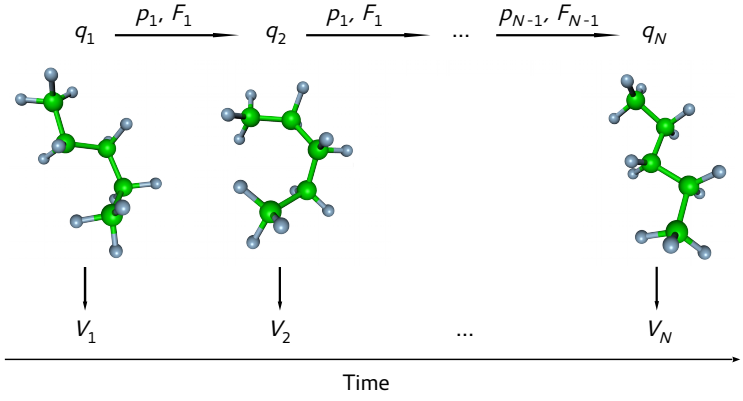


Figure 3.11: Construction of a molecular dynamics time series illustrated using pentane.

Taylor series.[121] The *velocity Verlet* algorithm

$$q(t+\Delta t) = q(t) + \Delta t M^{-1} p(t) + \frac{\Delta t^2}{2} M^{-1} \nabla_q V(q(t))$$
$$p(t+\Delta t) = p(t) + \frac{\Delta t}{2} V(\nabla_q V(q(t)) + \nabla_q V(q(t+\Delta t))).$$

is probably the most prominent integrator in common MD simulation software. It meets the requirement of time-reversibility and is characterized by an error in the order $\mathcal{O}(\Delta t^3)$ as well as a particularly low rounding error. As signified by Figure 1.4 in the introduction chapter, the discretized time step Δt is commonly set to about 10^{-15} s in order to capture the fastest molecular oscillations that are related to covalent bonds comprising hydrogen atoms. By applying bond constraints (described below) Δt can be set to larger values resulting in longer trajectories at the same number of energy/force evaluations. In the light of both the proportionality of velocity and momentum, $v(t) = M^{-1} p(t)$, and the equality of forces and potential gradients shown by Equation 3.19, there is an obvious analogy to Newton's solution to a conservative and constantly accelerated system (Equation 3.4). As illustrated in Figure 3.11, the iterative velocity Verlet algorithm typically provides a deterministic time series of pairwise coordinates and momenta, $(q(t), p(t))$ accompanied by potential energies.

Leap frog integrator

Another commonly used integrator coming into question for MD simulations is the *leap frog* algorithm

$$\dot{q}(t+\Delta t/2) = \dot{q}(t) + \frac{\Delta t}{2} M^{-1} \nabla_q V(q(t)) \tag{3.21}$$
$$q(t+\Delta t) = q(t) + \Delta t \dot{q}(t+\Delta t/2) \tag{3.22}$$
$$\dot{q}(t+\Delta t) = \dot{q}(t+\Delta t/2) + \frac{\Delta t}{2} M^{-1} \nabla_q V(q(t+\Delta t)) \tag{3.23}$$

which is in theory equivalent to velocity Verlet. Both are characterized by the same properties regarding time-reversibility, rounding error and the order of the error. However, in contrast to velocity Verlet, numerous implementations of the leap frog integrator do not provide positions and velocities at the same time but with an offset of a half time step.[15,120] However, one can apply the midpoint rule if they are desired at the time.

In their original formulation both integrators produce a microcanonical ensemble with constant total energy which is not convenient for biochemical systems as investigated in the framework of this thesis.[23,24] Thus, in order to yield other statistical ensembles able to exchange heat (and work), modifications to the integration scheme are necessary which we will discuss in the following section.

3.6 Temperature and pressure coupling

Ordinary MD calculations on the basis of a symplectic integration scheme yield a microcanonical NVE ensemble consisting of uniformly distributed microstates with constant energy E besides constant number of particles N and constant volume V. However, these thermodynamic boundaries hardly satisfy realistic, natural systems that are widely characterized by constant temperature T rather than constant energy E owing to the exchange of heat with the surrounding. Moreover, if we want to model systems under atmospheric conditions, we additionally need to switch to constant pressure p in lieu of the volume. Most convenient for such ubiquitous systems is the canonical (NVT) or isothermal–isobaric (NpT) statistical ensemble.[53,168] In order to meet these physical conditions, mathematical modifications to Newton's equation of motion and their solutions (integrators) as well as dynamics alternative to the Hamiltonian are conceivable. Stochastic as well as deterministic methods have been developed for that purpose. Algorithms designed for coupling the system temperature to a heat bath with some particular value T_0 are known as *thermostats*. Following Equation 2.18 derived from the equipartition theorem, the *instantaneous temperature*

$$T(t) = \frac{1}{k_\mathrm{B} N_\mathrm{f}} \sum_{i=1}^{3N} m_i v_i^2(t)$$

of an N-particle system with $N_\mathrm{f} = 3N - N_\mathrm{c} - N_\mathrm{r}$ internal degrees of freedom at time t can be calculated from the velocities v_i and the masses m_i (according to their matrix representation in Equation 3.9) associated with the $i^{\,\mathrm{th}}$ degree of freedom.[24] The degree of freedom is reduced by geometrical constraints N_c and depends on the system's boundary conditions via N_r which is set to $N_\mathrm{r} = 3$ or $N_\mathrm{r} = 6$ in case of periodic or vacuum boundary conditions, respectively. In the presence of stochastic and frictional forces N_r equals zero.[53]

First attempts to the constant temperature and/or constant pressure sampling were made by Andersen in 1980.[168] He proposed a stochastic randomization through randomly choosing a particle during simulation and generating its velocity from the Maxwell distribution (Equation 2.19). Andersen's method based on Hamiltonian equations of motion in combination with stochastic collisions generates a Markov chain of microstates in phase space and requires the system to be ergodic.[53] Probably due to the occurence of unphysical discontinuities in the trajectories, a poor efficiency and the lack of conserved quantity to be relied on, this thermostat did not become popular in a sense that it is less frequently implemented in current state-of-the-art simulation software. However, it must be considered as a forerunner of thermostat algorithms sketched

in the following.[169]

Berendsen thermostat

In the year 1984, Berendsen presented a popular thermostat denoted as *weak-coupling* algorithm that is widely integrated in common simulation software.[170] Essentially, the deviation of system temperature T from reference temperature T_0 decays exponentially according to the relation

$$\frac{\mathrm{d}T}{\mathrm{d}t} = \frac{T_0 - T}{\tau}$$

where the time or coupling constant τ steers the coupling strength. Starting with the addition of a friction and a stochastic term to Newton's formula (Equation 3.1) yielding a *Langevin equation*, the approach finally ends up with a modified equation of motion

$$m_i \dot{v}_i = F_i + m_i \gamma \left(\frac{T_0}{T} - 1 \right) v_i.$$

Here, the strength of the heat coupling is determined through the damping (or friction) constant γ chosen equal for each degree of freedom. In practice, velocities are scaled according to a time-dependent scaling factor

$$\lambda = \left[1 + \frac{\Delta t}{\tau_T} \left(\frac{T_0}{T} - 1 \right) \right]^{1/2}$$

describing the heat flow into or out of the system. A rigorous sampling from the canonical ensemble is prevented by the method's suppressive influence on fluctuations of the kinetic energy (as expressed by Equation 2.17) which are associated with constant temperature ensembles. However, for large systems (large N) the error becomes negligible since it scales with $1/N$.[171] Furthermore, Berendsen's method that is neither classified as stochastic nor as time-reversible quickly reaches thermal equilibration due to an exponential relaxation.[53,169]

Velocity rescaling thermostat

The *velocity rescaling* scheme developed by Bussi et al. in the past decade mainly operates like Berendsen's thermostat, though, it is able to generate a correct canonical ensemble. This is achieved through the presence of a *Wiener* (or *Brownian*) process $\mathrm{d}W$ as an additional stochastic term which introduces a fluctuation on the kinetic energy K[169]

$$\mathrm{d}K = (K_0 - K) \frac{\mathrm{d}t}{\tau_T} + 2\sqrt{\frac{K K_0}{N_\mathrm{f}}} \frac{\mathrm{d}W}{\sqrt{\tau_T}}.$$

The instantaneous kinetic energy K_0 is defined by multiplying Equation 2.18 with the number N_f of degrees of freedom

$$K_0 = \frac{N_\mathrm{f}}{2} k_\mathrm{B} T.$$

A validation of the NVT sampling as well as a reasonable choice of the integration time step is possible on the basis of a conserved quantity denoted as *effective energy* $\tilde{H}$ and corresponding to the heat flux between the system and its surrounding heat bath

$$\tilde{H}(t) = \int_0^t (K_0 - K(t')) \frac{\mathrm{d}t'}{\tau} - 2 \int_0^t \sqrt{\frac{K(t')\, K_0}{N_\mathrm{f}}} \frac{\mathrm{d}W(t')}{\sqrt{\tau}}.$$

Nevertheless, the velocity rescaling method according to Bussi et al. is classified as deterministic and does not cope with the physical requirement of time-reversibility.[53,169] According to the developers, their method quickly reaches equilibrium just like Berendsen's approach. Due to a first-order temperature decay, no oscillations are observed. However, for small systems or when the observables of interest are dependent on the fluctuations rather than on averages, this method cannot be used.[169] Almost all constant temperature MD simulations accomplished within this thesis employ Bussi's velocity rescaling method due to its generation of the correct canonical ensemble and its compatibility with several barostats (described below).

Other broadly used thermostats that produce the correct canonical ensemble particularly include the *extended ensemble Nosé–Hoover* scheme and *Langevin dynamics*. Regarding the former method, the capability of the correct canonical ensemble is due to the thermostat's deterministic nature and the lack of any stochastic influences. The term *extended ensemble* reposes on two additional degrees of freedom associated with the position and conjugate momentum of an imaginary heat reservoir where the velocity update includes an additional force proportional to the velocity and commonly referred to as *friction term*.[172] Probably due to its theoretical agreement with several physical properties (time-reversible, deterministic, correct canonical ensemble), the thermostat developed by Nosé and Hoover is widely used although it is not guaranteed to be ergodic and only slowly reaches the desired temperature due to an oscillatory relaxation.[53,169] However, further modifications of the equations of motion in association with multiple heat baths referred to as *Nosé–Hoover chain* have been proposed in order to remedy this problem.[173] In contrast, the Langevin dynamics also referred to as *stochastic dynamics* approach was shown to be ergodic.[53] The Langevin approach is characterized by an extension of Newton's equation of motion about two additional terms accounting for Brownian motion (steered stochastically by a white noise term) and for friction (drag

due to collisions with particles), respectively. Besides, these random collisions implicitly mimic solvent effects what makes this method suitable for implicit solvent simulations. According to the fluctuation-dissipation theorem, the friction coefficient depending on the solute's geometric and the solvent's physical properties such as viscosity is related to the strength or variance of random fluctuations.[53,174]

Berendsen barostat

Regarding pressure coupling of our biological host–guest systems, mainly two methods came into question constructed similarly to thermostats and sometimes considered as their mathematical counterparts. A popular example is Berendsen's weak-coupling scheme mentioned earlier in the context of temperature coupling.[170] Transferred to pressure coupling, atomic coordinates and box vectors are scaled every (N_{PC}) step(s) according to a first-order kinetic relaxation of the pressure

$$\frac{\mathrm{d}P}{\mathrm{d}t} = \frac{P_0 - P}{\tau_{\mathrm{P}}}$$

towards the reference temperature P_0 with time constant τ_{P}. An instantaneous pressure P is obtained as the sum of the kinetic energy and internal virial for pair-additive potentials

$$P = \frac{1}{3V}\left[\sum_i \frac{p_i^2}{m_i} + \sum_{i<j} r_{ij}F_{ij}\right]$$

where r_{ij} and F_{ij} denote the distance between particles i and j and, respectively, the force exerted on particle i by particle j. Box vectors and coordinates and are then adapted according to the scaling factor

$$\mu = \left[1 - \frac{\beta_{\mathrm{IC}}\Delta t}{3\tau_{\mathrm{P}}}\left(P_0 - P\right)\right]^{1/3}$$

making use of the isothermal compressibility β_{IC} which becomes a tensor in case of anisotropic triclinic systems.[170] The authors stated that the algorithm is most easily incorporated into the leapfrog integrator. Reportedly,[171] Berendsen's fast pressure control algorithm does not produce the exact NpT ensemble.

Parrinello–Rahman barostat

The extended ensemble Parrinello-Rahman approach to constant pressure simulations[175] exhibits technical similarities to the Nosé–Hoover extended system thermostat. Here,

the system volume $V = \det h = a\,(b \times c)$ defined on the basis of three box vectors a, b, and c that constitute the column vectors of matrix $h = (a, b, c)$ is taken as a further degree of freedom. By way of a coordinate transformation according to h, the position

$$r_i = hs_i = \xi_i a + \eta_i b + \zeta_i c.$$

of any particle $i \in \mathbb{R}^3$ is expressed in terms of the volume vectors and an additional vector $s_i = (\xi_i, \eta_i, \zeta_i)^\top$. An appropriate Lagrangian serving as basis for the equations of motion for a system with hydrostatic pressure p

$$\mathcal{L} = \frac{1}{2}\sum_i m_i \dot{s}_i^\top G \dot{s}_i - \mathcal{V}_N\,(r) + \frac{1}{2} W \mathrm{Tr}\left(\dot{h}_i^\top \dot{h}_i\right) - pV$$

is constructed using tensor $G = h^\top h$, potential $\mathcal{V}_N\,(r)$ and a stochastic term W steering the relaxation time.[175] The barostat developed by Parrinello and Rahman allows to change both size and shape of the box. It is useful whenever fluctuations in pressure or volume are important and gives the true NpT ensemble.[171]

Practical considerations

It is strongly recommended to subject any complex molecular system assembled for simulation purposes to an energy minimization followed by one or more successive equilibration procedures. Constant temperature systems NVT should initially undergo an equilibration of particle velocities at constant volume in order to eliminate excess heat and obtain homogeneous stress. If constant pressure is required as well (NpT ensemble), a proper barostat algorithm is turned on afterwards. For rigorously correct constant pressure ensembles, the velocity Verlet integrator is recommended.[171]

Depending on the purpose of an MD simulation, one or the other barostat and/or thermostat suits better. For systems far from equilibrium Berendsen's weak coupling thermostat and barostat algorithms are recommended exhibiting a fast and smooth approach to equilibrium. However, they turn out to be less reliable at equilibrium. As a consequence, for production runs targeting the prediction of thermodynamic properties of systems close to equilibrium the Parrinello-Rahman barostat in combination with the Nosé–Hoover or velocity rescaling thermostat is recommended since they, at least in theory, produce the correct ensemble.[171] Regarding our own MD simulations, we followed these recommendations and utilized Berendsen's thermostat/barostat for equilibration and the feasible combination of Parrinello-Rahman and velocity rescaling for the production of statistical averages.

3.7 Boundary conditions and geometric constraints

Prior to an MD run of some molecular system, its shape and the nature of its surroundings/boundaries must be properly defined. Depending on physical requirements, one can choose between vacuum, rigid, or *periodic boundary conditions*.[53] While the vacuum condition is not suitable for bulk biochemical systems surrounded by explicit solvent molecules and the rigid type mostly introduces artefacts due to nonphysical collisions, the periodic case enjoys great popularity in the biopolymeric MD community[24,53] as well as in the context of our calculations justifying its description in little more detail.

Cell unit and periodic boundaries

As indicated by Figure 3.12 for a quadratic simulation box in two dimensions, this type of boundary condition is specified by a space-filling regular arrangement of an infinite number of clones of a single unit cell containing the molecular system under investigation. Consequently, any particle i with position r_i of a simulation box defined by the cell vectors a, b, and c and cell lengths $L = (|a|, |b|, |c|)$ is copied into each cell clone $k = (\alpha, \beta, \gamma)^\top$ using a translation

$$r_i' = r_i + \alpha a + \beta b + \gamma c = r_i + \sum_\zeta k_\zeta L_\zeta.$$

The integer coefficients α, β, and γ specify the image cell's displacement in direction of respective box vectors. The major advantage associated with this approach is that system particles at the boundaries interact with "real" particles rather than with a wall yielding to artefacts. According to the *minimum image convention*, interactions of each particle i represented by the red sphere in the Figure's central unit are restricted only to the nearest image (black spheres) of any other particle j. That is, the nearest image of any other atom j exerting force F_i on atom i is determined as

$$k_{\min} = (\alpha, \beta, \gamma)_{\min}^\top = \underset{k}{\operatorname{argmin}} \left| r_i - \left(r_j + \sum_\zeta k_\zeta L_\zeta \right) \right|.$$

The set of $N - 1$ nearest images are encompassed by the dotted line which has the same shape and volume like the central cell unit, though, it is centered on atom j. Furthermore, if a particle leaves the central cell towards one side, it instantly reenters at the opposite side as illustrated by red arrows.

In practice, each vector of a simulation box is required to exceed $2R_c$ which is the (maximal) cutoff distance for nonbonded potentials in order to avoid the interaction

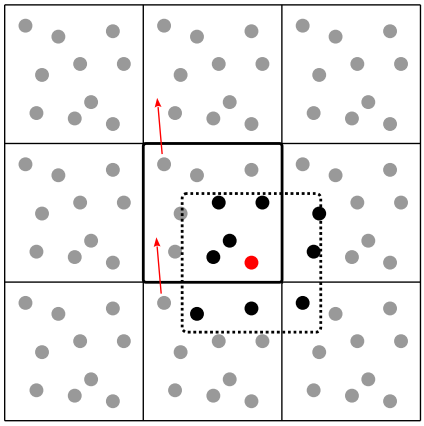


Figure 3.12: Periodic boundary condition: Central cell unit of a simulation box surrounded by its own clones. Particles leaving the a cell instantly reenter on the opposite side (red arcs). **Minimum image convention**: Only the nearest images (black spheres encompassed by dotted line) of any particle j are considered for interactions with particle i (red sphere) upon MD.

of a particle with its own image or more than one copy of another. Several shape types are known and often used for cell units that are compatible with periodic boundary conditions in bulk MD simulations. Apart from the cubic cell used in Figure 3.12, further special cases of triclinic such as truncated octahedral or rhombic dodecahedral cells are available in diverse MD software packages including GROMACS,[171] which was applied to all bulk simulations within the presented work. The latter two cells are closer to a spheric shape and therefore particularly suitable for approximately spheric molecules including all target molecules investigated throughout this thesis. In addition, these two cell types significantly save CPU-time (about 20-30 %) compared to a cube that would be required for the same spheric solute.[171]

Geometric constraint algorithms

During an MD simulation, one often wants to keep certain geometric properties like bonds and/or angles at some fixed value. This can be achieved through geometric *constraint algorithms*. The discretized time step Δt is usually chosen small enough, about $\Delta t = 1$ fs in classical systems, to yield sufficient resolution of the fastest oscillations which typically correspond to covalent bonds. Constraining all bond length to their reference values allows to increase the step size up to approximately 2 fs which in turn is required for resolving angle bending. Ensuing from a twice as large time step, the conformational space is sampled with double speed. This procedure is commonly accepted since the bond (and bond angle) contribution to energy as well as molecular conformation is small. Accordingly, further acceleration is achieved if constraints are applied

to bond angles as well.[24] Of course, one should be aware of the slight distortions of system dynamics that are introduced by constraints and judge whether the oscillations in question are necessary for the intended investigation or negligible. Most MD software use constrained algorithms that are based on Lagrange multipliers[176] exemplarily sketched in case of the SHAKE constraint algorithm. Within the scope of this thesis, bond constraints have been applied to all explicit water simulations.

SHAKE The SHAKE algorithm performs constrained MD using a Lagrangian $\mathcal{L}(q, \dot{q})$ extended by r additional constraints according to

$$\mathcal{L}^*(q, \dot{q}) = \mathcal{L}(q, \dot{q}) + \sum_{i=1}^{r} \lambda_i \sigma_i(q).$$

Each constraint i of the type $\sigma_i(q) = \xi_i(q) - \xi_i$ is associated with the Lagrange multiplier λ_i and required to become zero with ξ_i specifying the value to which $\xi_i(q)$ is fixed. Considering the leapfrog algorithm, Lagrange multipliers are determined iteratively right after the position update in Equation 3.22 of each numerical MD step until all constraints $|\sigma_i(q)|$ become smaller than some predefined tolerance.[177] The vector of displacements

$$\delta_{j+1} = \delta_j - J^{-1}\sigma_j$$

of step j is then calculated as the difference between the previous displacement vector and a second term on the right incorporating the inverse Jacobi matrix J. Due to the high computational demand of that quantity, SHAKE seems less suitable for large macromolecular systems. Moreover, it is inherently serial and cannot be parallelized.[176] Several extensions of the SHAKE algorithm have been developed including a non-iterative version[178] and RATTLE.[179] RATTLE was presented by Andersen as an alternative compatible with the velocity Verlet integrator which requires an additional constraining step.[180]

SETTLE The SETTLE constraint algorithm solves a system of non-linear equations analytically for three constraints and can only be used for one type of molecule. It is, therefore, particularly suitable for rigid water[176,181] and indeed applied to water molecules by the Gromacs simulation suite when SHAKE was chosen as constraint algorithm by the user.[171] The SETTLE algorithm is known to be fast and extremely stable.[176]

LINCS The linear constraint solver LINCS is another popular algorithm often used in MD simulations and published in 1997 by Hess and co-workers.[182] In contrast to

the iterative approach implemented in SHAKE, it performs in exactly two steps[171] and, in particular, quickly approximates the Jacobian J as a power series instead of computing and inverting it.[176] After an unconstrained update, the projection of new bonds to the old ones is set to zero followed by a correction of bond lengths. LINCS is much faster than SHAKE, though, it is advisable not to use it with bond *and* angle constraints if more than terminal angles including atoms with one binding partner only (hydrogens, halogens, keto oxygens, etc.) are considered. Due to its superiority over SHAKE regarding stability and efficiency, LINCS is set as default in GROMACS[171] and was used for all bond constraints in this thesis. Consequently, the time step size was increased to 2 fs for any MD run.

3.8 Markov chain Monte Carlo sampling

Monte Carlo (MC) methods designed for the calculation of high-dimensional integrals provide a completely different approach to the sampling of molecular geometries.[18,183] In contrast to deterministic MD simulations based on the numerical integration of Newton's equations of motion, MC methods perform some kind of *random walk* through configurational or phase space. In the application area of molecular simulations the terminology *random walk* emerges from the consideration of states $q^{(k)}$ as a random variable X sampled from some convenient probability distribution $P(X)$. In statistics a characteristic known as *Markov property*

$$P\left(X_k = q^{(k)} \mid X_{k-1} = q^{(k-1)}, \cdots, X_0 = q^{(0)}\right) = P\left(X_k = q^{(k)} \mid X_{k-1} = q^{(k-1)}\right)$$

states that the conditional probability distribution $P\left(X_k = q^{(k)}\right)$ of a new state $q^{(k)}$ depends (in case of a first order process) only on the immediately previous state $q^{(k-1)}$. The Markov property is a necessary condition for a stochastic process generating a *Markov Chain* of states. It is characterized by *transition probabilities* $P(q \to \tilde{q})$ from one state q to another $\tilde{q}$ and yields a non-deterministic time series $\left(q^{(1)} \to \cdots \to q^{(N)}\right)$ consisting of N states. This class of methods that is, consequently, referred to as *Markov Chain Monte Carlo* (MCMC) constructs a thermodynamic ensemble at thermal equilibrium whose distribution converges against the Markov chain's *stationary distribution* $\pi(q)$ with increasing N.[19,184] In order to meet the physical requirement of reversibility given at equilibrium and thereby show the uniqueness of the stationary distribution, an MCMC method must satisfy the *detailed balance* criterion

$$\pi(q)\, P(q \to \tilde{q}) = \pi(\tilde{q})\, P(\tilde{q} \to q) \tag{3.24}$$

Equation 3.24 says that, given state q, the conditional probability $P(\tilde{q} \mid q)$ of the transition from q to state $\tilde{q}$ must equal the conditional probability $P(q \mid \tilde{q})$ associated with the reverse transition.[185]

Metropolis–Hastings algorithm

Within the scope of molecular sampling, most MCMC algorithms perform two major steps. A new state is first proposed according to the underlying probability distribution and, afterwards, either accepted or rejected according to some acceptance probability. The Metropolis–Hastings algorithm from the class of MCMC methods was published in the middle of the 19th century.[19] Reasonably, its developers chose a probability distribution

$$\pi(q) = \frac{1}{Z_q} \exp(-\beta U(q)) \tag{3.25}$$

in accordance with the Boltzmann distribution of potential energies $U(q)$ derived from Equation 2.20. Considering the transition probability

$$P(q \to \tilde{q}) = P_{\mathrm{P}}(q \to \tilde{q})\ P_{\mathrm{A}}(q \to \tilde{q}) \tag{3.26}$$

as the product of a *proposal* (P_{P}) and an *acceptance probability* (P_{A}) and inserting Equations 3.25 and 3.26 into Equation 3.24 where the partition function Z_q cancels out yields the acceptance probability

$$P_{\mathrm{A}}(q \to \tilde{q}) = \min(1, \exp(-\beta \Delta U)) = \begin{cases} \exp(-\beta \Delta U) & \text{if } U(q) < U(\tilde{q}), \\ 1 & \text{else.} \end{cases} \tag{3.27}$$

To be more precise, if a uniformly distributed random value $r \in [0, 1]$ is less then $P_{\mathrm{A}}(q \to \tilde{q})$, one keeps the old state q and otherwise accepts the proposed state $\tilde{q}$. From Equation 3.27 it becomes obvious that new states with less energy than the current one are always accepted whereas the acceptance of higher energy states depends on chance.

In contrast to MD simulations that are often faced with trapping effects, MCMC methods are able to alter energy isosurfaces due to their stochastic character and overcome energetic barriers more easily. Their main disadvantage is the small step size in conformational space as well as the loss of determinism. In addition, it often suffers from low acceptance rates, in particular, in case of large and explicitly solvated systems with large energy fluctuations. In the following, we will therefore engage ourselves with a method combining the advantages of both deterministic as well as stochastic techniques for molecular sampling.

Hybrid Monte Carlo algorithm

Another popular MCMC approach introduced several decades after the Metropolis–Hastings algorithm is the *Hybrid Monte Carlo* (HMC) method.[116,183] It combines large MD steps in phase space with the stochastic character of MC methods which are known for an efficient compensation of the trapping effect associated with MD techniques.[23,183] Basically, each iteration of the MC procedure is accompanied by a short MD simulation leading to state proposals that are physically more reasonable, thereby, increasing their acceptance rate. In contrast to the original Metropolis–Hastings algorithm, the hybrid approach incorporates not only the potential energy as a function of coordinates but as well kinetic energies depending on velocities. Due to the resulting Hamiltonian's (Equ. 3.7) separability required by the HMC method, it is possible to decompose the Boltzmann distribution of states depicted by Equ. 2.20 into two factors

$$\pi\left(q,p\right)=\frac{1}{Z_q}\exp\left(-\beta U\left(q\right)\right)\ \frac{1}{Z_p}\exp\left(-\beta K\left(p\right)\right)=\pi_q\left(q\right)\ \pi_p\left(p\right)$$

representing the distribution of coordinates π_q and momenta π_p, respectively. Momenta sampled from the Boltzmann distribution $\pi_{\mathrm{P}}\left(p\right)$ of kinetic energies were shown to be a good choice for the proposal probability

$$P_{\mathrm{P}}\left(q\rightarrow\tilde{q}\right)=\pi_{\mathrm{P}}\left(p\right)=\exp\left(-\beta K\left(p\right)\right). \tag{3.28}$$

A short MD simulation starting with the old geometry and the proposed momentum yields a new point $(\tilde{q},\tilde{p})$ in phase space that is accepted with some probability. Considering the reversibility criterion of a molecular system at thermal equilibrium met by the Hamiltonian and given the equality $\pi_{\mathrm{P}}\left(p\right)=\pi_{\mathrm{P}}\left(-\tilde{p}\right)$ and inserting it into the detailed balance criterion depicted by Equation 3.24 yields the acceptance probability

$$P_{\mathrm{A}}\left(q\rightarrow\tilde{q}\right)=\min\left(1,\exp\left(-\beta\Delta\mathcal{H}\right)\right)=\begin{cases}\exp\left(-\beta\Delta\mathcal{H}\right) & \text{if } \mathcal{H}\left(q,p\right)<\mathcal{H}\left(\tilde{q},\tilde{p}\right),\\ 1 & \text{else.}\end{cases}$$

As already stated, another major advantage of the HMC method over an ordinary MCMC method is a significantly higher acceptance rate of about 80 % for small molecules which is due to velocities that are more physically distributed and therefore yielding an improved sampling of the conformational space. Properties of the Hamiltonian crucial for the use as an MCMC method are its reversibility yielding an invariant distribution, the energy conservation, as well as its symplecticity.[183] Setting the temperature to an extraordinarily high value yields extreme momenta which easily negotiate most energy barriers. As a consequence, the configurational space is explored more efficiently. Within the framework of this thesis, the HMC method was, due to these advantages,

applied to ligands under vacuum boundary conditions in order to obtain globally minimum energy conformers useful as initial structures for solvated MD simulations.

3.9 Third-party software and databases used in this thesis

All third-party software and databases used in this thesis are listed in Table 3.1.

Table 3.1: List of third-party software and databases used in this thesis.

Tool	purpose and source
ACPYPE	simplifies usage of the Ambertool Antechamber
	http://www.ccpn.ac.uk/v2-software/software/ACPYPE-folder
Ambertools v1.4 [186]	AMBER force field parameterization with Antechamber
	http://ambermd.org/AmberTools14-get.html
Amira [187]	inhouse molecular visualization tool
	https://amira.zib.de/
AutoDock-Vina v1.1.2 [98]	molecular docking
	http://vina.scripps.edu/
PDB [41]	protein structure database
	http://www.rcsb.org/pdb/home/home.do
Epos	Merck force field parameterization
	ZIB inhouse software
FADO [188]	molecular docking (Amira module)
	ZIB inhouse software
g_mmpbsa v1.1 [189]	free energy calculation according to MMPBSA method
	http://rashmikumari.github.io/g_mmpbsa/
Gromacs v4.0.7/v4.5.5 [190]	energy minimization and MD simulation
	http://www.gromacs.org/
MarvinBeans v5.5.0.1	2D drawing and 3D export of ligands
	https://www.chemaxon.com/download/marvin-suite/
Octave v3.2.4	numerical parameter estimation
	https://www.gnu.org/software/octave/
Openbabel [191]	cheminformatics toolbox
	http://openbabel.org/wiki/Main_Page
VMD v1.9.1 [192]	molecular visualization
	http://www.ks.uiuc.edu/Research/vmd/
ZIBGridFree	energy minimization and HMC simulation
	ZIB inhouse software

4 Development of systematic space discretization strategies

The molecular modeling techniques and results presented in this chapter have to a notable extent been published in the following articles. A republication of its content in the framework of this thesis was kindly permitted by the publishers:[193,194]

- M. Weber, R. Becker, V. Durmaz, R. Köppen: Classical hybrid Monte-Carlo simulation of the interconversion of hexabromocyclododecane stereoisomers. *Molecular Simulation*, 34(7):727–736, 2008.
- V. Durmaz, S. Schmidt, P. Sabri, C. Piechotta, M. Weber: A hands-off linear interaction energy approach to binding mode and affinity estimation of estrogens. *Journal of Chemical Information and Modeling*, 53(10):2681–2688, 2013.

Due to obstacles regarding MD time scale differences and, in particular, the trapping in basins of the highly complex conformational space, an exhaustive sampling of large macromolecular systems within reasonable time remains a difficult not to mention impossible task. Nevertheless, there exist strategies to an extensive conformational scanning especially of small molecules, and the space complexity issue can be tackled using space discretization strategies. Since it is in the interest of all concerned scientists to easily gain a quick overview over the conformational diversity of a molecular system and identify suitable representatives for further investigations including binding free energy calculations, we will elaborate in the following a couple of convenient algorithms. The first presented strategy is related to the identification of a single conformational representative associated with the highest statistical weight (global energy minimum) of a chemical compound which served as initial structure for host–guest binding analysis in upcoming chapters. The second part of this chapter describes the development of an algorithm designed for clustering conformational ensembles with intend to yield a minimal set of representatives per substance covering all regions of the conformational space that are physically accessible. Finally, a simple approach to ligand binding pose generation is presented based on an uniform decomposition of the space of relative host–guest orientations. For an as reliable as possible *in silico* estimation of

protein–ligand binding affinities, these aspects must be taken into account. Accordingly, the methods presented here constitute preliminary steps generating geometries that are used as structural basis for binding affinity models elaborated in subsequent chapters. Altogether, a fully automated pipeline is presented describing the estimation of binding affinities ensuing from a ligand geometry and the spatial position vector of an active site.

4.1 High-temperature HMC approach to global minima

Significant conformational changes during a typical MD or MC simulation are rare events. That is molecular conformations sometimes referred to as *metastable subsets* within a molecular conformational space are typically separated by high energetic barriers that are scarcely negotiated. Since within the framework of this thesis global minimum conformations are intended to serve as structural basis for subsequent binding affinity estimations, we are interested in an efficient and reliable way to their identification. Consider the dihedral angle spanned by two vicinal bromine atoms of the additive brominated flame retardant hexabromocyclododecane (HBCD) as illustrated by Figure 4.1. During an HMC sampling, the torsional angle represented by the black plot mainly resides in two small ranges around $\pm 70°$. Even using HMC rather than MD, the transition from one of these two metastable sets to the other occurs only once per several thousands of iterations. If further dihedral angles are considered, combinations of particular ranges of these angles become even rarer. Two major implications follow from that observation: An extensive sampling of the entire conformational space

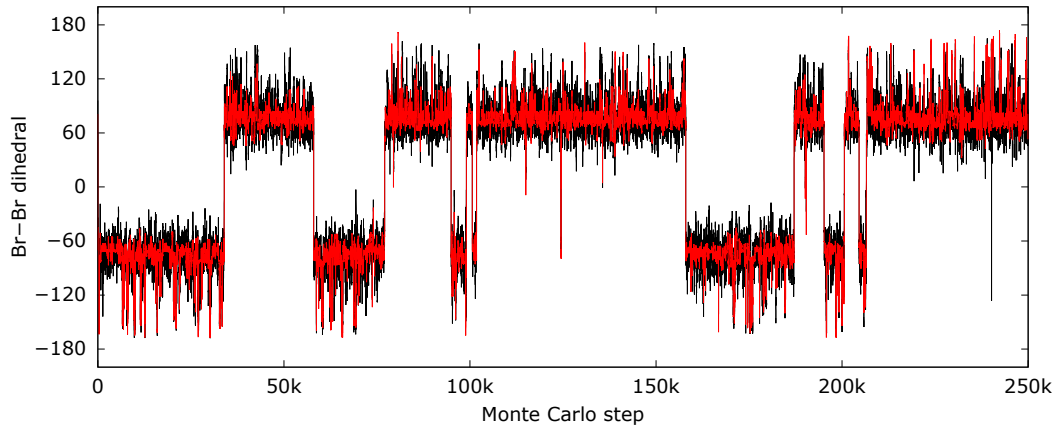


Figure 4.1: Conformational trapping effect associated with MD and MC simulations: $(-)$-β-HBCD dihedral sampled using an HMC scheme (black) and after CG minimization (red).

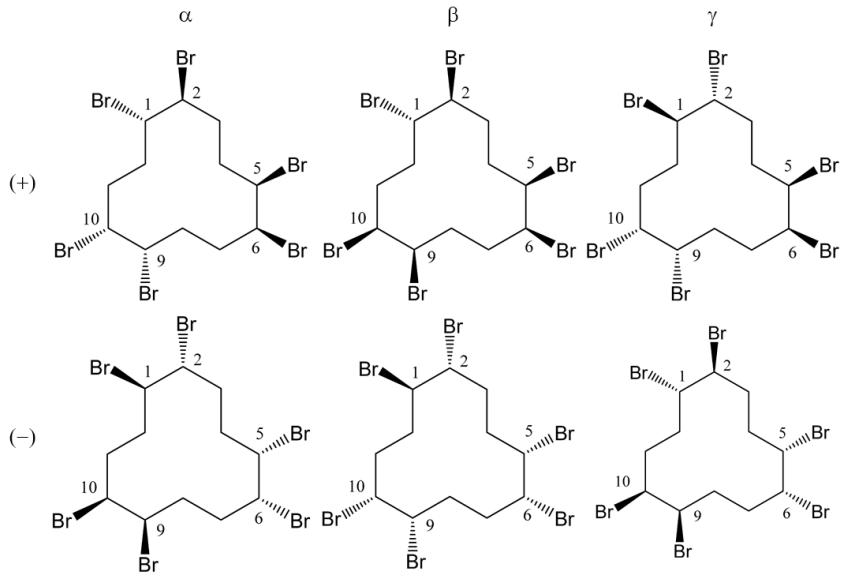


Figure 4.2: Six major HBCD stereoisomers: three diastereomeric pairs of enantiomers. Reprinted from the original publication of Durmaz et al. 2012.

will most probably take more time than we are willing to spend. Ensuing from that, it might often be advisable to focus on the most likely states that are associated with particularly low potential energies and, consequently, the by far highest statistical Boltzmann weights.

In the light of these considerations, a proper choice of a preferential conformation of the guest molecule is a key step to the estimation of binding affinities for host–guest systems. For this reason we will, by taking the example of HBCD isomers, elaborate in the following how global minimum conformations of small molecules for the purpose of binding affinity calculations can easily be determined and verified. Until 2013, HBCD was widely used in upholstery textiles and polystyrene foams,[195,196] though, it was increasingly considered as an *persistent organic pollutant* (POP) that accumulates in many environmental compartments as well as in biota including humans.[197–200] Consequently, HBCD underwent a risk assessment commissioned by the *European Chemicals Agency* that finally lead to its *Annex A* consideration for elimination by the *Stockholm Convention on POPs* in the year 2013.[201]

Due to technical reasons, there exist six major HBCD isomers that are able to interconvert one into another.[199,200] As depicted in Figure 4.2, each of the three HBCD diastereomers α, β, and γ from left to right can appear in a form denoted as $(-)$-HBCD or as its mirror-inverted $(+)$-enantiomer. HBCD stereoisomers are characterized by three bromine pairs each forming a Br–C_i–C_{i+1}–Br moiety where $i \in \{1, 5, 9\}$. According to the absolute configuration associated with six HBCD stereo centers listed in

Table 4.1: Absolute configuration and CIP nomenclature of the six major HBCD stereoisomers. Reprinted from the original publication of Durmaz et al. 2012.

Stereoisomer	CIP nomenclature	Stereoisomer	CIP nomenclature
$(-)$-α-HBCD	1R,2R,5S,6R,9R,10S-HBCD	$(+)$-α-HBCD	1S,2S,5R,6S,9S,10R-HBCD
$(-)$-β-HBCD	1R,2R,5S,6R,9S,10R-HBCD	$(+)$-β-HBCD	1S,2S,5R,6S,9R,10S-HBCD
$(-)$-γ-HBCD	1S,2S,5S,6R,9R,10S-HBCD	$(+)$-γ-HBCD	1R,2R,5R,6S,9S,10R-HBCD

Table 4.1 along with their Cahn-Ingold-Prelog (CIP) nomenclature, every Br–C_1–C_2–Br moiety shows either *R-R* or *S-S* chirality whereas the two other moieties associated with C_5–C_6 and C_9–C_{10} are of a mixed type (*R-S* or *S-R*). This system of stereoisomers is particularly suitable for the evaluation of sampling as well as global energy minimization routines due to several levels of symmetry that must be reflected by theoretical energies if the sampling is supposed to be converged against the thermodynamic equilibrium distribution. One the one hand the two enatiomeric counterparts $(\pm)$ of any HBCD diastereomer must yield analogous internal energy distribution. On the other hand the distribution of states should reflect the radial symmetry given in case of α- and γ-HBCD stereoisomers.

High-temperature HMC sampling and convergence diagnostics

The trace plot shown in Figure 4.1 is based on five 50.000 step MCMC chains resulting from an HMC sampling of $(-)$-β-HBCD at 1500 K. Setting the temperature to such artificially high values yields large kinetic energies and momenta increasing the efficiency of an conformational sampling. This idea resembles the key feature of the *replica exchange* or *parallel tempering* method which performs several independent MC runs at distinct temperatures and occasionally swaps configurations in order to improve the dynamic properties of the sampling algorithm.[202] Moreover, Figure 4.1 indicates that simulations comprising 50.000 iterations might be insufficient for a rigorous sampling of drug-sized molecules, although the HMC method is known for its more efficient conformational sampling compared to MD.[183] Substantial structural changes due to large torsional flippings are observed few times only in Figure 4.1. Nevertheless, by using the HMC method, one is usually interested in the stationary distribution of states that, once achieved, does not change significantly any more during further simulation. Whether a sampling process has reached the stationary distribution or not can be evaluated using mathematical tools for convergence diagnostics. The convergence evaluation method applied here is attributed to Gelman and Rubin who proposed to monitor convergence by running multiple MCMC chains and comparing *within* and *between chain*

variances with respect to some parameter θ.[203,204]

Ensuing from the original paper, nine out of twelve dihedrals forming the cyclododecane ring were chosen as basis for the convergence check. Due to the cyclic structure, the ring geometry is unambiguously defined by those nine internal degrees of freedom. In order to remedy issues with variances of cyclic entities, each dihedral ϕ_k was represented twice among the parameters, by its cosine and sine function

$$\theta^{(k,\sin)} = \sin(\phi_k); \quad \theta^{(k,\cos)} = \cos(\phi_k)$$

resulting in a set of 18 parameters in total. Since $\cos(x) = \cos(-x)$, using the cosine expression enables to "close the circle" at $\pm 180°$ and, thus, to reduce the fallacious variance associated with cyclic dihedrals around 180°. Indeed, a typical torsion potential about a bond between two sp3-hybridized carbons as sketched for butane in Figure 1.5 reveals an energetically favorable dihedral at this critical value. The two other preferential conformations correspond to dihedrals at $\pm 60°$ which holds for HBCD as well as sketched in Figure 4.1. However, as a consequence of the cosine function's axial symmetry, it would erroneously consider these two major conformations identical. This type of error in turn is compensated by the sine expression which yields significantly different function values for $-60°$ and $+60°$ but, fortunately, zero for angles around 180° where the cosine function is intended to dominate. Thus, the maximum of the two trigonometric functions would provide a good indication of any dihedral's variance. Prior to the high-temperature HMC sampling, all compounds were parametrized according to the Merck Molecular Force Field (MMFF) which was particularly designed for small drug-sized molecules.[127] On the basis of $m = 5$ HMC Markov chains built at $T = 1500$ K including $n = 10^5$ iterations with 30 MD steps per iteration, the convergence of each parameter θ was investigated independently following the proposal of Gelman and Rubin. Having calculated the mean within-chain variance

$$W = \frac{1}{m} \sum_{j=1}^{m} \frac{1}{n-1} \sum_{i=1}^{n} \left(\theta_{ij} - \bar{\theta}_j\right)^2, \tag{4.1}$$

which likely underestimates the true variance unless all points in the conformational space are reached, and the variance between chains

$$B = \frac{n}{m-1} \sum_{j=1}^{m} \left(\bar{\theta}_j - \bar{\bar{\theta}}\right)^2, \tag{4.2}$$

the variance of the stationary distribution was estimated as a weighted average of Equations 4.1 and 4.2

$$\hat{\mathrm{Var}}(\theta) = \left(1 - \frac{1}{n}\right) W + \frac{1}{n} B$$

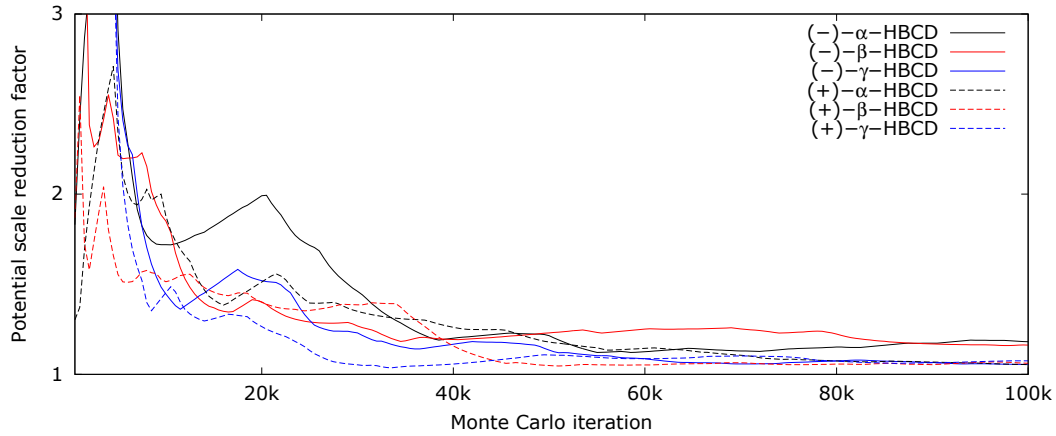


Figure 4.3: HMC convergence diagnostics of the six major HBCD stereoisomers according to Gelman and Rubin: PSRF trace plot using five Markov chains of 100 k steps each.

used to determine the potential scale reduction factor (PSRF)

$$\hat{R} = \sqrt{\frac{\hat{\mathrm{Var}}\,(\theta)}{W}}$$

which is interpreted as a relative distance from the stationary distribution. The larger this value, for which a cutoff value in the range of 1.1-1.2 is recommended, the more additional sampling is required for a proper convergence. Convergence was evaluated after every 500 HMC iterations and separately for the 18 chosen parameters. At each time, the largest PSRF representing the dihedral angle that more than any other prevents from converging to the stationary distribution was determined. Figure 4.3 shows the PSRF behaviour for the six HBCD stereoisomers under investigation. It leads to the assumption that no significant improvement occurs anymore after about 50k HMC steps and that we do well with at least that number of iterations for the purpose of a global minimum search of small compounds. However, it should be noted that the PSRF value can mislead. For instance, if all MCMC chains reside in the same tight subdomain of space by collectively neglecting all other parts of space, they will still most likely yield an erroneously promising value close to one which indicates convergence.

Global minimization and evaluation

At this point, symmetry properties of HBCD stereoisomers mentioned above come into play as they offer an additional level for convergence diagnostics. To be precise, the energy distribution with respect to a particular internal coordinate should be similar to that of another coordinate if both degrees of freedom are embedded in identical physical

Table 4.2: Global potential energy minima of *anti* and *gauche* subspaces given in [kJ/mol] with respect to all C_iBr–$C_{i+1}Br$-moieties of the six major HBCD stereoisomers.

		α		β		γ	
Enantiomer	Dihedral	anti	gauche	anti	gauche	anti	gauche
$(-)$	C_1C_2	253.7	238.7	263.6	249.5	256.7	256.7
	C_5C_6	272.6	238.7	285.3	249.5	275.1	256.7
	C_9C_{10}	272.6	238.7	276.2	249.5	275.1	256.7
$(+)$	C_1C_2	253.7	238.7	263.6	249.5	256.7	256.7
	C_5C_6	272.6	238.7	285.3	249.5	275.1	256.7
	C_9C_{10}	272.6	238.7	276.2	249.5	275.1	256.7

environments due to structural symmetries. Instead of comparing energy distributions directly, one can as well select geometries representing particular critical points such as "locally global" minimum conformations. For this purpose, all geometries generated by the HMC algorithm underwent a conjugate gradient (CG) minimization with at most 5000 iterations if the maximum force at some iteration step had not ended up below the tolerance value of 2 kJ mol^{-1} nm^{-1} before. Figure 4.1 exemplarily shows how the variance associated with one particular dihedral angle of $(-)$-β-HBCD is narrowed down to less states (red plot) that are related to local minima compared to the broader HMC sampling (black plot). Afterwards, these optimized geometries were divided into different domains according to some internal coordinate (dihedral) and the lowest minimum energy conformation of each domain was identified. Table 4.2 shows a pair of optimal potential energy values per stereoisomer and Br–C_i–C_{i+1}–Br-moiety denoted as *anti* and *gauche* and resulting from the decomposition of each bromine-bromine dihedral ϕ into two subspaces representing *anti* ($|\phi| > 120°$) and *gauche* ($|\phi| < 120°$) conformations, respectively. Obviously, the enantiomeric counterparts $(+/-)$ of any diastereomer correctly yielded identical optimal energies regarding all Br–C_i–C_{i+1}–Br-moieties. Furthermore and as indicated by yellow blocks in Table 4.2, the radial symmetry given in case of α- and γ-HBCD is well reflected as well because *anti*/*gauche* optima of both the Br–C_5–C_6–Br and Br–C_9–C_{10}–Br moieties yielded the same pair of values, 272.6/238.7 kJ mol^{-1} for α-HBCD and 275.1/256.7 kJ mol^{-1} for γ-HBCD. Both observations clearly confirm a sufficient sampling in terms of convergence. Interestingly, except for three values, all localized global minima presented in Table 4.2 based on 10^5 HMC iterations had already been calculated using 10^4 HMC steps only. The exceptions were related to the Br–C_9–C_{10}–Br group of $(-)$-β-HBCD with 278.4 instead of 276.2 kJ mol^{-1} and of $(\pm)$-γ-HBCD with 289.5 instead of 275.1 kJ mol^{-1}. After the short sampling, the exceptional stereoisomers' maximum PSRF value ranged

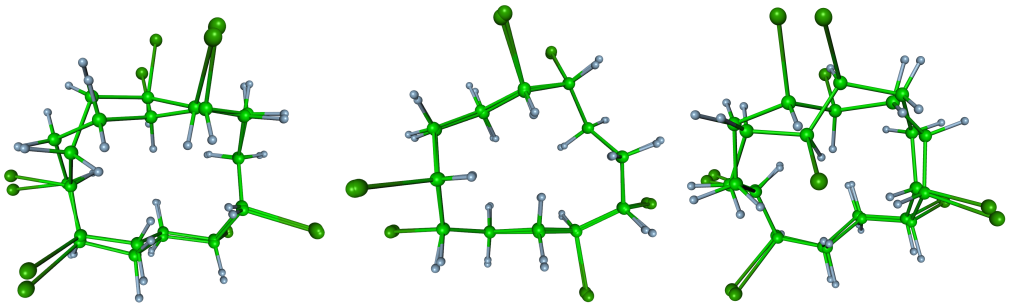


Figure 4.4: Alignment of globally minimized major HBCD conformations to respective X-ray structures. Ensemble was parametrized according to the AMBER force field, sampled using HMC, and minimized with the CG method. From left to right: α, β, γ diastereomer. Reprinted from the original publication of Durmaz et al. 2012.

between 1.4 and 1.7.

Finally, for the three HBCD diastereomers (α, β, and γ) crystallographic data was available in the Cambridge Structural Database (CSD) under the IDs 633325, 617557, and 633326, respectively,[205] which were used for an evaluation of theoretical minima. Figure 4.4 shows structural alignments of global minimum conformations of HBCD diastereomers with respective crystallographic data. In addition to the visual inspection, the similarity of predicted and crystallographic structures was quantified by means of both Cartesian and torsional root mean square deviations (RMSD) as summarized in Table 4.3. In general, the RMSD value is defined as

$$\mathrm{RMSD} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\xi_i^2}$$

where ξ_i was either equated with Cartesian coordinates directly or, in case of dihedral angles, computed as

$$\xi_i = \begin{cases} \phi_i^{\mathrm{sim}} - \phi_i^{\mathrm{exp}} & \text{if } |\phi_i^{\mathrm{sim}} - \phi_i^{\mathrm{exp}}| \leq 180^\circ, \\ 360^\circ - |\phi_i^{\mathrm{sim}} - \phi_i^{\mathrm{exp}}| & \text{if } |\phi_i^{\mathrm{sim}} - \phi_i^{\mathrm{exp}}| > 180^\circ. \end{cases} \tag{4.3}$$

The case differentiation is necessary since any rotational distance can be expressed by two values, ϕ and $360^\circ - \phi$. Equation 4.3 guarantees that the smaller one (at most 180°) of these two distances is used for similarity measurement. Both AMBER as well as the Merck force field yielded comparably reasonable Cartesian RMSD values that are particularly remarkable for β-HBCD evaluated as less than 0.2 Å. Indeed, this diastereomer displays perfect matching as depicted in Figure 4.4 whereas the other two diastereomers show considerable deviations in certain regions. The visual observations better correlate with the torsional measurement on the basis of AMBER. With MMFF,

Table 4.3: Cartesian and angular RMS deviation of computed global HBCD minima from crystallographic geometries.

	Cartesian RMSD [Å]		Angular RMSD [°]	
Diastereomer	MMFF	GAFF	MMFF	GAFF
α-HBCD	0.49	0.49	52.8	52.8
β-HBCD	0.08	0.16	2.0	5.0
γ-HBCD	1.23	0.73	95.8	40.3

a significantly worse conformational similarity of γ-HBCD to the corresponding crystal structure was computed. All in all, the observations indicate a high suitability of both force fields to the description of small organic molecules where AMBER achieves slightly better results as a whole and, especially, regarding the high conformational complexity due to many rotatable bonds coupled with an assumingly notable strain caused by the cyclic sructure. The significant deviations associated with α/γ-HBCD may be attributed to several causes, for instance, solvent or chemical effects during the crystallization process or simply due to the fact, that crystallographic conformations do not necessarily need to match those associated with global energy minima under solvation or vacuum conditions.

4.2 Efficient clustering of molecular conformations

Sometimes, a single global minimum conformation as the only representative of a drug-sized molecule is not sufficient for further analysis. One might be interested in an exhaustive sampling for the purpose of extensive conformational analysis or free energy estimations regarding small molecules or large host–guest systems. As already pointed out, complex systems hardly exhibit significant structural changes during MD simulations. However, instead of running one trajectory for a very long time one could run multiple time series simultaneously from a set of initial structures corresponding to significantly differing regions of the configurational space. Cluster algorithms addressing this task are commonly termed *meshless discretization* methods.

We will in the following elaborate an as simple as efficient centroid-based algorithm for the generation of a k-split, the selection of a set $C^{(k)}$ of k cluster representatives c_j out of n states (geometries/frames) of a molecular trajectory/ensemble where $k \ll n$. The cluster centers c_j are iteratively selected possibly from a high-temperature sampling unless at least one of two termination criteria is met: either a particular preassigned number k of centers is obtained or the Euclidean distance $d_{c_j}(i)$ of every frame i to

its *nearest* representative c_j lies beneath a given torsional distance limit d_{max}. As a consequence, *every* domain of the conformational space physically accessible at extremely high temperatures is represented by some geometry regardless of its likeliness to occur in a statistical ensemble. Favorable local minima as well as rarely sampled conformers often associated with transition states are taken into account for further analysis possibly including path reconstruction, molecular kinetics, and free energies. Insofar, this approach is entirely contrary to clustering tools such as *k*-means that rather try to identify points amidst many others representing particularly dense domains.[206] Suitable input for the presented strategy can be constructed from an extensive high-temperature HMC or MD sampling as described in the previous section since it is likely to cover most of the relevant conformational space.

Algorithmic details

Given an ensemble consisting of n molecular geometries, the algorithm's central data structure is a membership array M_j of length n updated with each cycle j of an iterative process. M_j consists of torsional Euclidean distances $d_{c_j}(i)$ of each frame i to its currently nearest cluster center c_j. The procedure is illustrated in detail by Figure 4.5 exemplarily using an eight frames subset of an HMC sampling of pentane. At the very beginning, array M is initialized with extraordinarily large values, $M_0 := (\infty)$. Generally, the frame associated with the largest value of M_j is chosen as next cluster center c_{j+1}. In the special case of the first iteration where $M_0 := (\infty)$, an arbitrary state may be selected. With no loss of generality, we decided for the first state of the ensemble, $c_1 = 1$. Afterwards, a vector $d_{c_{j+1}}$ of distances $d_{c_{j+1}}(i)$ from any frame i to the new centroid c_{j+1} is calculated. An updated version M_{j+1} is then constructed by taking element-wise minima of the two vectors $d_{c_{j+1}}$ and M_j. In other words, if the current cluster center c_{j+1} is closer to some state i than the closest of all the previous centers $\{c_1, \cdots, c_j\}$, then state i is assigned to cluster $j+1$ represented by centroid c_{j+1} and the corresponding distance $d_{c_{j+1}}(i)$ is allocated to the i^{th} field of M_j. Certainly, the values of M decrease monotonically along iteration j since

$$M_j(i) \leq M_{j-1}(i) \quad \forall\, i \in \{1, n\} \text{ and } \forall j \in \{1, k\}$$

Finally, after k iterations, the minimal array M_k contains for each frame the distance to its nearest representative c_j. Since during each iteration j the frame corresponding to the maximum distance value of array M_j is chosen as next centroid, the presented method will be referred to as *maxdist* algorithm in the following. Thus, by simply selecting the frame with the largest distance to its nearest centroid as the next one,

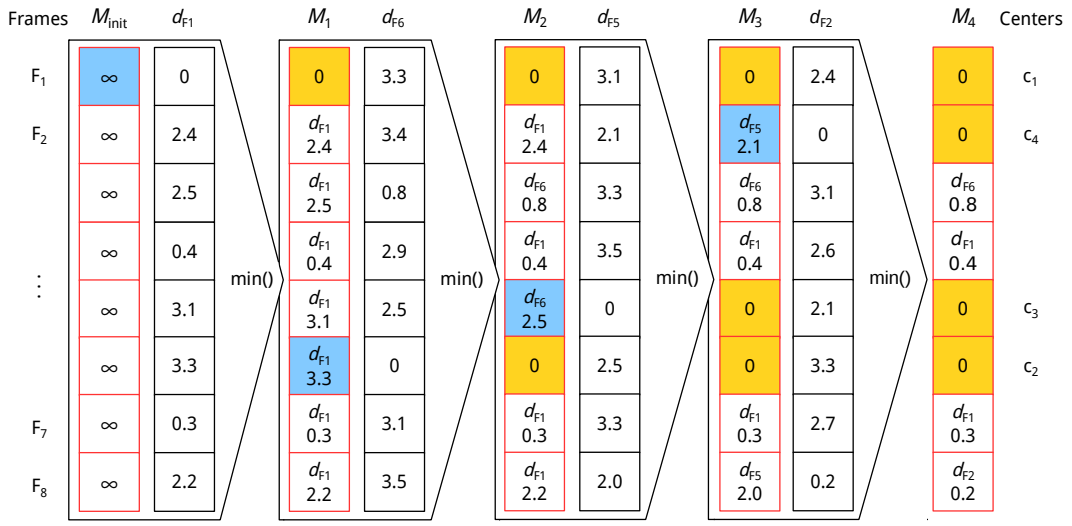


Figure 4.5: Meshless discretization strategy of *maxdist* algorithm: Vector M_j keeps track of each frame's Euclidean distance to its nearest centroid. It is constructed as the element-wise minimum of its predecessor M_{j-1} and frame distances d_{c_j} to the current centroid c_j.

a complete representation of space in particular at complete disregard for physically irrelevant domains is guaranteed.

Regarding the HMC sampling of pentane (Figure 4.5), the clusters have been identified on the basis of its two most relevant dihedral angles ϕ_1 and ϕ_2 and a maximal torsional distance d_{max} set to $\pi/3 \approx 1.05$. In analogy to butane sketched in Figure 1.5, ϕ_1 is related to the angle spanned by $C_1{-}C_2{-}C_3{-}C_4$ and, in addition, ϕ_2 related to $C_2{-}C_3{-}C_4{-}C_5$. First, M_0 was initialized with infinity. To simplify matters we chose the first frame (F_1) as the first representative, $c_1 := F_1$ corresponding to the top row and left-most blue-marked sample in Figure 4.5 and calculated d_{F_1} afterwards. Taking the element-wise minimum of M_0 and d_{F_1} yielded an updated array M_1 which is identical to d_{F_1} since all of the calculated distances are less than the corresponding ones of array M_0 initialized with ∞. Frame F_6 (blue-marked sample in vector M_1) associated with the largest distance to the current (initial) center is then chosen as next representative c_2. Again, for each frame i its distance $d_{F_6}(i)$ to this new representative is computed and element-wisely compared to the current array M_1 of least frame-wise distances yielding an once again updated array M_2. After $k = 4$ iterations we arrived at a set of four representatives, $\{F_1, F_2, F_5, F_6\}$, (yellow-colored fields in M_4) to which all other frames are assigned with some minimal distance below $d_{\text{max}} = \pi/3$. Figure 4.2 shows a scatter plot of the same set of eight pentane samples within its conformational space projected onto two internal dimensions (ϕ_1 and ϕ_2). If we had terminated the procedure after $k = 3$ iterations (left subfigure), the maximum torsional distance between any pair of representative and sample would have amounted to 2.1 radians which is related to the

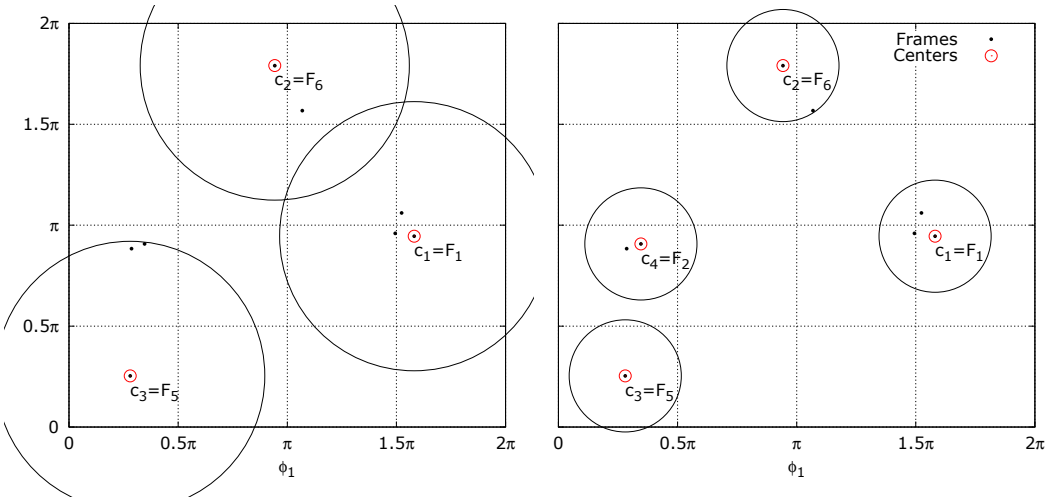


Figure 4.6: Meshless discretization strategy of *maxdist* algorithm illustrated by a 3-split (left) of eight two-dimensional conformational objects plus one successive iteration cycle yielding a 4-split (right).

second frame F_2 and its nearest center c_3 (=F_5). According to that 3-split, frames F_1, F_6, and F_5 were chosen as cluster centers. They are marked by red circles and surrounded by grey-colored spheres defining the maximum torsional distance. Moreover, F_2, the righter one of the pair of points located above c_3 at $\phi_2 \approx 3$ right at the encompassing circuit, constituted the next cluster candidate. With its selection during the final iteration, all large encompassing circles immediately became significantly smaller namely 0.8 related to the distance between cluster c_2 and the frame below. A naive look at the 3-split in Figure 4.2 might raise the question why the third representative $c_3 = F_5$ was not selected earlier since the distance between the two previous centers, c_1 and c_2, is smaller than their distances to c_3. However, the visual appearance is deceiving because we are dealing with periodic quantities that repeat every 360° in each dimension so that the shortest circular distance between two points might be less than what the figure suggests. In short, the algorithm performs as follows:

1. Initialize vector M_0 of minimum distances per frame with ∞.
2. Select next cluster center c_j either as the frame associated with the largest value in M_{j-1} or, in the initial case only, arbitrarily from M_0.
3. Calculate vector $d_{c_j}(i)$ of distances between every frame i and the current representative c_j.
4. Update distances in vector M with the element-wise minimum of the two vectors computed previously, $M_j = \min\left(M_{j-1}, d_{c_j}\right)$.
5. Repeat steps 2.-4. unless stopping criterion is reached.

Due to nk evaluations of distances and pairwise minima, the presented algorithm's time complexity is in the order of $\mathcal{O}(nk)$. It can even be considered as linearly scaling, $\mathcal{O}(n)$, since in realistic examples $k \ll n$. However, the complexity may increase if either only few states have been sampled or, depending on the stopping criterion and the number of torsional degrees of freedom, a large number of clusters is required. In a worst case scenario, if all frames are chosen as cluster centers ($k = n$), the time complexity would amount to $\mathcal{O}(n^2)$. In practice, though, given a large set of n geometries, k will be negligible compared to n as illustrated by Figure 4.7 for *n*-pentane as a representative for molecules with two torsional degrees of freedom. It shows the number of required clusters depending on the number of frames randomly selected as input out of a 400 k step HMC sampling of pentane. For each input size ranging from 1 to 5000 states the number of clusters was averaged over ten random sets. In addition, the rotational distance cutoff d_{max} was varied from $\pi/6$ to π. With an increasing number n of samples, the number of clusters seem to, sooner or later, converge against some value k which clearly depends on d_{max}. The maximal number of clusters is indeed limited since from a certain point on the entire conformational space (respectively all input frames) will be covered by spheres encompassing the cluster centroids c_j at radius d_{max}. Of course, large d_{max} values enhance convergence as illustrated by lower plots in Figure 4.7). Regarding space complexity, the entire procedure requires to keep track of only two arrays (M_j and d_{c_j}) consisting of n floating point values each.

Stopping criteria and convergence diagnostics

As already stated, a critical part of the cluster algorithm is related to the stopping criterion for which *maxdist* offers two options. On the one side, one can predefine a

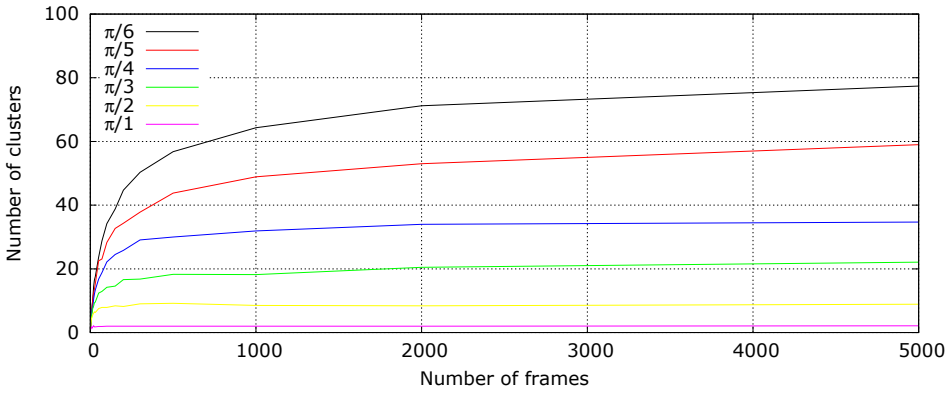


Figure 4.7: Number of centroids vs. number of frames using the *maxdist* algorithm with different rotational distance cutoffs.

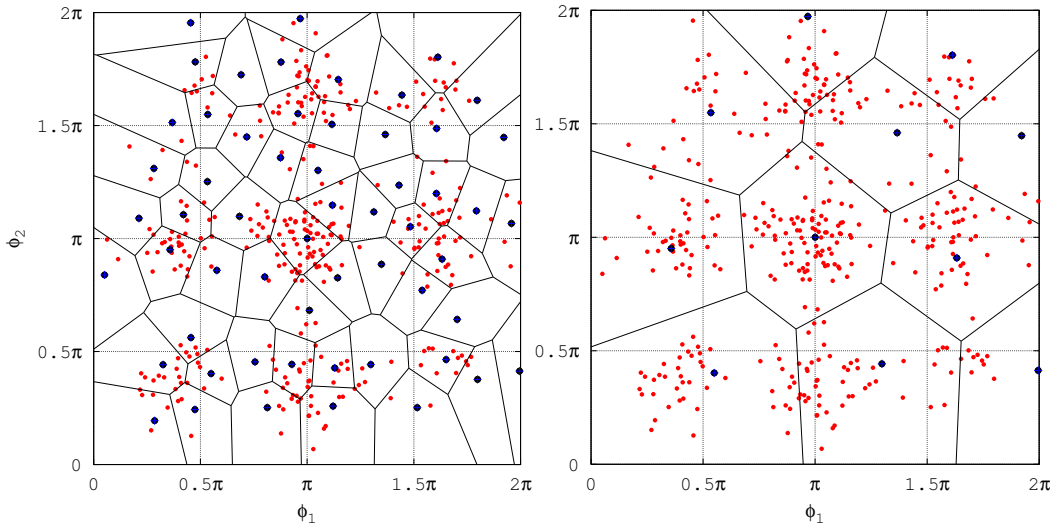


Figure 4.8: Meshless space discretization and Voronoi diagram in two torsional pentane dimensions according to *maxdist* algorithm: $d_{\max}=\pi/6$ yields 57 representatives (left), $d_{\max}=\pi/2$ yields 11 (right).

particular number k of clusters causing the algorithm to abort exactly after k iterations. On the other side, it is possible to limit the maximum allowed distance between any frame and its nearest cluster to a predefined value $d_{\max}$. Which criterion suits more the user's needs is up to the problem. If one wants to be sure that no possible conformer has an Euclidean distance to its representative larger than a particular value, he would probably predefine a distance limit $d_{\max}$. If, in contrast, the number of representatives shall be limited, this is easily achieved by a proper setting. The two options do not exclude each other. Rather, the same sequence of representatives is calculated (up to and theoretically beyond the point where the first criterion is met). Again, we consider a 1500 K HMC sampling consisting of 400 k pentane conformers. As we already know, the pentane conformation is well approximated by the two all-carbon dihedral angles ϕ_1 and ϕ_2. Accordingly, each red dot of the 2D-scatter plot in Figure 4.2 is associated with a particular pentane geometry. The distribution of 2000 randomly chosen states on the basis of five Markov chains as described earlier obeys the Boltzmann distribution of pentane energies since low energy of the potential energy surface (PES) are covered more densely than unfavorable regions of higher energies.[207] At least nine densely sampled regions have been identified which obviously correspond to the nine major out of eleven known pentane conformations. Frames chosen as cluster representatives are marked through large dark dots and associated with a Voronoi cell in Figure 4.2. A 57-split and a 11-split were necessary in order to fulfill the $d_{\max} = \pi/6$ (left subfigure) and, respectively, the $d_{\max} = \pi/2$ criterion (right subfigure). It should be

noted that the relationship

$$C^{(k)} \subset C^{(l)} \quad \forall \; k < l,$$

is valid for any two sets $C^{(k)}$ and $C^{(l)}$ of clusters corresponding to a k and, respectively, l-split generated with *maxdist*. Consequently, 11-split representatives form a subset of those associated with the 57-split. In any case every point is assigned to at least one cluster centroid at a distance less than $d_{\max}$. Obviously the *maxdist* Voronoi partitioning does not tend to match a physical decomposition of space at constant temperature (following the Boltzmann distribution). Since the intention of this algorithm is to represent the entire conformational space rather than dense regions preferentially, the two given domain decompositions seem more or less reasonable. Both related subsets $C^{(k)}$ include lone geometries from low-density transition regions as well as geometries located in quickly descending high-probability areas around local pentane minima. However, the coarser decomposition on the right of Figure 4.2 exhibits at least one minimum (e. g. at 300°,60°) that is not satisfactorily represented. A distance limit set to $d_{\max} = \pi/4$ yielded 33 centers that sufficiently cover two dimensions associated with torsional degrees of freedom by using much less centers than the 57-split. Considering a typical periodic distribution of rotational barriers with a maximum either every 120 or every 180 degrees, it appears reasonable to choose a distance limit $d_{\max}$ less than half the *smallest* possible periodic offset,

$$d_{\max} \leq \frac{\pi}{3} = 60°. \tag{4.4}$$

With increasing dimensions, though, this algorithm quickly identifies a great many number of clusters. Solid lines in Figure 4.9 representing *maxdist* results show possible tradeoffs between the number of clusters k and the maximum allowed distance for a couple of torsional degrees of freedom N_{f} ranging from one to five. Each setting's value was averaged on the basis of ten *maxdist* runs on independent sets of 5000 states randomly selected from the 500 k high-temperature HMC sampling of HBCD. Torsional angles used here originate from the cyclic carbon scaffold of $(-)$-β-HBCD as shown in Figure 4.2. Obviously, the best up-rounded $\lceil\cdot\rceil$ approximation of k follows hyperbolic functions

$$k = \left\lceil \frac{1}{(d_{\max}/2\pi)^{N_{\mathrm{f}}}} \right\rceil \tag{4.5}$$

of the distance limit where the function order corresponds to the dimension N_{f}. The progress of k calculated with Equation 4.5 and using five different degrees of freedom is depicted in Figure 4.9 through dashed lines. The normalization of $d_{\max}$ by 2π is necessary for the following reason: Due to its discrete nature, the function value (the

number of chosen clusters) cannot be less than one, $k \geq 1$. It evaluates approximately as one for $d_{\max}$ values equal to or larger than 2π. Since all hyperbolic functions of the type x^{-n} intersect at (1,1), a scaling of the horizontal axis is necessary in order to shift the intersection point (1,1) to (2π,1). For a $d_{\max}$ domain ranging from 0.01 to 2π, Equation 4.5 yields remarkable coefficients r_{N_f} of correlation: $r_1 = 0.986$, $r_2 = 0.984$, $r_3 = 0.999$, $r_4 = 0.998$, and $r_5 = 0.987$. However, with increasing function values the hyperbolas' (dashed lines) deviations from respective simulation results (solid lines) grow increasingly, particularly affecting systems associated with higher degrees of freedom. The consequent introduction of a correction factor $(1 - \alpha)$ into the exponent of Equation 4.5 yields a model

$$k = \left\lceil \left[\frac{2\pi}{d_{\max}} \right]^{N_f(1-\alpha)} \right\rceil \tag{4.6}$$

that copes better with the algorithm's results for large N_f as illustrated by dotted lines if $\alpha = 0.06$: $r_1 = 0.982$, $r_2 = 0.989$, $r_3 = 0.998$, $r_4 = 0.999$, and $r_5 = 0.991$. Intuitively, α might reflect structural and dynamic properties of the molecular system under investigation. For instance, it could be considered as the deviance of the sampling from an *ideal distribution* of geometries in the conformational space which might be characterized by infinitely dense and uniformly spread points. That is, the smaller the domain in space the internal molecular flexibility is restricted to the less cluster will be required leading to a larger value for α. Either way, in order to be able to use Equation 4.6 for an estimation of the expectable number of clusters given a particular

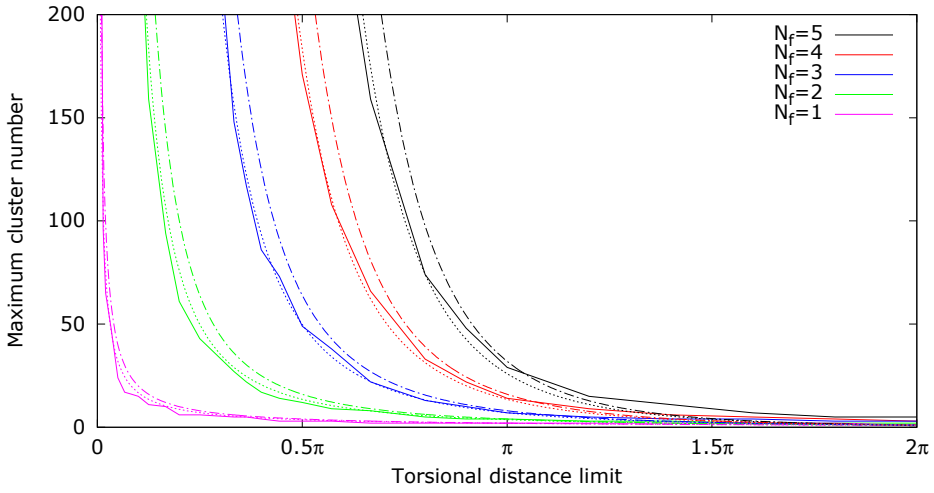


Figure 4.9: Maximum number of clusters computed for 500 k HBCD geometries dependent on the distance limit and the number of torsional degrees of freedom N_f using the *maxdist* algorithm. **Solid lines**: clustering results, **dashed lines**: hyperbolic fit according to Equation 4.5, **fine dotted lines**: corrected hyperbolic fit according to Equation 4.6.

N_f and d_{max}, α could, in advance, be parametrized generally or specifically using a set of molecules with differing or, respectively, similar torsional flexibilities. An efficient implementation of the linearly scaling algorithm is available within the framework of the software *ZIBgridree*[208] which was designed for conformational analysis of small as well as macromolecules. It should be noted that the clustering method focusses on an exhaustive representation of the conformational space by few cluster centers rather than iteratively choosing the optimal representative of each cluster as carried out by the *k*-means algorithm which is widely used in spite of its poor time performance under certain conditions.[206]

4.3 Host–guest binding mode decomposition

As already pointed out and illustrated through Figure 4.1, significant conformational changes are rare events during an MD simulation. This observation is also valid regarding the relative orientation of a host and guest molecule in complex systems including receptor–hormone and enzyme–substrate complexes. It is therefore advisable, to decompose the conformational space for independent simulations in order to identify the most preferential binding mode(s) for the purpose of affinity or conformational analysis. In the following, we will describe a simple and straight-forward strategy to the decomposition of conformational space and use respective representatives for the prediction of quantitative or relative binding affinities in successive chapters. These are calculated on the basis of either the most preferential mode that needs to be determined with some convenient method or a weighted sum of all modes. The regular decomposition is based on symmetry properties of an icosahedron consisting of twelve vertices and 20 faces. A rotational decomposition according to the icosahedron which is (besides its dual, the dodecahedron) the Platonic solid with the highest order of symmetry yields 60 uniformly distributed orientations (binding modes) of the guest molecule at the host molecule's binding site. They serve as initial conformations for independent MD runs of the molecular complex. Though this procedure highly increases the computational effort, we are confident to snatch the preferential binding mode(s) in contrast to ordinary molecular docking algorithms. Figure 4.10 illustrates the idea using the antibiotic sulfamethoxazole depicted on the left as an example. We will be faced with this drug again in Chapter 7. Of course, there have already been attempts to estimate the binding affinity on the basis of multiple binding modes.[84] However, a systematic investigation of the entire space of binding modes for automatization purposes if no experimental information about the correct binding mode exists is still missing.

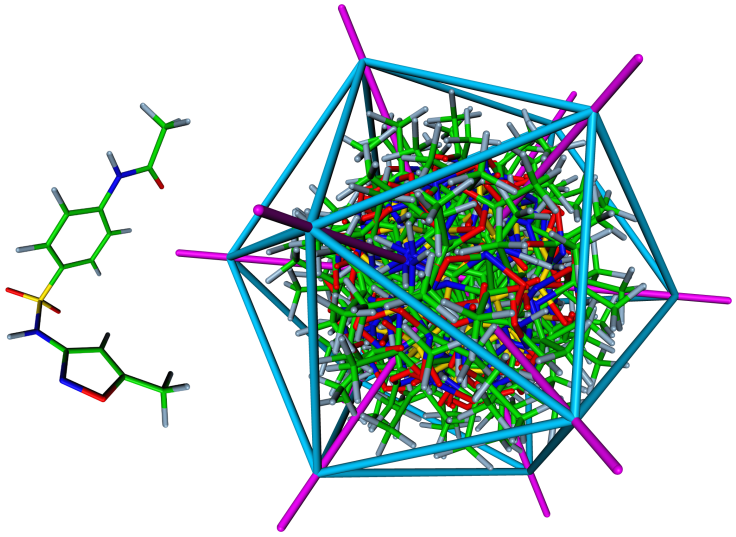


Figure 4.10: 60 uniformly distributed rotational binding modes of sulfamthoxazole (left) according to the icosahedron on the right providing starting conformations for MD simulations

Algorithmic details

One of the possible ways to quickly generate a set of 60 uniformly distributed binding modes entails in using rotation matrices R_i which are of dimension 3×3 for our aim. Any position vector $v \in \mathbb{R}$ can be rotated

$$v' = R_i\left(\alpha\right)\, v$$

by an angle α about some axis i resulting in a transformed vector v'. In analogy to the concept of *Euler* angles, there exist three rotation matrices

$$\overbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & -\sin\alpha \\ 0 & \sin\alpha & \cos\alpha \end{pmatrix}}^{R_x\left(\alpha\right)} \quad \overbrace{\begin{pmatrix} \cos\alpha & 0 & \sin\alpha \\ 0 & 1 & 0 \\ -\sin\alpha & 0 & \cos\alpha \end{pmatrix}}^{R_y\left(\alpha\right)} \quad \overbrace{\begin{pmatrix} \cos\alpha & -\sin\alpha & 0 \\ \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{pmatrix}}^{R_z\left(\alpha\right)} \tag{4.7}$$

corresponding to the three spatial axes x, y, z. Successive rotations about axis i and j (in this order) are achieved by multiplying vector v with one transformation matrix after another from left

$$v' = R_j\left(\beta\right)\, R_i\left(\alpha\right)\, v.$$

One should bear in mind that temporarily, during any rotation process, the object under consideration must be geometrically centered on the coordinate system's origin

since we want to keep its initial position in space, most likely the binding site. Now considering any ligand molecule as an icosahedron, each pair (i, i^*) of two opposing vertices provides one rotational axis L_i, $i \in 1, \ldots, 6$ (magenta-colored in Figure 4.10) about which the object can be rotated in five 72 degree steps per full cycle. Having chosen some axis as the first one, this yields five binding modes if we save a snapshot of the molecule after each step. Five further orientations associated with the same axis are gained after having swapped the two corresponding vertices through a respective 180 degree rotation of the icosahedron resulting in ten binding modes altogether per rotational axis. Since there exist five further rotational axes defined by ten further vertices where each axis provides ten further orientations we arrive at a total number of 60 binding modes as sketched within the icosahedron on the right of Figure 4.10. All that has to be done is aligning one of these axes after another to the first axis in a convenient way. Let us first define the (arbitrary) initial position according to Figure 4.11. It is characterized by three assumptions:

- The icosahedron is geometrically centered on the point of origin.
- Both vertex 1 and its opposing vertex 1^* lie on the x-axis serving as the central rotation axis from which poses are sampled.
- Vertices 5 and 6 define a line parallel to the xy-plane at a unique negative z-value.

Algorithm 1 illustrates the detailed procedure in terms of pseudo code. As stated above, before any rotation the molecule's geometric center must be translated to the point of origin (carried out in lines 4, 15, and 22) and moved back to the active site (lines 13 and 20) for binding pose sampling. Lines 7-10 are associated with the alignment of any other axis $i \in 2, \ldots, 6$ formed by two opposing vertices (i, i^*) to the first one $i = 1$). Orientation 1 corresponding to the initial orientation is additionally sketched in Figure 4.11. It graphically demonstrates rotations according to lines 8-10 in Algorithm 1 required to explicitly align the second axis L_2 formed by vertices 2 and 2^* to the initial one L_1 right after having sampled two times five rotations about the first x-axis and arrived at the initial position again. Basically, this transformation expressed as $L_2 \rightarrow L_1$ is ensured by aligning vertices (2,6,4^*) to (1,5,6). Sampling from any other rotation axis L_i, $i \in 3, \ldots, 6$ of the icosahedron requires only one preceding rotation of $(i - 1) \cdot 72°$ (line 7 in Algorithm 1) about the x-axis in order to have it aligned with L_2. After this $L_i \rightarrow L_2$ transformation one would perform the same steps as associated with $L_2 \rightarrow L_1$. The $L_2 \rightarrow L_1$ transformation sketched by Figure 4.11 and the determination of proper angles is described in more detail in the following.

First, according to the blue-colored lines and coordinate axis (z) in Figure 4.11, the

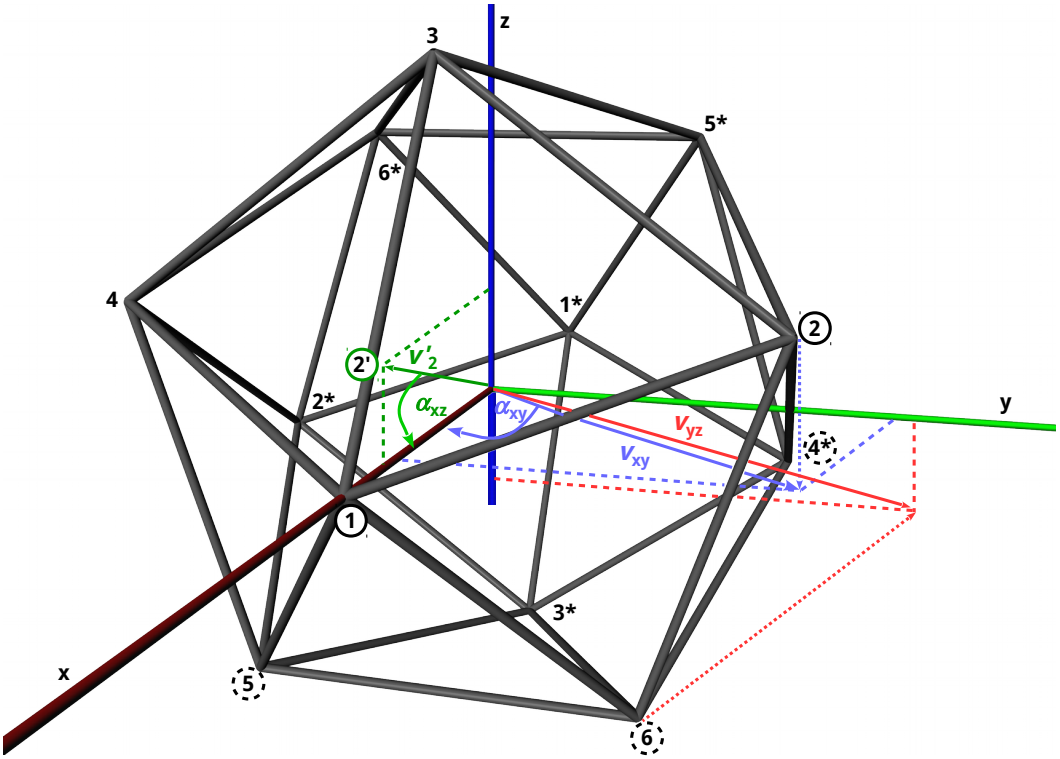


Figure 4.11: Alignment of the next rotational icosahedron axis (represented by the vertices {2-6-4*}) to the initial (x) axis (vertices {1-5-6}) using three Euler rotations about the z (blue dashed lines), y (green dashed lines), and x (red dashed lines, illustrative only) axis.

object was rotated by an angle α_{xy} about the z-axis using matrix $R_z\left(\alpha_{xy}\right)$ until vertex 2 represented by its position vector v_2 lied on the xz-plane (line 8 in Algorithm 1). The respective angle

$$\alpha_{xy} = \operatorname{acos}\frac{\langle e_x, P_{xy}\left(v_2\right)\rangle}{\|P_{xy}\left(v_2\right)\|} = -62.256^\circ$$

was determined using the normalized inner product $\langle\ ,\ \rangle$ of the unit vector $e_x = [1, 0, 0]^\top$ associated with the x-axis and the orthogonal projection $P_{xy}\left(v_2\right)$ of vertex 2 onto the xz-plane (blue fine dotted arrow v_{xy} in Figure 4.11). Any projection of a position vector on some plane spanned by two coordinate axes was carried out by setting the component associated with the third spatial axes (y in the given example) to zero. For the next transformation sketched through green-colored lines and coordinate axis, the projected image

$$v'_2 = R_z\left(\alpha_{xy}\right)\ v_2$$

of vertex v_2 (referred to as $2'$ in Figure 4.11) lying on the xz-plane was (certainly along with all other vertices of the icosahedron) rotated about the y-axis onto the x-axis using

matrix $R_y(\alpha_{xz})$. The respective angle α_{xy} was determined as

$$\alpha_{xz} = \text{acos}\frac{\langle e_x, v_2' \rangle}{\|v_2'\|} = 16.047^\circ$$

(line 9 in Algorithm 1). Now, vertex 2 (and its opposite 2*) is exactly aligned to vertex 1 (and 1*) and a last adjustment was necessary in order to align vertices 6 and 4* to vertices 5 and 6, respectively (line 10 in Algorithm 1). This was achieved by a final rotation about the x-axis using $R_x\left(\alpha_{yz}\right)$. The angle

$$\alpha_{yz} = \text{acos}\frac{\langle P_{yz}(v_6), P_{yz}\left(v_{4^*}''\right)\rangle}{\|P_{yz}(v_6)\| \, \|P_{yz}\left(v_{4^*}''\right)\|} = -26.289^\circ$$

was determined on the basis of $P_{yz}(v_6)$ and $P_{yz}\left(v_{4^*}''\right)$ representing yz-projections of both vertex 6 and the second-order image

$$v_{4^*}'' = R_y(\alpha_{xz}) \; R_z\left(\alpha_{xy}\right) \; v_{4^*}.$$

of 4*. For the sake of convenience, only the projection of vertex 6 onto the yz-plane resulting in vector v_{yz} is illustrated using red colors in Figure 4.11) whereas the projection of the second-order image of vertex 4* is neglected.

The rotations elaborated here are necessary for the alignment of vertices (2,6,4*) to (1,5,6) in order to switch (lines 7-10 of the algorithm) to the next rotation line L_i from which 2×5 modes are sampled. The first five poses are produced through five 72° rotations (for-loop in lines 12-17) about the x-axis which is currently equivalent with L_i. As indicated earlier, before the second set of five poses can be sampled from L_i (for-loop in lines 19-24), the corresponding vertices i and i^* need to be swapped through an 180° rotation about y (line 18) and revoked afterwards (line 25). After having processed any axis L_i, the icosahedron is first transformed back to its initial orientation (lines 27-30) before the next axis L_{i+1} is addressed. Preceding transformations of axes L_i with $i \in${3,4,5,6} to (the original location of) L_2 are achieved by the same 72 degree rotational steps about x as used for sampling poses from L_i.

Ensuing from Equation 4.3 any series of n successive rotations $R_i \in \mathbb{R}^3$ can be expressed by one single transformation matrix

$$R = R_n \cdots R_1$$

performing a single rotation about a rotational axis that is specified by the eigenvector associated with the eigenvalue 1. Using the tray $\text{tr}(R)$ of matrix R, the respective angle is calculated as

$$\phi = \text{acos}\left[\frac{1}{2}\left(\text{tr}(R) - 1\right)\right].$$

Algorithm 1 Algorithm creating 60 uniformly distributed binding modes

procedure ICOSAHEDRON(mol)
  $v \leftarrow$ geometric center of mol
  $T \leftarrow [\,]$ ▷ Initialize empty list of geometries
  translate mol by $-v$ ▷ Translate to origin before any rotation
  for $i = 1$ to 6 **do** ▷ Loop over six icosahedron axes
    if $i > 1$ **then** ▷ Align next axis to initial one
      rotate mol by $R_x\left((i-1)\cdot 72^\circ\right)$
      rotate mol by $R_z\left(-62.256^\circ\right)$
      rotate mol by $R_y\left(16.047^\circ\right)$
      rotate mol by $R_x\left(-26.289^\circ\right)$
    end if
    for $j = 1$ to 5 **do** ▷ Loop over five symmetric rotations per axis
      translate mol by v ▷ Translate back to original site
      push mol to T ▷ **Save current binding mode**
      translate mol by $-v$ ▷ Translate back to origin
      rotate mol by $R_x\left(72^\circ\right)$
    end for
    rotate mol by $R_z\left(180^\circ\right)$ ▷ Swap for further five rotations per axis
    for $j = 1$ to 5 **do** ▷ Loop over five symmetric rotations per axis
      translate mol by v ▷ Translate back to original site
      push mol to T ▷ **Save current binding mode**
      translate mol by $-v$ ▷ Translate back to origin
      rotate mol by $R_x\left(72^\circ\right)$
    end for
    rotate mol by $R_z\left(180^\circ\right)$ ▷ Swap back to initial orientation of current axis
    if $i > 1$ **then** ▷ Retransform back to initial orientation
      rotate mol by $R_x\left(26.289^\circ\right)$
      rotate mol by $R_y\left(-16.047^\circ\right)$
      rotate mol by $R_z\left(62.256^\circ\right)$
      rotate mol by $R_x\left(-\left(i-1\right)\cdot 72^\circ\right)$
    end if
  end for
  return T
end procedure

Accordingly, a torsional 60×60 distance matrix between any pair of two vertices was calculated and rounded after two decimals in order to get rid of machine precision issues. A 30 bin histogram (Figure 4.12) of $[0, 2\pi]$ reveals the discrete nature of all 3600 rotational distances. Distances to oneself associated with the main diagonal are represented by the first bar at distance 0. In particular, several hundred pose distances are located at small values like 63° and 72°. They ensure that for any pair of two final poses out of 60 a path possibly comprising further intermediate poses can be constructed where no intermediate distance is larger than 72°. In other words, the "true" binding mode of some ligand is substantially less than 72° away from the next representative if having decomposed its space of modes in the described manner. From that point of view, the problem is related to the meshless discretization approach presented in the previous section for which a maximum distance cutoff of 60-90° was recommended in order to have the entire available (conformational) space represented.

Application to hormone receptors

The regular rotational domain decomposition method was evaluated on the basis of several molecular systems with practical relevance. In any case, the binding modes served as initial structures for atomistic MD simulations of molecular complexes and underwent a prioritization according to physical properties. However, some results concerning the evaluation of the rotational domain decomposition constitute the basis for binding affinity estimations of host–guest systems described in the following chapters and, therefore, are presented along with those results. Nevertheless, an evaluation of the

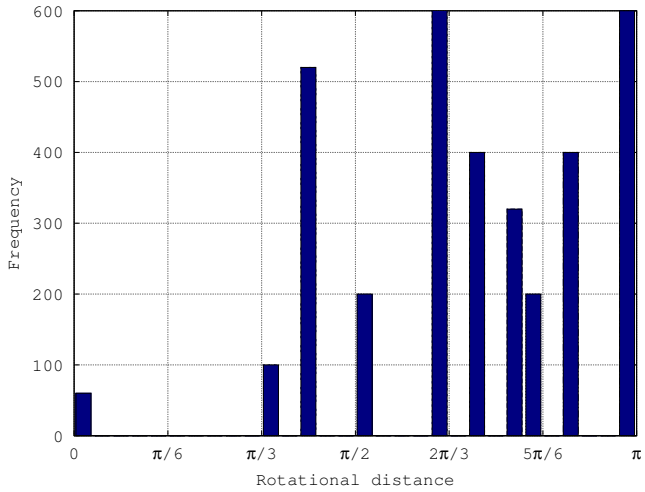


Figure 4.12: Histogram of rotational distances between 60 binding poses according to the symmetry of icosahedrons.

method in terms of a binding mode prediction is presented in this section using the prominent human estrogen receptor alpha (ERα) system by way of example.

The intracellular receptor ERα belongs to the family of nuclear hormones and is activated by the hormone 17-β-estradiol, also referred to as estrogen (E_2). After activation, the complex forms a homo or heterodimer (associated with a second type of the estrogen receptor (ER) denoted as ERβ and coded by another gene) which translocates into the nucleus. Here, it acts as transcription factor that binds to the DNA in order to regulate various different genes.[209] Thus, both types of ER are composed of several sections including a ligand binding domain, DNA-binding domain as well as a domain responsible for the dimerization. ER which plays a crucial role in cell differentiation is expressed in various tissues.[210] Its expression level and activation pattern have been shown to correlate significantly with various types of cancer and other diseases.[211–213] In the face of these results and, in addition, high estrogenic activities of several synthetic compounds, the risk of endocrine disruption by xenoestrogens has been elucidated in the early eighties already.[214] For all these reasons, this target system has been undergoing many investigations including the prediction of binding affinities. In the framework of this thesis, a predictive model for binding affinities regarding ERα was developed and will be presented in Chapter 6. Here, we will only present the prediction of a favorable binding mode which served as input for the affinity estimation. Many attempts have been made for the prediction of binding affinities to biopolymers as we will see later. For instance, on the basis of several host–guest systems, van Lipzig and co-workers achieved excellent squared coefficients of correlation around 0.9 ± 0.04 for ERα using classical MD simulations refined with respect to the number of hydroxy groups of 19 ligands. In advance, four ligand orientations had been chosen manually inspired by crystallographic data.[84] However, an automatized predictive model should abstain from any preliminary information about the ligand's orientation and, in particular, avoid any manual or random choice of some favorable pose. From this point of view, a systematic search of the preferential binding mode on the basis of a uniform decomposition seems to be a necessary approach to the estimation of binding affinities.

Data preparation and computational methods

A reasonable structure file of ERα including a co-crystallized E_2 molecule was retrieved from the PDB under the ID 1GWR[215] providing structural information about the protein's ligand binding domain (LBD) ranging from amino acid (AA) 306 to AA 549 and the nuclear receptor box (AA 742-750). Due to missing residues (AA 332-334 and AA 462-464) the structure underwent a protein building step described in more detail in

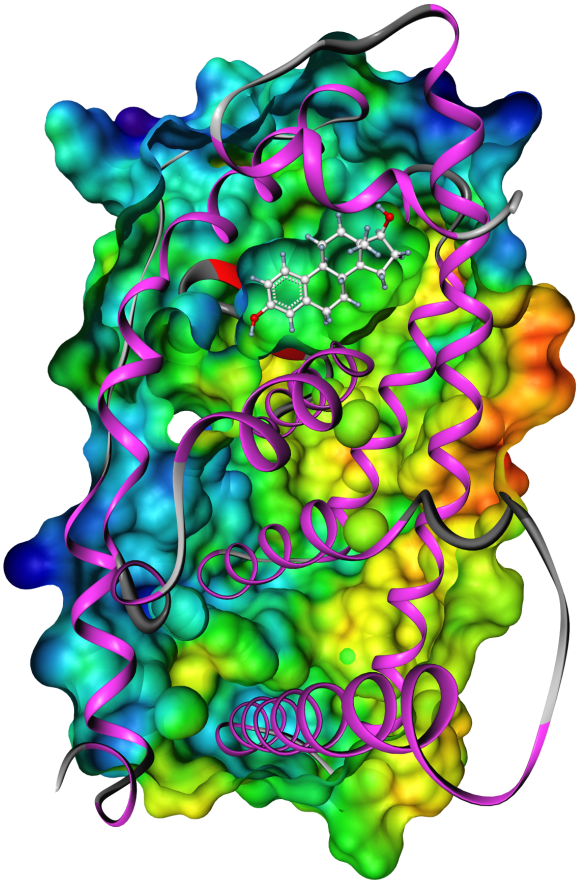


Figure 4.13: X-ray structure of the estrogen receptor α in complex with the natural binder 17-β-estradiol (white carbon scaffold) represented by its secondary structure and clipped electrostatic solvent excluding surface. Reprinted from the original publication of Durmaz et al. 2013.[194]

Section 6.2. Six different molecules could be identified in the PDB that were available as co-crystallized ligands at the LBD of ERα. For each of those ligands that are listed in Table 4.4 the global minimum was determined as described in Section 4.1. To be more precise, their structures had not been taken from the PDB but sketched, cleaned in 3D and exported to the pdb file format using the program *MarvinSketch* v5.5 and subsequently provided with MMFF parameters. These served as input for a high-temperature (1500 K) HMC sampling with a convergence check on the basis of five Markov chains consisting of 10^5 states each. Afterwards, all frames of each ligand underwent a CG minimization procedure. The lowest energy frame was chosen as an estimate for the global minimum conformation and served as initial structure for subsequent MD simulations. For the purpose of binding mode and affinity estimation, the ERα target molecule retrieved from the PDB was parametrized according to the amber99sb force field[134] which is particularly convenient for biopolymers such as proteins. In contrast,

ligand force field parameters were determined on the basis of the General Amber Force Field[123] (GAFF) using *Antechamber* from the *AmberTools* v1.4 package.[186] Charge assigment was carried out with the am1bcc method[141,153] approximating restrained electrostatic potential (RESP) charges.[151,152] Of each ligand's global minimum conformation 60 initial orientations were generated as described above and positioned at the ligand binding site of 1GWR in such a way that their geometric centers were aligned to that of the co-crystallized E_2 molecule of 1GWR. Explicit water solvation was provided by the Amber tip3p model[216] which is part of the Amber–Gromacs MD interface denoted as *amber ports*.[217] Using the *Gromacs* v.4.0.4 simulation package,[190,218,219] MD simulation was performed in three essential steps: Initially, the complex underwent 7000 steepest descent energy minimization steps if the maximum force acting on any atom had not ended up below 300 kJ/(mol nm) before. During a subsequent 200 ps equilibration phase, all but solvent atoms and ions were restrained in their positions and the pressure was coupled weakly using Berendsen's algorithm.[170] Afterwards, the entire system was simulated for 400 ps without position restraints but with constraints on all bonds according to the LINCS approach[182] allowing to set the discretized time step size to 2 fs. In accordance with human physiology, the simulation temperature was coupled to 310 K by stochastically rescaling atomic velocities.[169] Interaction energies were computed on the basis of smooth particle mesh Ewald summation[220] for Coulomb potentials with a 10 Å cutoff and a van der Waals cutoff set to 14 Å.

Optimal binding mode identification and evaluation

An estimation of preferential binding modes of ligands within the LBD of ERα was carried out on the basis of interaction enegies incorporating van der Waals and electric contributions. These energies typically provided by classical force fields are particularly suited for this purpose due to a physical quantification of pairwise atomic repulsions and attractions. Since weighted averages of these two contributions to the potential energy have been shown to significantly correlate with the binding affinity of host–guest systems[84] they were as well applied to various systems described in upcoming chapters accordingly. A considerable fraction of the time series generated by the MD production run must be regarded as an equilibration phase and consequently ignored upon average calculation because this run is the first entirely unrestrained one. It is still to be clarified, to what extent the trajectory's beginning shall be omitted. To that end, we had a closer look at the dynamics of the drug tamoxifen in complex with the protein which seems to be the most flexible and therefore, in terms of equilibration, the most demanding compound of our set of ligands. For each of its 60 MD time

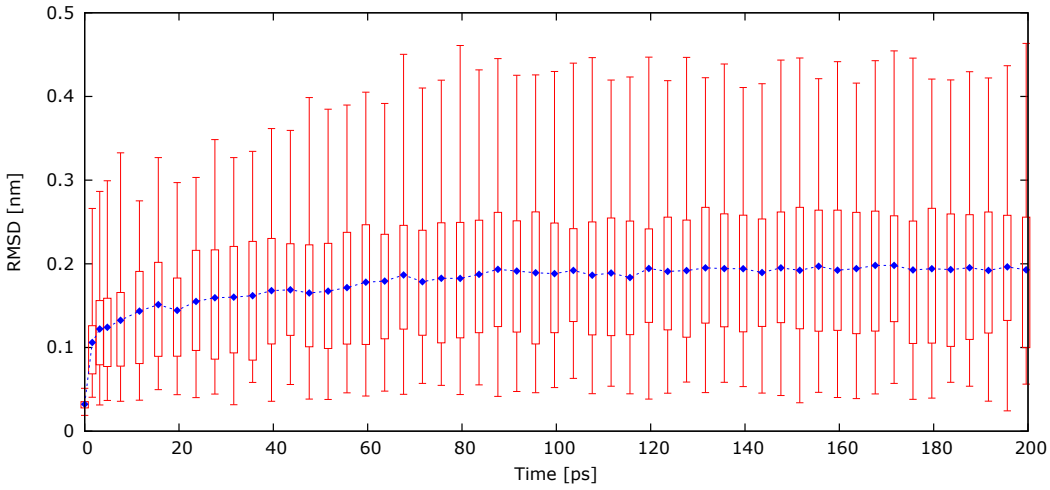


Figure 4.14: Boxplot of tamoxifen RMS deviations from initial states during MD after according least squares fit of protein backbone. Each time step is represented by the minimum/maximum as well as the first, second (median), and third quartile regarding 60 independent trajectories.

series, the protein backbone of every other MD time step was least-squares-fitted to the respective initial frame. Afterwards, we calculated root mean square (RMS) deviations from corresponding initial poses during simulation for each of the trajectories. Figure 4.14 shows a boxplot summarizing RMS deviations of these 60 poses in the course of the first 200 ps. Thus, each point in time is associated with basic statistical values of the entire set of modes including its minimum and maximum RMSD value (lower and upper whiskers), first and third quartile (lower and upper bound of sticks), and in particular its mean RMSD value (blue labels). Obviously, largest changes are related to the maximum RMSD of each point in time which can be considered converged after about 80 ps (20 %) of the MD run. For this reason, the first 80 ps were excluded from further analysis.

For the purpose of binding mode prediction, the interaction energy E_i of any time step i was constructed in a couple of different ways: first, as the unweighted sum of van der Waals E_i^{vdW} and electronic E_i^{elec} contributions,

$$E_i = E_i^{\mathrm{vdW}} + E_i^{\mathrm{elec}},$$

and in addition using each of the two energy terms solely. In combination with the estimation of binding affinities, the highest correlation has been achieved on the basis of electronic interactions only, $E_i = E_i^{\mathrm{elec}}$. An comparative evaluation is therefore presented in Chapter 6. Ensuing from this, the binding mode associated with the lowest

Table 4.4: Heavy atom root mean square deviation of predicted from crystallographic binding modes of after structural backbone alignment of respective PDB entries. Reprinted from the original publication of Durmaz et al. 2013 with slight modifications.

Compound	Receptor	PDB id	RMSD
17-β-Estradiol	ERα	1GWR	0.44
Bisphenol A	ERα	3UU7	1.52
Estriol	ERα	3Q95	0.51
Estrone	ERα	3HM1	0.58
Genistein	ERα	1X7R	1.22
(4-Hydroxy-)Tamoxifen	ERα	3ERT	2.04
Ponasterone A	EcR	1R1K	0.91

interaction energy

$$E_{\text{opt}} = \min_{j \in [1,60]} \left[\frac{1}{N - 200} \sum_{i=200}^{N} E_i \left(q_{i,j} \right) \right] \tag{4.8}$$

averaged over $N = 1000$ time frames deducting the initial 80 ps (200 frames) was chosen as the favorable (most likely) one. As illustrated by Table 4.4, the predictive model was evaluated by comparison with crystallographic structures from the PDB. RMS deviations less than 1.5 Å are considered acceptable.[86] This condition was met by all ligands except of the particularly flexible molecule tamoxifen having yielded a negligibly higher deviation. In particular, the natural binders E_2, estriol, and estrone yielded excellent RMSD values followed by the insect metamorphosis-regulating steroid hormone ponasterone A. The latter's binding mode was predicted against the ecdysone

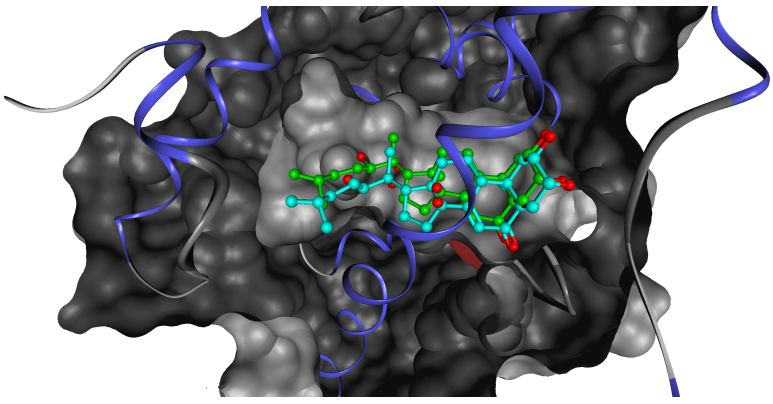
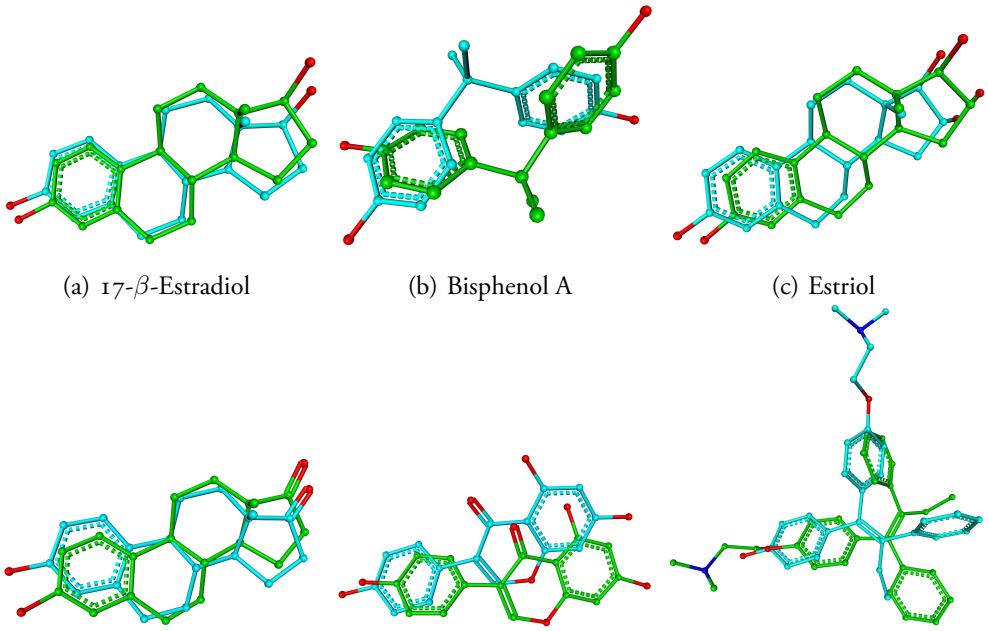


Figure 4.15: Crystallographic tertiary structure of the ecdysone receptor originating from Heliothis virescens (PDB id 1R1K) and including one co-crystallized ponasterone A molecule (light blue carbon scaffold) as well as its optimal binding mode predicted using MD simulations (green).



(a) 17-β-Estradiol (b) Bisphenol A (c) Estriol

(d) Estrone (e) Genistein (f) Tamoxifen

Figure 4.16: Alignment of predicted binding modes (green carbon scaffold) to respective co-crystallized (xeno)estrogens retrieved from the Protein Data Bank. Oxygens and nitrogens are colored red and dark blue, respectively. Reprinted from the original publication of Durmaz et al. 2013.

hormone receptor (EcR) originating from the moth breed Heliothis virescens (PDB id 1R1K). In analogy to ERα, the ecdysone-inducible mammalian expression system is responsible for cell differentiation and metamorphosis.[221] However, a visual inspection allows more reliable judgement of the achieved results (see Figure 4.16 and 4.15). Before RMSD computation, the protein backbones of each simulation's last time frame (green carbon scaffold in both figures) had been aligned to protein backbones of respective crystallized complexes (light blue scaffold). For all steroids as well as the phytoestrogen genistein and ponasterone A, the native pose was successfully estimated as the preferential one. Due to the flexibility of ERα atoms during MD, slight deviations of these ligands at the binding site have been expected, whereas bisphenol A and, in particular, tamoxifen show considerable deviations from the natural binding mode. Note, that in contrast to our simulations, the PDB entry 3ERT contains the 4-hydroxylated form of tamoxifen. For this reason, the comparison of prediction and crystal data has a limited validity. However, chemical properties and groups such as benzene or hydroxy groups match well corresponding crystallographic groups. In the face of these results, it seems likely that the prediction of compounds with increasing number of rotational degrees of freedom will tend to produce wrong binding modes. Nevertheless, binding pose estimates according to the presented strategy constitute a solid basis for host–guest binding affinity estimations described in the next chapters.

4.4 Concluding remarks

This chapter entirely addressed the decomposition of the conformational space which is exceedingly complex for most molecular systems and, in particular, in case of explicitly solvated macromolecules including biological host–guest systems. Already the state space of small drug-sized molecules with less than 100 atoms generates great computational effort when it comes to the determination of a globally geometry-optimized initial structure for conformational analysis or further investigations including binding affinity estimation. In order to remedy the trapping problem associated with both deterministic MD simulations as well as random Monte Carlo samplings, and calculate the global energy minimum of such small molecules, we have elaborated a simple two-stage strategy on the basis of the HMC method. After an extraordinarily high-temperature HMC sampling a gradient-based local minimization routine was applied to each geometry. The conformer associated with the lowest energy minimum was selected as global minimum. Conformational analysis of the highly symmetric system of six major HBCD stereoisomers exhibits the correct distribution and, thus, indicates a sufficient spacious sampling from the "entire" space of HBCD using five separate 10^5 step Markov chains. According to the convergence criterion proposed by Gelman and Rubin, convergence had already been achieved using only $5 \cdot 10^4$ steps per chain whereas 10^4 steps per chain did not result in a satisfying convergence as evaluated on the basis of HBCD symmetry properties. A structural comparison of the three major HBCD diastereomers with crystallographic data indicates excellent predictions for α and β-HBCD and a less satisfying estimation in case of γ-HBCD. This visual observation is by far better reflected by torsional instead of Cartesian RMS deviations. In the light of the large number of rotational degrees of freedom and high energetic barriers due to the strained cyclic structure of HBCD, the convergence and, in particular, the prediction of globally minimum energy geometries by way of the presented method must be considered as remarkable.

Sometimes, one is interested only in a small subset of a large sampling representing the conformational space as largely as possible. Spacious representations are, for instance, useful for investigations of molecular kinetics of conformational changes and binding events that are often described by several (intermediate) states including unfavorable conformers. We have presented the *maxdist* algorithm for the selection of k representatives out of an n-frames sampling such that the Euclidean distance of each frame to the nearest cluster representative based on internal (torsional) coordinates is below a particular limit $d_{\max}$. Due to an expectedly high number of samplings and since usually $k \ll n$, the quasi-linear algorithm's time complexity must be categorized in the

order of $\mathcal{O}(n)$. The choice of $d_{\max}$ is up to the user and recommended to be chosen significantly less than 2π divided by the multiplicity of a typical torsion angle associated with a rotatable bond between two sp3 hybridized carbon atoms, i. e. less than 120°. Thus, reasonable values for $d_{\max}$ might be 90° or even 60°. The latter value substantially increases k which is well approximated by a hyperbolic function of $d_{\max}$. According to the mathematical model, the number of torsional degrees of freedom defines the exponent of $d_{\max}$. Using a 500 k step HMC sampling of HBCD as a case example and k values calculated on the basis of 29 different values for $d_{\max}$ ranging from 0.0005π to 2π, the coefficient of Pearson correlation with the mathematical model amounts to excellent values larger than 0.984 regarding one to five degrees of freedom. A higher correlation in particular regarding increasing degrees of freedom and small values for $d_{\max}$ is achieved if introducing a correction factor into the exponent. This quantity was interpreted as a measure for the sampling's deviance from an ideal (uniform) distribution due to structural restraints. As an alternative termination criterion, the straight forward algorithm allows to set a certain number of cluster representatives instead of the maximum allowed distance to the nearest one.

Finally, an as systematic as simple strategy to a rotational space discretization has been presented that was used for the prediction of guest binding modes at ligand binding sites of target molecules. Such systematic approaches are necessary in order to remedy issues related to trapping effects during MD simulations. According to the symmetry properties of the highest order Platonic solid, namely the icosahedron, the globally minimum energy conformation of several ligand molecules were placed at the LBS of ERα and, respectively, EcR in 60 uniformly distributed orientations. This resolution guarantees, that the rotational distance of the "true" (or most likely) binding mode to the nearest one of the 60 given modes is less than 72 degrees. Each of these complexes served as an initial structure for an 400 ps MD simulation of which the first 80 ps were omitted as they had to be considered as an unrestrained relaxation and equilibration phase according to the ligand RMS deviation. On the basis of the remaining frames, time averages of ligand interaction energies with its surroundings have been calculated in order to predict a preferential binding mode. Electric interactions represented by pairwise Coulomb potentials yielded results superior to van der Waals interactions as modeled by a Lennard-Jones potential. An evaluation was possible by comparing predicted binding modes with crystallographic data of several receptor–ligand complexes available at PDB. RMS deviations from these reference structures amounted to promising values below 1.5 Å. Further evaluation of the described method is carried out in upcoming chapters where those preferential binding modes are used for a quantification of binding strength regarding various host–guest systems.

5 Modeling chromatographic separation of HBCD stereoisomers

The molecular modeling techniques and results presented in this chapter have to a notable extent been published in the following article.[164] A republication of related contents in the framework of this thesis was kindly permitted by its publisher:

- V. Durmaz, M. Weber, R. Becker: How to Simulate Affinities for Host-Guest Systems Lacking Binding Mode Information: Application in the Liquid Chromatographic Separation of Hexabromocyclododecane Stereoisomers. *Journal of Molecular Modeling*, 18(6):2399–2408, 2012.

Within the frame of this thesis, this is the first chapter that brings together two major parts elaborated during previous chapters: the uniform decomposition of conformational host–guest space for an estimation of preferential binding modes on one side and, on the other, the prediction of corresponding binding affinities by developing a suitable empirical linear model related to the LIE method. Albeit, the molecular system under investigation is not about biological host–guest systems but concerning the interaction of small chemical compounds with a material used by a method called *High-performance liquid chromatography* (HPLC) for the separation of aqueous compound mixtures. The major aim of this chapter was the development of an automated as well as robust method to the prediction of HPLC results at a minimum number of manual operations and decisions.

5.1 Introduction

HPLC is an analytical (and preparative) technique in chemistry that is used for the separation of a mixture of compounds. Using reference chemicals, it is possible to determine the structure as well as quantity of probes. In few words, a liquid mixture (*mobile phase*) of the compounds under observation flows through a column filled with some sorbent material (*stationary phase*). Due to different degrees of interaction with

the stationary phase, substances dissolved in the mobile phase reveal differing flow rates and, therefore, leave the column (*elute* from the column) after different *retention times*. The retention time and in particular the elution order of compounds is mainly determined by the choice of the two phases. Depending on the combination of those phases, the separation of substances can be due to various physicochemical properties such as their size, polarity, and formal charge.[222] A computational method for the estimation of the elution order must take several aspects into account. The compounds' conformational flexibility has to be considered since their binding affinity to the stationary phase or host molecule strongly depends on conformational changes. As we have already pointed out in Chapter 2, classical MD simulations on the basis of the two reaction end states (bound and unbound) provide an acceptable tradeoff between accuracy and computational expense. From a physical point of view, the chromatographic elution order of analytes depends on the strength of their interaction with both stationary and mobile phase and can be interpreted as a binding affinity. Further, we have seen that in terms of thermodynamics the binding affinity is related to the concentration ratio of two states of a host–guest system at thermal equilibrium: the bound and the unbound state. It can be derived from the free energy difference of these two states comprising enthalpic as well as entropic contributions. Classical MD simulations provide estimates for both of them. In the last years, a number of different methods for the calculation of the elution order have been developed. A mathematical model for estimating enantiomeric resolutions from molecular mechanics simulations of chiral separations was developed by Zhang et al.[223] Issaraseriruk et al.[224] derived binding free energies for enantiomeric separation with a combination of molecular docking using AutoDock[96] and semi-empirical Parametric Model 3 (PM3)[225] calculations where the latter method had substantially more discriminating power than AutoDock energies.[224] Pérez-Garrido et al. achieved excellent correlations and cross-validation values with a quantitative structure-activity relationship (QSAR) model used to predict complexation of a series of organic molecules with β-cyclodextrin (β-CD).[226] Of course, MD simulations as well have been employed in order to describe host–guest interactions and separation phenomena.[227] However, these models suffer either from a manual choice of (initial) binding mode or from the lack of explicit solvent molecules. That is, in case of the latter issue, many strategies for the simulation of elution orders concentrate on modeling the interaction between the compound and the stationary phase, though, neglecting explicit interactions with the mobile phase. This also holds for MD-based investigations of chromatographic separation systems either modeled in gas phase only[228] or incorporating an implicit solvent as published few years ago.[227] However, apart from the crude simplification of solvent effects through implicit solvent, many

solvents have not been modeled implicitly so far.

In this chapter, different aspects related to explicit solvent simulations of interactions between chemicals and HPLC stationary phases are discussed. On the one hand, the suitability of classical force fields and solvent models will be evaluated regarding the estimation of a chromatographic elution order. If they are applicable, the simulated data is supposed to comprise all necessary information describing the host–guest interaction. To that effect, only physically meaningful force field terms will be extracted from the data in accordance with thermodynamic principles. In the course of comparing several descriptors for the retention behaviour, we will in particular compare energies averaged over certain MD time ranges with single-step energies as known from ordinary molecular docking. Besides and in contrast to the mainstream trend, a high value is set on consistency in the observed behaviour. Appropriate correlations of simulated with experimental results are supposed to be robust regarding the time range under consideration, especially since pretending to simulate molecular systems at chemical equilibrium. The procedure is illustrated using by way of example the separation of HBCD stereoisomers on a chiral stationary phase. HBCD seems well suitable for the predictive approach outlined in the following because it displays a complex cohort of diastereomers and enantiomers and is therefore regarded as ideal starting point re-

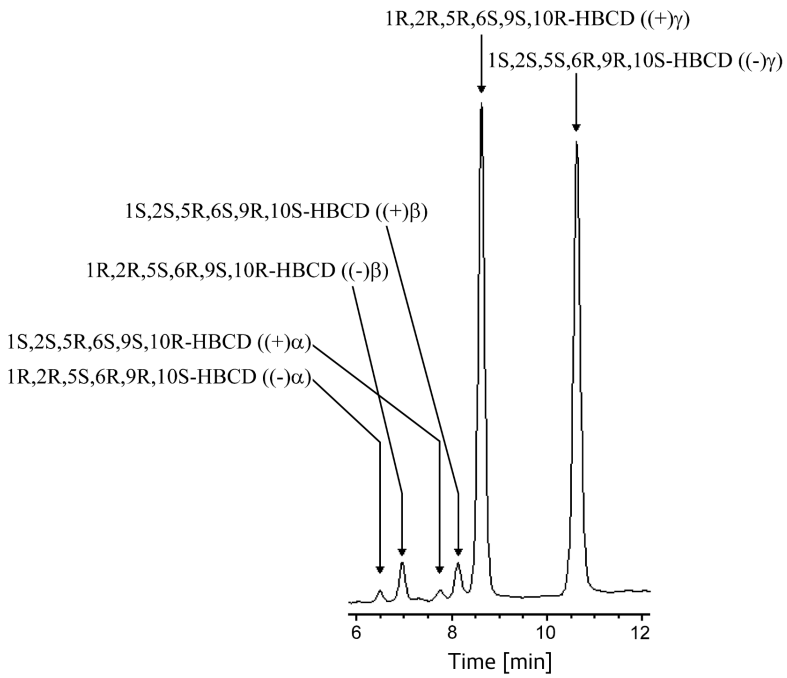


Figure 5.1: Separation profile of the six major HBCD stereoisomers on a chiral β-pmCD column. This figure was taken from [229] by courtesy of the main author. Reprinted from the original publication of Durmaz et al. 2012.

garding computational challenge and practical significance.[164] In addition, the entire molecular system is substantially smaller than typical biopolymers such as enzymes and receptors resulting in much less computational effort.

A separation of the six major HBCD stereoisomers (Figure 4.2) by HPLC was accomplished in the year 2007 on the basis of the technical mixture. The analytical challenge was conclusively described with the assignment of the absolute configurations of the enantiomers by Köppen et al.[229] In short, an analytical column packed with permethylated β-cyclodextrin (β-pmCD) as an unpolar stationary phase (see Figure 5.2) was combined with a polar solvent gradient consisting of water and acetonitrile (ACN). The separation of enatiomeric pairs on that stationary phase is possible due to the chiral nature of β-pmCD. Because of the hydrophobic character of HBCD, its interaction with β-pmCD and, as a consequence, separation increases in contact with water. In contrast, the less polar co-eluent ACN reduces host–guest interactions and, thus, enhances HBCD elution. Figure 5.1 shows the corresponding chromatogram with retention times.[164]

5.2 Data preparation and computational methods

The β-pmCD crystal structure was retrieved from the Cambridge Structural Database (CSD)[230] under the id COYXET20.[231] In analogy to HBCD stereoisomers, it was parametrized according to the generalized AMBER force field (GAFF) using Antechamber from the AmberTools v1.4 package. Prior to the determination of globally minimum geometries of HBCD in accordance with the procedure described in Section 4.1. charges were assigned using the am1bcc method. As described in the previous section and illustrated by Figure 5.2, each of the 60 orientations of any HBCD stereoisomer was placed inside the β-pmCD cavity such that host and guest were centered geometrically. In addition, each system was solvated explicitly in two different ways: once in pure water as provided by the Amber tip3p model[217] and once again in pure acetonitrile according to a model developed by Nikitin et. al.[232] Starting with these initial complex structures and using Gromacs v.4.0.4, all systems underwent 5000 steepest descent energy minimization steps unless the maximum force acting on any atom and in any spatial direction had ended up below 100 kJ mol^{-1} nm^{-1} before. During a subsequent 400 ps canonical equilibration phase, the positions of all but solvent atoms were restrained. Afterwards, the whole system was simulated for at least 400 ps without position restraints but with constraints on all bonds according to the LINCS approach and allowing to set the discretized MD step size to 2 fs. In accordance with HPLC condi-

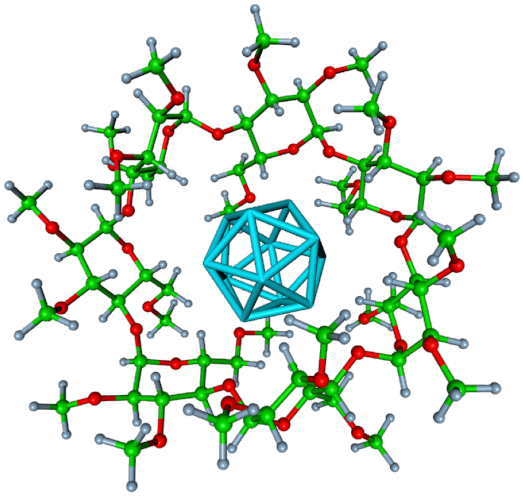


Figure 5.2: Permethylated β-cyclodextrin stationary phase used for HPLC separation of HBCD stereoisomers. The icosahedron in the β-pmCD cavity represents 60 uniformly distributed rotational binding modes of HBCD providing starting conformations for MD simulations. Reprinted from the original publication of Durmaz et al. 2012.

tions, the temperature of the NpT ensemble was coupled to 310 °C by stochastically rescaling atomic velocities, and the pressure was set to 25 bar.[229] Interaction energies were computed using the Gromacs setting `PME-Switch` on the basis of the smooth partical mesh Ewald summation[233] for Coulomb potentials with a cutoff at 11 Å and the `shift` setting for van der Waals interactions within a dual range switched after 9 Å and a cut off at 10 Å.

5.3 Optimal binding mode analysis

In analogy to the host–guest binding mode estimation in Section 4.3, the question arose which part of the MD trajectories to use in the course of further interaction analysis. To answer this question, center of mass distances between each HBCD stereoisomer and β-pmCD during the MD simulations were calculated and smoothed using Matlab's `filtfilt()` function along with the Octave software. A distance trace plot of both solvents, water (left plot) and ACN (right plot), is shown in Figure 5.3. HBCD orientations shown here were approved as predominant for reasons specified below. Due to high repulsive forces within the β-pmCD cavity caused by unfavorable atomic collisions at the beginning, host–guest distances of all complex MD runs show a steep initial increase until some equilibrium is reached after approximately 50 ps. On that account, the first 80 ps of the time series were omitted again when it came to the computation

of average interaction energies. After 80 ps, host–guest distances behave substantially differently regarding the two solvents. In case of water, comparatively little fluctuations are observed such that the intermolecular distance of each β-pmCD–isomer complex stays within a region of 1 Å during the remaining trajectory. In contrast, MD runs of HBCD solvated in ACN exhibit more inconsistent distances showing large jumps of up to 1.5 Å within few tens of picoseconds and even beyond 0.5 ns. Moreover, these large fluctuations are mostly directed from the host cavity outward, towards the solvent ACN. This observation might be caused by the significantly larger size (molecular weight) and smaller molar density compared to water. As a consequence, ACN dampens HBCD collisions less uniformly than water which is able to form a smoother barrier.

Center of mass distances between HBCD and β-pmCD averaged over time, initial binding modes and stereoisomers after 400 ps were significantly larger in ACN amounting to 8.0 Å than in water with 5.2 Å. This confirms on the one hand the hydrophobic character of HBCD that would rather prefer to reside in the less polar solvent ACN and, on the other hand, experimental observations implying that chromatographic separation (interaction with the stationary phase) is superior in water whereas solubility and, therefore, HBCD elution rate is advanced by more hydrophobic solvents such as ACN. These results are supported by interaction energies of HBCD with its surrounding derived from MD simulations incorporating explicitly solvated HBCD stereoisomers without the stationary phase β-pmCD. With $-170.7\,\mathrm{kJ\,mol^{-1}}$, ACN yielded substantially lower interaction energies than water which amounted to $-148.5\,\mathrm{kJ\,mol^{-1}}$. According to this observation, HBCD prefers to reside in ACN rather than in water which

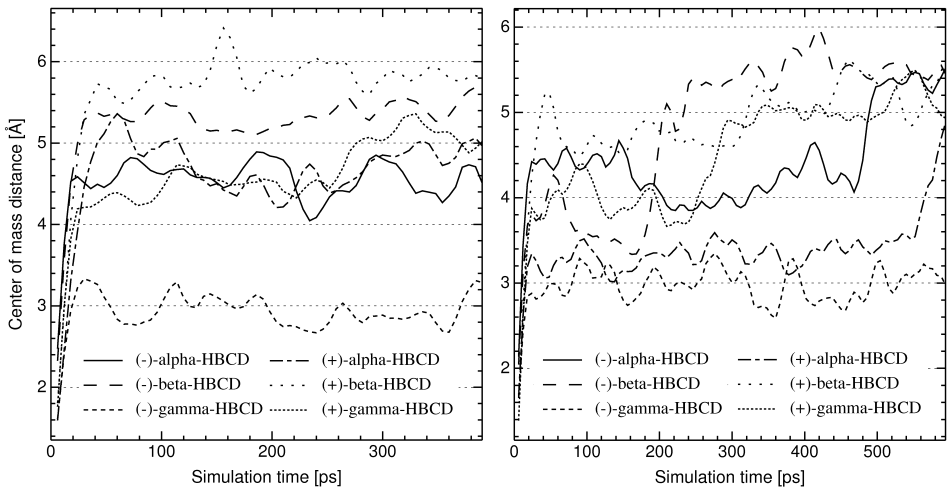


Figure 5.3: Center of mass distances between HBCD stereoisomers and β-pmCD in water (left) and acetonitrile (right) during MD simulations. Reprinted from the original publication of Durmaz et al. 2012.

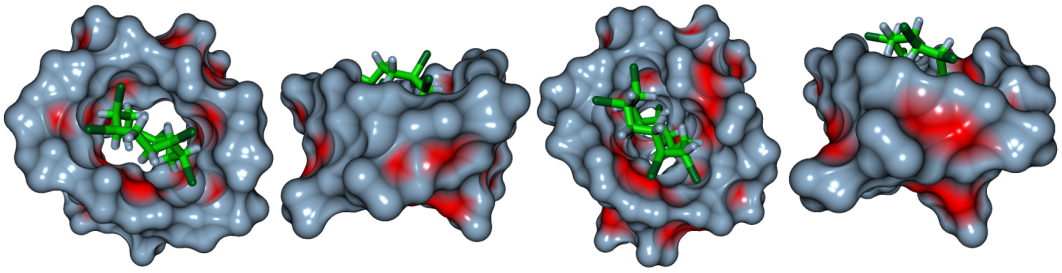


Figure 5.4: Low energy conformations of $(-)$-γ-HBCD within the β-pmCD (represented by its solvent-excluded surface) cavity with a mass center distance of 3 Å on the left-hand and of $(+)$-β-HBCD with a distance of 5.8 Å on the right-hand fetched from the equilibrium region of an MD simulation with explicit water. Reprinted from the original publication of Durmaz et al. 2012.

clearly copes with experimental results originating from HPLC separation.[229] Typical low energy modes of $(-)$ -γ-HBCD (smallest distance) and $(+)$ -β-HBCD (largest distance) isomers within β-pmCD simulated in water are depicted in Figure 5.4. At equilibrium, about half of each isomer's surface is encompassed by C_2- and C_3-methoxy moieties of β-pmCD glucopyranose units. Despite that, HBCD interactions concern the entire cyclodextrin molecule since nearly all intermolecular carbon distances are below the smallest chosen cutoff distance that is 10 Å in case of van der Waals forces.

An exhaustive sampling of the conformational space for the sake of computation of binding free energy differences associated with host–guest complexes at chemical equilibrium is generally hindered by high energetic barriers forming an extremely complex energy landscape. This issue was to a large extent remedied by decomposing the space of binding modes for 60 independent simulations. The major question arising from this strategy concerns the derivation of a single value quantifying the host–guest interaction probably along with the identification of a preferential binding mode. Several solutions to this question come into consideration and have been compared on the basis of force field energies E between HBCD and β-pmCD. A first reasonable approach that comes into mind is related to the (single) lowest energy state

$$E_{\text{min}} = \min_{i \in [1,N]} \left(\min_{j \in [1,n]} E_i \left(q_{ij} \right) \right) \tag{5.1}$$

of $N = 60$ MD time series per isomer each consisting of $n = 160\,\text{k}$ states with coordinates q (after having omitted the first 40,000 steps). The idea behind Equation 5.1 is comparable with that of molecular docking approaches as these typically quantify binding affinities on the basis of as well one preferential binding pose using some scoring function. Alternatively, a favorable binding pose can be determined based on the lowest

of 60 *time-averaged* energies

$$E_{\text{mean}} = \min_{i \in [1,N]} \left(\frac{1}{n} \sum_{j=1}^{n} E_i \left(q_{ij} \right) \right). \tag{5.2}$$

Obviously, Equation 5.2 better copes with thermodynamic principles since it clearly reflects the inner energy as an average of a thermodynamic ensemble comparable with *in-vitro* and *in-vivo* measurements. From this point of view, an even more rigorous approach would take into account the average not only of the most preferential binding mode but of all of them for the quantification of host–guest interactions. Consequently, this third ansatz was accomplished by summing up the 60 modes' time-averaged energies weighted according to the Boltzmann probability distribution for canonical ensembles and yielding

$$E_{\text{prob}} = -\frac{1}{\beta} \ln \left(\frac{1}{N} \sum_{i=1}^{N} \exp \left(-\frac{\beta}{n} \sum_{j=1}^{n} E_i \left(q_{ij} \right) \right) \right). \tag{5.3}$$

$\beta = [RT]^{-1}$ represents the inverted product of the gas constant, $R = 8.3145\,\text{J}\,\text{mol}^{-1}\,\text{K}^{-1}$ and temperature $T = 310\,\text{K}$. Since this strategy does not directly identify a preferential complex geometry, the binding mode was predicted in accordance with the highest statistical weight. As described in Chapter 2, this is equivalent with choosing the representative by means of the lowest average energy of 60 trajectories just like in case of Equation 5.2. Consequently, both Equations 5.2 and 5.3 propose the same preferential binding mode.

5.4 Validation of physical descriptors

Besides defining strategies for selecting representative (favorable) orientations, those energy terms needed to be determined that considerably contribute to the correlation between the simulated and the experimental elution order. In a rigorously thermodynamic manner, one would rather use the total inner energy of a molecular system for the computation of free energy differences. However, due to the high number of water molecules, the inner energy reveals high fluctuations among which the small contributions of intermolecular interactions between the two core molecules would get lost. For this reason, the inner energy is neither expected to be reproducible nor to correlate well with the experimental elution order. A more convenient choice that better reflects the interaction of interest related to β-pmCD and HBCD is concerned with force field terms describing nonbonded molecular interactions. Typically, these are electronic and

van der Waals forces modeled by classical force fields in terms of the Coulomb and Lennard-Jones potential, respectively. In order to figure out which one or composition of these two contributions correlates best with the elution order of HBCD isomers, the individual values of these terms as well as their sum were fed to Equations 5.1-5.3. To be precise, respective energy differences between the bound (including β-pmCD) and unbound system (excluding β-pmCD) were assigned to $E\left(q_{ij}\right)$ in order to compute the enthalpic part of the free energy difference. This physically reasonable approach of deriving free energy differences only on the basis of interaction energy terms has already been successfully applied to the estimation of host–guest binding affinities.[234] According to thermodynamic studies on chromatographic retention behaviour, several thermodynamic quantities including the inverse temperature, enthalpy, and in particular free energy scale linearly with the natural logarithm of HPLC capacity factors[235,236]

$$k_\xi = \frac{t_\xi - t_0}{t_0}. \tag{5.4}$$

Thus, for evaluation purposes, k_ξ was calculated for each stereoisomer ξ on the basis of its retention time t_ξ according to the chromatogram in Figure 5.1. The chromatographic dead time t_0 reflects technical and physical properties of the involved HPLC system and buffers.[229] Table 5.1 shows averaged squared coefficients for the running correlation of the natural logarithm $\ln\left(k_\xi\right)$ of experimental capacity factors with various interaction energy compositions computed in accordance with the three scoring strategies described above (Equations 5.1-5.3). In order to figure out the physical model's robustness, these coefficients were averaged over multiple time ranges starting at succeeding time frames with 20 ps offsets and always ending at 360 ps. In addition, we distinguished between both solvents and the discriminating power for both the entire set of six stereoisomers (Iso.) as well as among each pair of enantiomers (Ena.) only. The correlation coefficient associated with the enantiomer-specific separation was calculated as $R^2\,(\text{Ena.}) = \frac{1}{3}\left(R_\alpha^2 + R_\beta^2 + R_\gamma^2\right)$ where subscripts address HBCD diastereomers.[164]

Taking the sum of both nonbonded energy terms, Coulomb and Lennard-Jones potential representing electronic and, respectively, van der Waals interactions clearly provided the best overall separation performance (correlation). This particularly holds for systems solvated with water for which all models (Equations 5.1-5.3) yielded excellent R^2 values between 0.8 and 0.87 in case of the entire set of stereoisomers (Iso.). However, with squared coefficients around 0.86 and 0.87 both models based on statistical averages in turn performed notably better than the one relying on a single isomer state (evaluated as 0.8). Regarding only single pairs of enantiomers (Ena.), all models achieved $R^2 = 1$, that is, the elution order among any pair of enantiomers was

correctly predicted using the sum of both energy contributions. Considering only one of these energy terms yielded substantially lower values for simulations in water. For ACN as solvent, in contrast, significantly less correlation with experimental observations was encountered. The largest Pearson coefficients having achieved correlations of 0.73 (Iso.) and 1 (Ena.) are associated with the minimum energy state out of the pool of all 60 trajectories (E_{min} according to Equation 5.1).[164] However, since we are particularly interested in the prediction of the correct elution order, a correlation of $R^2 = 0.73$ seems unsatisfactory for a predictive model as will be discussed below. All in all, the solvent influence agrees very well with experimental HPLC observations since better HBCD separation is indeed achieved with dominant water concentrations of the eluent whereas elevated ACN concentrations rather reduce separation efficiency and enhance elution of HBCD isomers.[229] For water systems and using the sum of both interaction energy terms, the quality of all three models is showcased in Figure 5.5. In order to get an impression of the models' consistency, correlation coefficients have been plotted against different ranges of the trajectories starting with 20 ps offsets and collectively ending frame $t = 400$ ps. The left diagram reveals consistently high squared correlation coefficients at about 0.8 associated with the minimum energy state (E_{min}) or even better with about 0.85 using one of the average-based equations (E_{mean} or E_{prob}). By all models, enantiomer-specific separation (right diagram) in water was, nearly over the full trajectory range, correctly estimated. On top of this, comparing the two latter, average-based models indicates a satisfactory approximation of the Boltzmann-weighted sum of all orientations (corresponding to Equation 5.3) by the preferential orientation

Table 5.1: Squared Pearson coefficients R^2 of three approaches (columns) for the running correlation of logarithmized experimental capacity factors with HBCD interaction energies regarding its chemical environment. It was distinguished between two solvents and the discriminating power for all isomers (Iso.) and for enantiomers (Ena.) only. All coefficients were averaged over successive 20 ps MD time ranges. Reprinted from the original publication of Durmaz et al. 2012 after modifications.[164]

		$\langle R^2\,(E_{\text{mean}})\rangle$		$\langle R^2\,(E_{\text{min}})\rangle$		$\langle R^2\,(E_{\text{prob}})\rangle$	
Force field potential	Solvent	Iso.	Ena.	Iso.	Ena.	Iso.	Ena.
Coulomb	ACN	0.05	0.11	0.69	0.11	0.38	0.11
	Water	0.09	0.11	0.71	0.11	0.45	0.11
Lennard Jones	ACN	0.66	0.11	0.72	1.00	0.70	1.00
	Water	0.52	0.11	0.73	1.00	0.58	0.11
Coulomb. & L.-J.	ACN	0.64	0.31	0.63	0.11	0.67	0.11
	Water	0.86	1.00	0.80	1.00	0.87	1.00

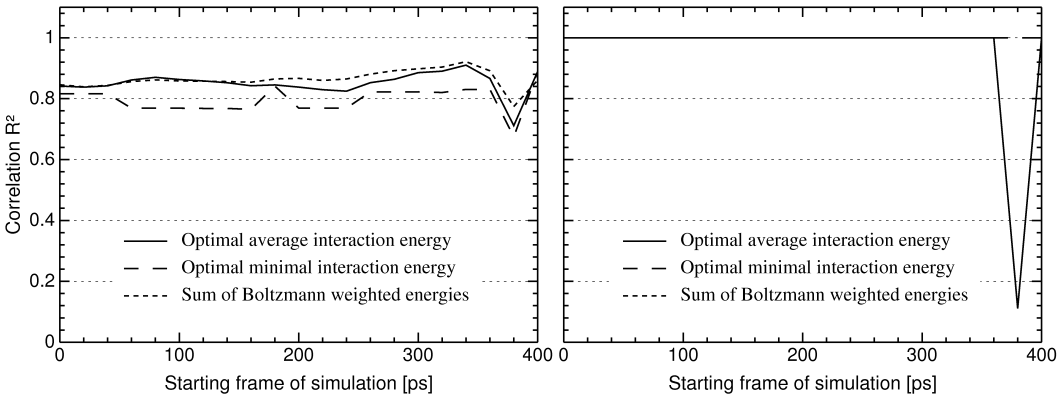


Figure 5.5: Running plot of squared coefficients for the correlation of HBCD HPLC capacity factors with interaction energies from explicit water simulations depending on the running starting frame with 20 offsets. The HPLC elution order estimation for all stereoisomers (left) and, respectively, three pairs of enantiomers only (right) was carried out using three models. Reprinted from the original publication of Durmaz et al. 2012.

associated with the highest statistical weight (Equation 5.2). Apart from distortions within the last 50 ps, these results clearly approve the robustness of the ensemble-based multi-mode approach over a wide time range – at least for this hydrophobic class of compounds separated on a stationary phase associated with molecules that exhibit cavities large enough for small compounds. The system's thermodynamic equilibrium is well reflected by coefficients of correlation which have a consistently high value independent from the time range under consideration. Interestingly, according to another study[237] on chiral separation using a β-cyclodextrin column, 20 ns were necessary in order to obtain the correct order of two enantiomers, though, the implicit solvation simulations started with 15 Å distance between host and guest rather than a guest directly nested in the host molecule. It should be noted, that indeed the likeliness for randomly choosing the pairwisely correct order among three independent pairs (separation of HBCD enantiomers) is $p = 0.5^3 = 0.125$, whereas guessing the correct order of six stereoisomers at once is about 100 times more unlikely with $p = (6!)^{-1} = 0.0014$.[164]

5.5 Computation of the HPLC elution order

A direct quantitative translation of force field energies into chromatographic retention times is hard to realize since these depend on many physical and process parameters including flow rate, temperature, pressure as well as the length, density and diameter of the column. However, deriving a relative order of elution directly from host–guest interaction energies is possible and provides the information needed by the analyst. Moreover,

in the light of a linear dependency of capacity factors on energy/enthalpy[235,236] it seems practicable to implement an empirical linear model on the basis of training data with known capacity factors (retention times). Such a parametrized linear equation allows to predict retention times for unknown substances separated under the same conditions as the compounds building the training set. Consequently, the linear model was constructed for the most promising system that is the one solvated in water and evaluated according to Equation 5.2 addressing the binding mode associated with the highest statistical weight. As justified by the mass center distances shown in Figure 5.3, all frames within the time range from 80 ps through 400 ps of that binding mode were taking as data basis for the development of a predictive model. First, we constructed a system of $n = 6$ linear equations

$$\underbrace{\begin{pmatrix} \ln(\mathrm{k}_1) \\ \vdots \\ \ln(\mathrm{k}_n) \end{pmatrix}}_{y} \approx \underbrace{\begin{pmatrix} E_{\mathrm{mean},1} & 1 \\ \vdots & \vdots \\ E_{\mathrm{mean,n}} & 1 \end{pmatrix}}_{A} \underbrace{\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}}_{x} \tag{5.5}$$

associated with the six major HBCD stereoisomers and comprising the sum of both interaction energy terms as well as a constant as parameters in matrix A. Ensuing from that, weights x were fitted using the *least-squares* method

$$x = \left(A^{\mathrm{T}}A\right)^{-1} A^{\mathrm{T}}y \tag{5.6}$$

such that the squared deviation of simulated (Ax) from experimental capacity factors (y) became minimal

$$\min_x \|y - Ax\|_2^2 \,. \tag{5.7}$$

In other words, optimal interaction energies were scaled and shifted in order to minimize the deviation. We will briefly discuss the suitability of the least-squares method to this type of matrices in Chapter 6. In any case, we obtained $-0.015\,\mathrm{mol\,kJ^{-1}}$ and -0.661 for x_1 and x_2. For the examined time range (80-400 ps) and using these weights, squared coefficients of the Pearson correlation between y and Ax amounted to 0.86 and 0.85 considering the optimal average and, respectively, weighted sum approach which are considerably higher than 0.77 achieved by the single step approach (see Table 5.2). By the way, these squared coefficients would have increased to 0.92, 0.91, and 0.82, respectively if having had fitted the energies to $k_x i$ instead of its logarithm which makes no sense physically. The parameterized linear equation

$$\ln(k_\xi) = -0.015 \frac{\mathrm{mol}}{\mathrm{kJ}} E_{\mathrm{mean},\xi} - 0.661 \tag{5.8}$$

is exemplified using energies E_{mean} related to the optimal average approach which had yielded the highest correlation. Equation 5.8 is easily rearranged in order to obtain an estimate of k_ξ or, with the aid of Equation 5.4, the retention time

$$t_\xi = t_0 \left[1 + \exp\left(-0.015 \frac{\text{mol}}{\text{kJ}} E_{\text{mean},\,\xi} - 0.661 \right) \right] \tag{5.9}$$

of some substance ξ. Experimental (HPLC) capacity factors k_ξ and respective elution orders[229] are listed in Table 5.2 along with those obtained by the fitting procedure. The two average-based approaches not only yielded the highest Person correlation R^2 of energy values but as well the same high rank correlation coefficient R_s^2 according to Spearman which was calculated on the basis of the predicted elution order compared to the experimental order. Table 5.2 illustrates that the elution order predicted by these two models is correct except for one single exchange associated with the two adjacent isomers $(+)$-α and $(+)$-β-HBCD. Indeed, the smallest difference in HPLC capacity factors k between any pair of two stereoisomers is related to these two compounds and amounts to $\Delta_{\text{min}} k = 0.63$. This observation also holds for force field energies possibly explaining the only failure regarding the computed relative order of these two isomers. If E_{mean} (E_{prob}) of $(+)$-α-HBCD had just been calculated larger by 1.3 kJ mol^{-1} (0.5 kJ mol^{-1}), the elution order would have been exactly predicted. The predictive model's quality was evaluated through *leave-one-out cross-validation* (LOOCV). In detail, the capacity factor $\ln(k_\xi)$ of each "left out" isomer ξ was predicted on the basis of parameter coefficients x that had been trained as described above (by least-squares fitting) using the set of five compounds left over. This procedure guarantees that no

Table 5.2: Optimal interaction energies simulated using a β-pmCD stationary phase along with capacity factors and the corresponding elution order obtained from both HPLC experiment and through least-squares-fitting of optimal interaction energies. Modified reprint of the original publication of Durmaz et al. 2012.[164]

	Interaction energy $\left[\frac{\text{kJ}}{\text{mol}}\right]$			Capacity factor k				Elution order			
Isomer	E_{mean}	E_{min}	E_{prob}	HPLC	E_{mean}	E_{min}	E_{prob}	HPLC	E_{mean}	E_{min}	E_{prob}
$(-)$-α	-200.2	-247.1	-28.6	10.01	10.83	11.42	10.84	1	1	2	1
$(-)$-β	-201.9	-241.7	-29.1	10.80	11.11	10.60	11.39	2	2	1	2
$(-)$-γ	-231.7	-275.6	-33.5	17.02	17.47	16.92	17.60	6	6	6	6
$(+)$-α	-207.1	-255.6	-29.8	12.16	12.02	12.84	12.21	3	4	5	4
$(+)$-β	-205.9	-247.6	-29.4	12.79	11.80	11.50	11.74	4	3	3	3
$(+)$-γ	-209.3	-253.6	-30.1	13.62	12.43	12.49	12.58	5	5	4	5
Squared correlation coefficients:				R^2:	0.86	0.77	0.85	R_s^2:	0.89	0.69	0.89
Leave-one-out cross correlation:				R^2_{LOO}:	0.81	0.69	0.81				

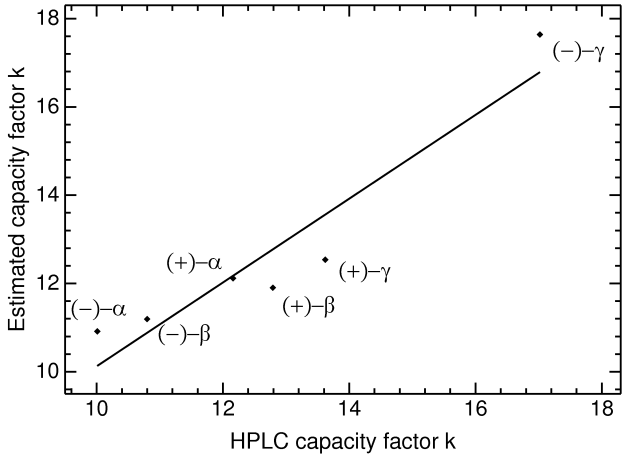


Figure 5.6: Correlation of Amber force field interaction energies (after least-squares fitting) with HBCD stereoisomer capacity factors from chiral HPLC separation. Reprinted from the original publication of Durmaz et al. 2012.

compound ξ contributes to the training set used for predicting compound ξ itself. The resulting (squared) correlation coefficient R^2_{LOO} describes the dependence of those entirely predicted values on the experimental ones. In general and as indicated by Table 5.2, this coefficient is smaller than the one quantifying the correlation of fitted values with experimental data which is denoted as R^2 above. The LOOCV method confirms on the one hand the suitability of the presented predictive method and, on the other, what has already been indicated by Pearson's correlation coefficient: ensemble-based models exhibit significantly more predictive power in terms of thermodynamic systems than algorithms that rely only on a single state. Figure 5.6 shows capacity factors k_ξ predicted on the basis of E_{mean} (with water as solvent) and plotted against respective HPLC results. Interestingly, the extraordinarily high affinity was correctly estimated for $(-)$-γ-HBCD which is indeed eluted a noticeable time period after all other stereoisomers as sketched in the chromatogram (Figure 5.1). In explicit water, the enantiomer-specific separation was estimated correctly with all approaches. However, it should be noted, that the chromatographic separation in pure water or pure ACN does not yield differing elution orders and does not lead to the separation of all six stereoisomers. Already for this reason, we cannot expect an exact agreement from simulations using one pure solvent. Regarding the three models handling multiple binding modes for affinity analysis, statistical approaches on the basis of mean potential energies (inner energies) turn out to be more convenient than single-geometry approaches as represented by E_{min} since they better reflect the microscopic variability associated with thermodynamic ensembles. [164]

5.6 Concluding remarks

An empirical approach to *in silico* determination of elution orders as well as retention times of HPLC runs on the basis of explicit solvent force field simulations has been presented. In order to remedy the trapping problem which is inherent to MD simulations, the space of host–guest binding modes was uniformly decomposed as described in Section 4.3. Prior to this, global minimum energy conformations of the guest molecules had been determined according to the HMC approach sketched in Section 4.1. For these reasons, the observations made in this chapter can be considered as an additional successful evaluation of methods developed in the previous chapter.

Again, the system consisting of six major HBCD stereoisomers has been taken as an example for which results of an analytical HPLC separation with a water/ACN gradient through a β-pmCD column were available for the purpose of comparison. Consequently, calculations were carried out in each solvent. In addition, three physical model descriptors have been investigated based on force field energies either of a single time step, of the preferential binding mode's time-average, and of the Boltzmann-weighted sum of all orientations. The effect of various force field potentials particularly suitable for the quantification of intermolecular interactions has been investigated as well. Particular emphasize was laid on consistency of the results regarding the time range of the MD trajectories under consideration.

Having skipped the first 80 ps of the time series due to equilibration events obvious from center of mass distance calculations between host and guest molecules, the data obtained from simulations in pure water achieved significantly better all in all correlations with wet lab experiments than those associated with ACN. This particularly holds for the sum of Coulomb and Lennard-Jones potentials which was compared to HPLC capacity factors. Nevertheless, the results as well indicate that the host–guest interaction is dominated by van der Waals rather than electronic forces. No reliable reproduction of the HBCD elution order was possible with the solvent ACN. Moreover, regarding distances of these two molecules in ACN, no steady state was achieved during 600 ps of simulation due to the occurrence of large jumps distributed over the entire time range. In contrast, calculations incorporating water instead reveal a satisfying equilibration reached after about 50 ps. Besides, the center of mass distance between host and guest averaged over time, initial binding modes, and stereoisomers after 400 ps were about 60 % larger in ACN and tendentially increasing. In addition, the simulation of HBCD in pure solvent (without β-pmCD) clearly revealed significantly lower interaction energies concerning ACN than water. These observations indicate that HBCD prefers

to reside rather in ACN than in water and consequently confirm experimental results. For all these reasons, the calculations agree very well with wet lab results where indeed elevated concentrations of the highly polar solvent water at the beginning of the eluent gradient enhance interaction between the two nonpolar substances β-pmCD and HBCD and, thus, HBCD separation whereas an increasing fraction of the rather hydrophobic solvent ACN favors the elution of analytes, specifically, in case of this host–guest combination.

Considering MD simulations performed in water, all descriptive models achieved very high correlations with HPLC results that were considerably consistent regarding the underlying time range. All of them performed accurately in the prediction of an enantiomer-specific elution order. In particular, for the two models based on statistical averages, exceptionally and similarly high squared Pearson coefficients of correlation have been calculated. The elution order of the six HBCD stereoisomers estimated upon these two models had the smallest possible combinatoric deviation from the correct elution order since only the two compounds with the smallest difference in HPLC retention times had been interchanged by the predictive models. This is well reflected by high rank correlations according to Spearman. The strong correspondence of results obtained from these two models clearly implies a sufficient approximation of the "entire" conformational space (Boltzmann-weighted sum of 60 binding modes) by the representative subspace associated with the lowest inner energy (preferential binding mode) yielding a similar accuracy at substantially less computational costs. The approximation makes even more sense since the respective scientist usually is interested in a picture of one single preferential binding mode. All these results clearly serve as a positive validation of the previously described strategy dealing with a uniform decomposition of the relative host–guest orientation. The method is highly suitable for molecular docking to predefined binding sites of host molecules. However, as can be seen from the inferior performance of the physical model based on only one single conformer instead of time-averages, the estimation of binding modes as well as affinities benefit from information of molecular dynamics. Indirectly, the successful reproduction of wet lab results confirms the convenience of the AMBER force field for this type of organic solvated systems.

An empirical linear model for the prediction of quantitative retention times was constructed by fitting of the sum of intermolecular force field interaction energies to HPLC capacity factors. The resulting parameter coefficients can be used for the prediction of retention times of substances that underwent an HPLC separation under exactly the same conditions, i. e., that were part of the same mixture. The linear model's predictive quality and suitability for substances other than those comprising the training set was

conclusively evaluated as very well by means of the LOOCV method. The strategy is useful whenever experimental assignment of peaks to stereoisomers is impossible or for selecting suitable stationary phases for a given mixture of compounds. However, the presented approach is, due to the way binding modes are determined, only applicable to host-like stationary phases.

6 Novel ERα binding affinity model

The molecular modeling techniques and results presented in this chapter have to a notable extent been published in the following article.[194] A republication of related contents in the framework of this thesis was kindly permitted by its publisher:

- V. Durmaz, S. Schmidt, P. Sabri, C. Piechotta, M. Weber: A hands-off linear interaction energy approach to binding mode and affinity estimation of estrogens. *Journal of Chemical Information and Modeling*, 53(10):2681–2688, 2013.

This chapter aims at the development of a physical model suitable for the prediction of binding affinities associated with biological host–guest systems. Using, by way of example, the estrogen hormone receptor alpha (ERα) already introduced in Section 4.3 and a series of diverse natural and synthetic ligands collected from various sources, we will describe in detail how differences in binding free energies can be derived from physical descriptors on the basis of classical force field simulations. The quantification of the host–guest binding strength elaborated here constitutes the second major part of a two-step procedure as it directly builds upon the determination of a preferential ligand binding pose as devised in Section 4.3. In conjunction, these two methods are supposed to provide, for some small molecule under observation, highly accurate binding affinities in a fully automatized fashion with no further a priori information than a crystallographic structure file of the target molecule and a spatial specification of its binding site.

6.1 Introduction

An as accurate as possible prediction of binding affinities related to biological protein–ligand systems is still a challenging task in the area of pharmaceutical and toxicological research. In particular, structure-based design of drugs was substantially accelerated due to fast virtual screenings and lead optimization supported by *in silico* methods.[238,239] In this sense, *in silico* experiments often significantly reduce the setup of biological assay experiments. Moreover, predictive computational methods turn out to be the only access

to toxicity assessment on metabolites (transformation products) of anthropogenic substances if no structure determination or synthesis for experimental analysis is possible. However, up to now and in spite of the vast (parallel) computing power available nowadays, the prediction of binding modes and affinities for complex host–guest systems remains a time-consuming and highly non-trivial challenge.[194] In order to achieve reliable results, several computational tasks need to be carried out of which the complexity increases drastically with the number of atoms.[240] And since upon the binding process both molecules adopt a proper conformation, it is important to consider the flexibility of both protein and ligand.[238] Furthermore, it should be noted that the binding affinity on its own does not necessarily make a useful statement about the ligand's qualitative influence on metabolism. Relating to receptor proteins, the ligand may either act as an *agonist* (activating the protein's function) or an *antagonist* (inhibiting its function through competitively preventing the binding of natural ligand). We have already discussed the theoretical background of biological association/dissociation reactions in Chapter 2. Nevertheless, in any case such small molecules would be classified as *endocrine disrupting chemicals* (EDC) as they interfere with the hormone system and consequently considered potentially harmful.[241] A prominent example of critical human target proteins is the hormone receptor ERα shown in Figure 4.13. In the face of high estrogenic activities of several synthetic compounds, the risk of endocrine disruption by xenoestrogens has already been elucidated in the early eighties.[214] Consequently, this target system has been undergoing many investigations regarding the prediction of binding affinities by means of computational methods. As already stated in Section 4.3, van Lipzig and co-workers achieved impressive coefficients of correlation around 0.9 ± 0.04 for ERα using an extended LIE model with 19 structurally similar ligands.[84] In advance, however, four ligand orientations had been selected manually inspired by crystallographic data. High squared coefficients q^2_{Loo} of leave-one-out cross-validation up to 0.71 was achieved with pure 3D-QSAR methods investigating ERα and xenoestrogens with initial conformations selected through comparison with known binding modes.[109] Using a QSAR model along with partial least square regression on affinity prediction for ERα, Wang and co-workers achieved a correlation of $r^2_{\text{Train}} = 0.92$ and $r^2_{\text{Test}} = 0.84$ for the training and test set, respectively, but poor cross-validation with $q^2_{\text{Loo}} = 0.43$ which seems somewhat surprising in comparison with the test set's correlation.[242] Many other models for the estimation of host–guest affinities do either yield poor cross-validation values or include the manual selection of an initial binding mode.[243,244]

In case of cross docking, the binding affinity estimation is preceded by the determination of a favorable host–guest binding mode. As pointed out in Chapter 2, common

methodology for the second step ranges from extensive perturbation approaches (FEP and TI) using large sets of MD simulations up to much faster but less accurate QSAR methods. Besides their high computational cost, TI and FEP cannot be applied to highly diverse compounds. In the light of these considerations and due to a moderate computational effort in conjunction with its proven satisfactory accuracy,[83–85] an extension of the original LIE model[78] on the basis of classical force field simulations was opted for all binding affinity calculations presented in this chapter. Besides, this method copes well with physical principles since it considers ensembles of Boltzmann-distributed microstates and, therefore, the involved molecules' flexibility. Only two ensembles representing the two endpoints of the binding process are required for an affinity estimation: the unbound ligand in a solvent box on the one side and the ligand in complex with the target molecule, solvated as well, on the other side. Again, we want to point out that an ideal predictive model should abstain from any preliminary information about the ligand's orientation and in particular, avoid the manual or random choice of some favorable pose. Using a suitable thermodynamic model of a particular active site, the majority of all compounds coming into consideration and representing a wide range of affinities and structural properties should, ideally, be covered by the same parameter and training set. Consequently, we are going to develop an as simple as efficient predictive linear model that is, in conjunction with the systematic sensing of the space of binding modes described earlier, able to estimate highly accurate binding affinities of some chemical compound in a fully automatic manner starting from a crystallographic protein structure and a binding site definition in terms of three Cartesian coordinates.

6.2 ERα modeling and force field simulations

As ensuing from the preferential binding mode prediction described in Section 4.3, all binding affinity calculations presented in the following base on the same PDB entry 1GWR[215] of ERα available in the PDB data base and particularly comprising the ligand binding domain in complex with its natural binder 17β-estradiol (E_2). Initially, all components except for amino acid (AA) atoms of chain A were removed from the coordinates file. Due to an incomplete resolution of chain A corresponding to the ligand binding domain, this PDB entry is missing atomic coordinates associated with AA residues Asp332, Pro333, and Thr334 as well as Leu462, Ser463, and Ser464. For the sake of structure completion, the PDB entry 3ERD[245] served as a template contributing coordinates of two respective sequences: Ala322-Ala340 and Ile452-Leu469.

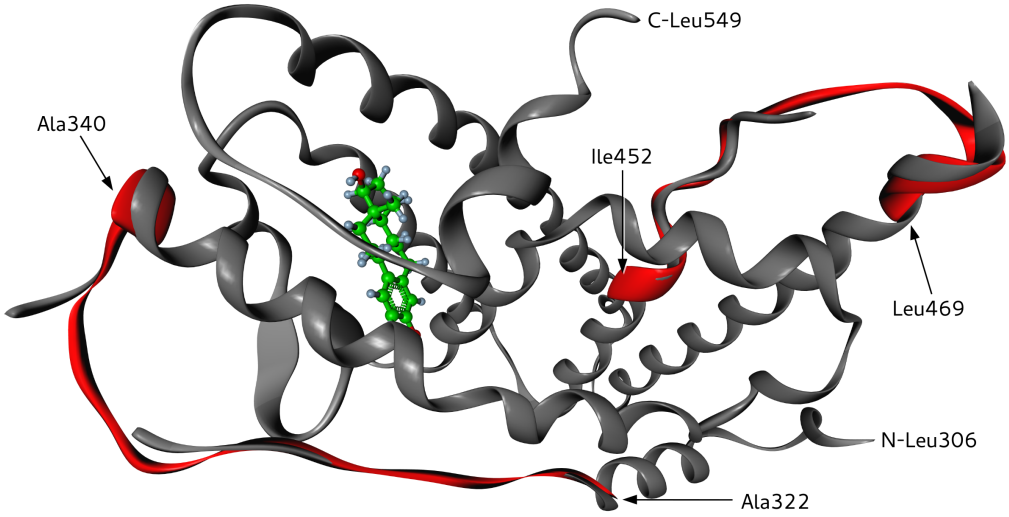


Figure 6.1: ERα protein model based on PDB entry 1GWR (grey secondary structure) originally complexed with the natural hormone 17-β-estradiol (green carbon scaffold) and completed using two substitutes (Ala322-Ala340, Ile452-Leu469) originating from PDB file 3ERD after a backbone alignment of both crystal structures.

Model building is illustrated in Figure 6.1 using the secondary structure of 1GWR (grey) and red-colored sequences originating from 3ERD after a structural alignment of both backbones. In order to avoid the usage of loose end coordinates associated with those incomplete regions, the two 3ERD substitutes included a couple of additional vicinal AAs such that a satisfactory overlap with 1GWR characterized by a minimal atomic deviation was guaranteed. We want to refer to the fact that the minimal distance between any atom of E_2 and the substitute amounts to 6.2 Å which is related to the hydroxy-hydrogen attached to the aromatic ring of E_2 and one of the hydrogens attached to a primary carbon atom of AA Leu327. As already pointed out, the complete coordinate file was provided with amber99sb force field parameters before being CG energy-minimized and used in complex with ligands for explicit solvent MD simulation.

A set of 31 ligands (depicted in Figure 6.2) including diverse scaffolds and associated with respective binding affinities to ERα spread over 10^7 magnitudes served as training as well as cross-validation set for the empirical model. They were collected quasi-randomly from three different sources (Table 6.1): twelve compounds originated from Kuiper's set,[246] another ten compounds had been published by Blair et al.,[247] and binding affinities of further eight substances had been determined by the Federal Institute for Materials Research and Testing (BAM) in Germany using binding assay studies.[194] In order to be able to compare binding affinities from different sources, relative binding affinities (RBAs) were derived from pharmacological IC_{50} (*inhibitory constant*)

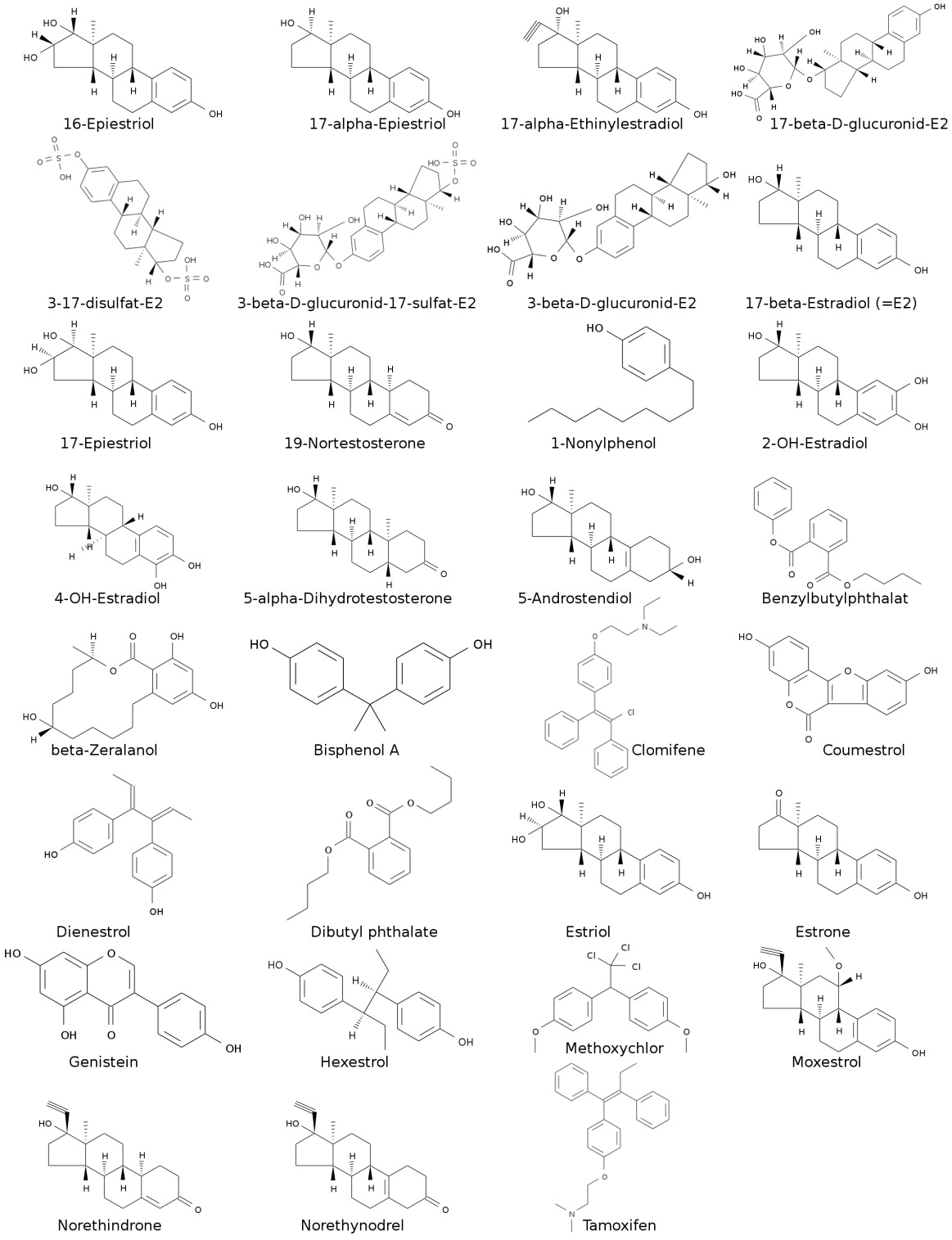


Figure 6.2: Training and model evaluation set of compounds originating from three different sources with known (relative) binding affinities to the estrogen hormone receptor ERα.

Table 6.1: Training and model evaluation set of compounds originating from three different sources.

Kuiper et al., 1997	Blair et al., 2000	BAM, 2010–2011
17-β-Estradiol (E_2)	17-α-Estradiol	17-α-Ethinylestradiol
17-β-D-glucuronid-E_2	19-Nortestosterone	2-Hydroxy-E_2
3-17-Disulfat-E_2	5-Androstenediol	4-Hydroxy-E_2
3-β-D-Glucuronid-E_2	5-α-Dihydrotestosterone	1-Nonylphenol
3-β-D-Gluc.-17-sulfat-E_2	β-Zearalenol	Coumestrol
16-Epiestriol	Bisphenol A	Dienestrol
17-Epiestriol	Clomifene	Estrone
Benzylbutylphthalate	Estriol	Hexestrol
Dibutyl phthalate	Genistein	Norethynodrel
	Methoxychlor	Tamoxifen
	Moxestrol	
	Norethindrone	

values in relation to the natural binder E_2 which obtained the value $\mathrm{RBA}(E_2) = 100$. Consequently, the RBA value of any other compound L was expressed in terms of the IC_{50} values of E_2 and substance L,

$$\mathrm{RBA}(L) = \mathrm{RBA}(E_2)\,\frac{IC_{50}(E_2)}{IC_{50}(L)} = 100\frac{IC_{50}(E_2)}{IC_{50}(L)}.$$

Basically, the system setup and molecular simulations of ERα with the set of 31 ligands followed the protocol in Section 4.3. In short: each ligand was subjected to a hybrid Monte Carlo sampling from which the global energy minimum (as described in Section 4.1) was selected as input geometry for a series of host–guest molecular mechanics simulations. After explicit solvation and charge neutralization, each of the 60 target–ligand complexes per ligand underwent a local geometry optimization as well as an MD simulation for equilibration and unrestrained production purposes according to the description in Section 4.3.[194] During an additional production run for the sake of comparison, the backbone atoms of ERα experienced position restraints keeping fixed the tertiary structure retrieved from PDB. Finally, the most favorable binding mode associated with the lowest time-averaged interaction energy (according to Equation 4.8) was chosen out of the set of 60 orientations per ligand for further analysis and model development. The limitation on the state with the highest statistical weight had been justified in the previous chapter regarding the HPLC elution order of HBCD stereoisomers. That is, results obtained with the likeliest state were nearly identical to the Boltzmann-weighted sum of all binding modes.

6.3 Monte Carlo approach to conformational entropies

As we have already learned in an earlier chapter, the Gibbs free energy $G\,(N, p, T)$ not only depends on enthalpic but as well on temperature-dependent entropic contributions of the molecular system under observation. In other words, apart from repulsive and attractive atomic interactions, the number of possible system manifestations at a given temperature influences the value of G, too. The multiplicity of possible states is mainly attributed to the conformational diversity as well as solute–solvent arrangements. A couple of studies revealed some linear relationship between the system's enthalpy and certain entropic contributions obtained from MD simulations.[248,249] This relation is commonly referred to as *entropy–enthalpy compensation*. In spite of that and in contrast to most LIE applications, we decided to explicitly include and investigate an entropy estimate representing conformational diversity within the statistical ensemble of a ligand. An auspicious theoretical framework for calculating conformational entropies on the basis of MD trajectories has recently been proposed by Weber et al.[250,251] Essentially, the method is about a Monte Carlo approach to the estimation of *conformational entropies* S based on the variance of internal atomic coordinates (conformers) $q_i \in \mathbb{R}^{3N}$ over time frames i. We will, in the following, have a detailed look at its implementation in the context of binding affinity calculations using the LIE method. The central idea is to express S through the fraction of conformers q_i, whose RMSD value r_{ij} to some properly chosen reference states q_j is beneath a certain cutoff value r_{ref}. In other words, one expects that stiff compounds prefer residing in close vicinity of some reference whereas flexible ones are supposed to spread widely (Figure 6.3). Since all entropy calculations

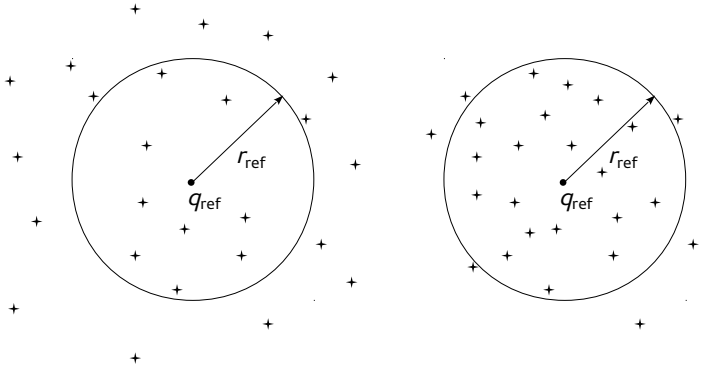


Figure 6.3: Monte Carlo estimator for conformational entropies exemplified for a two-dimensional case: the entropy is derived from the fraction of MD states (given in internal or external coordinates) with an RMS deviation larger than a certain cutoff r_{ref} from some reference state q_{ref} which has approximately average energy. The system on the left is associated with higher entropy.

presented here are related to ligand molecules only, it was necessary to omit all other components from the trajectories of complex (B) as well as unbound (U) production runs. Potential energies U were recalculated for each time step i of the remaining N_{L} ligand atoms associated with coordinates $q'_i \in \mathbb{R}^{3N_{\mathrm{L}}}$. Hence, given such a narrowed MD trajectory in the form of a set (indicated by curly brackets)

$$T = \{q'_i\} \quad \forall\, i \in [1, n]$$

of n frames along with a list of corresponding ligand potential energies $U\left(q'_i\right)$, the implemented algorithm first determines the time-averaged geometry

$$\overline{q'} = \frac{1}{n} \sum_{i=1}^{n} q'_i$$

and its RMS deviation

$$r_i = \mathrm{rmsd}\left(q'_i, \overline{q'}\right)$$

from every frame i. Afterwards, the cutoff distance

$$r_{\mathrm{ref}} = \frac{\sigma^{\mathrm{U}} + \overline{\sigma^{\mathrm{B}}}}{2} \tag{6.1}$$

valid for the entire set of 61 (60 bound and one unbound) simulations per compound was expressed in terms of two standard deviations: one

$$\sigma^{\mathrm{U}} = \sigma\left(\left\{r^{\mathrm{U}}\right\}_{\mathrm{U}}\right)$$

representing the uncomplexed system and its counterpart

$$\overline{\sigma^{\mathrm{B}}} = \frac{1}{60} \sum_{l=1}^{60} \sigma\left(\left\{r^{\mathrm{B}}\right\}_{l}\right)$$

associated with RMSD values component-wisely averaged over 60 complex binding modes. The functionality of r_{ref} is demonstrated by Figure 6.3 for a two-dimensional case. A system with a higher conformational entropy value would be less restrained and consequently exhibit a higher variance of coordinates. Such a system would correspond to the distribution on the left characterized by larger RMS deviations from some reference states q_{ref}. In contrast, a bound ligand molecule residing at the binding site of a target molecule would be substantially restrained and yield a lower conformational entropy which is represented by the distribution on the right. In practice, $k = 10$ representatives

$$q'_j \in T \quad \forall\, j \in [1, k]$$

with approximately average potential energies[250]

$$U\left(q'_j\right) \approx \langle U\left(q'_i\right)\rangle_i$$

(states associated with lowest deviations from the average energy) had been selected out of $n = 50,000$ frames (200 ps) as illustrated by black circles in Figure 6.4. Afterwards, for each such representative q_j the number $n_j \in [1, n]$

$$n_j = \left| \left\{ q'_i\,,\ \text{rmsd}\left(q'_i, q'_j\right) < r_{\text{ref}} \right\} \right|$$

of states q_i characterized by RMS deviations from the reference states less than r_{ref} was calculated. Using the fraction n_j/n averaged with respect to k representatives, the estimated conformational entropy derived from an MD trajectory amounts to

$$S \approx -R \ln\left(\frac{1}{k}\sum_{j=1}^{k}\frac{n_j}{n}\right).$$

R denotes the gas constant used instead of the Boltzmann constant k_{B} which is proposed in the original paper. The entropy value clearly depends on $n_j \in [1, n]$ and consequently ranges from $S = -R\ln(1/n)$ in case of maximum entropy ($n_j = 1$) to $-R\ln(1) = 0$ if no variance at all is given ($n_j = \text{n}$). It should be noted that in the original publication of the extended LIE model for estrogens,[194] the reference distance calculated according to Equation 6.1 was equated with the standard deviation of the unbound system only, $r_{\text{ref}} = \sigma^{\text{U}}$. However, since we are interested in entropy *differences* ΔS between the bound and each of the 60 bound ensembles, it seems more appropriate to consider all the 61 standard deviations in Equation 6.1. As an acceptable consequence, one obtains a highly pronounced entropy discrimination of all involved trajectories. That is, binding mode samplings associated with the limits $n_j = 1$ (*maximal entropy*, all points except for reference state itself outside the circle in Figure 6.3)

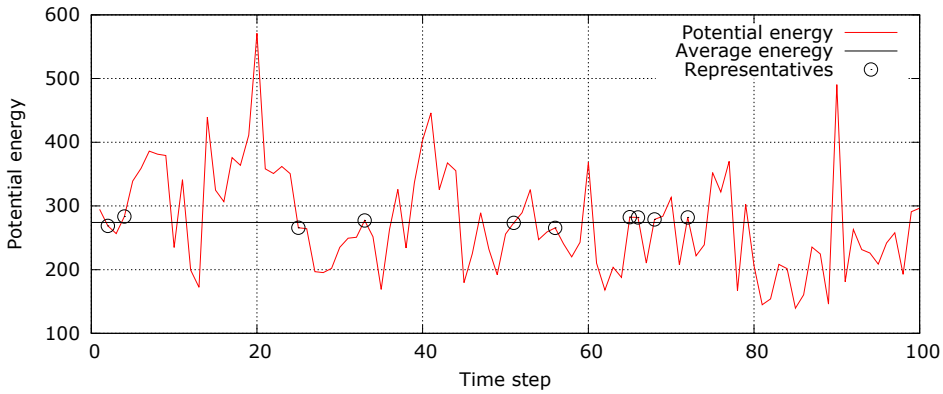


Figure 6.4: Selection of reference states associated with approximately average energy for the Monte-Carlo estimation of conformational entropies.

or $n_j = n$ (*zero entropy*, all points within the circle) are expected to appear less frequently. We want to point out, that theoretically r_{ref} could have as well been set to any proper constant with a highly discriminating power. Another modification of the original entropy model concerns the type of coordinates used for RMSD calculations. Rather than on the basis of internal coordinates defined by torsion angles and related to conformational changes only, we decided to comprise external (Cartesian) coordinates in order to include translational/rotational degrees of freedom in addition.

6.4 Extended LIE model and cross-validation

We had published[194] a first LIE-based predictive model not only on the basis of classical interaction energy terms but of structural descriptors as well. To be more precise, our final $m = 6$ parameter LIE model incorporated four extra parameter terms in addition to time-averaged van der Waals $\Delta \left\langle E^{\text{vdW}} \right\rangle$ and Coulomb $\Delta \left\langle E^{\text{elec}} \right\rangle$ contributions: average potential energy differences $\Delta \left\langle U^{\text{Lig}} \right\rangle$ (*strain energy* taken up upon binding), the conformational entropy TS^{conf} of the ligand derived from the unbound system and multiplied with temperature T. Finally, the model comprised two boolean structural descriptors, $\delta^{\text{benz}} \in \{0, 1\}$ and $\delta^{\text{phen}} \in \{0, 1\}$, indicating the presence of a benzene ring and, respectively, of a hydroxy phenyl group which would be typical for QSAR methods. Due to physical reasons, however, it seems more convenient to use entropy differences instead of entropies of the unbound ligand because the related entropy loss is more significant in terms of thermodynamics (see Equation 2.4). Using a combination of these descriptors, the new functional form of the linear model looked like

$$\begin{aligned} \Delta G^{\text{comp}} = {} & x_1 \Delta \left\langle E^{\text{elec}} \right\rangle + x_2 \Delta \left\langle E^{\text{vdw}} \right\rangle + x_3 \Delta \left\langle U^{\text{lig}} \right\rangle + x_4 T \Delta S^{\text{conf}} \\ & + x_5 \delta^{\text{benz}} + x_6 \delta^{\text{phen}} \end{aligned} \tag{6.2}$$

where Δ accounts for the difference between any ligand's bound and unbound sampling. According to the original LIE model based on thermodynamic principles, the first two summands of Equation 6.2 address all pairwise intermolecular force field interactions involving ligand atoms. These are ligand-target and ligand-solvent interactions in the bound case and, respectively, solely ligand-solvent interactions in case of the unbound system. Following the procedure described in Chapter 5 regarding HPLC retention times, the coefficients x_i where calculated on the basis of $n = 31$ ligands.

Consequently, a system of 31 linear equations

$$\underbrace{\begin{pmatrix} \Delta G_1^{\text{exp}} \\ \vdots \\ \Delta G_n^{\text{exp}} \end{pmatrix}}_{y} \approx \underbrace{\begin{pmatrix} \Delta \left\langle E_1^{\text{elec}} \right\rangle & \Delta \left\langle E_1^{\text{vdw}} \right\rangle & \dots & \delta_1^{\text{phen}} \\ \vdots & \vdots & \ddots & \vdots \\ \Delta \left\langle E_n^{\text{elec}} \right\rangle & \Delta \left\langle E_n^{\text{vdw}} \right\rangle & \dots & \delta_n^{\text{phen}} \end{pmatrix}}_{A} \underbrace{\begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix}}_{x} \tag{6.3}$$

was constructed resulting in a vector $y \in \mathbb{R}^n$ of experimentally determined free energies of binding and a matrix $A \in \mathbb{R}^{n \times m}$ consisting of structural descriptors and MD averages weighted by $x \in \mathbb{R}^m$. While the empirical model of the underlying publication [194] comprised decadic logarithms of RBA values for the construction of y, we decided, this time, to operate on the correct physical formulation of binding free energies

$$y \;=\; \Delta G^{\text{exp}}\left(L\right) \;=\; RT \ln \left[K_{\text{d}}\left(E_2\right) \frac{\text{RBA}\left(E_2\right)}{\text{RBA}\left(L\right)} \right] \tag{6.4}$$

which had been derived for each ligand L from its RBA value. The expression inside the logarithm of 6.4 is related to dissociation constants K_{d} amounting to 0.2 nM. Apart from its physical relevance, results obtained through this strategy are better comparable with other published methods. All values accounting for matrix A are listed in Table 6.2 where overall uncertainties

$$u_O = \sqrt{\left(\frac{\partial O}{\partial O_{\text{B}}}\right)^2 \sigma_{O_{\text{B}}}^2 + \left(\frac{\partial O}{\partial O_{\text{U}}}\right)^2 \sigma_{O_{\text{U}}}^2} = \sqrt{\sigma_{O_{\text{B}}}^2 + \sigma_{O_{\text{U}}}^2}$$

of any time-averaged parameter O have been propagated on the basis of standard deviations $\sigma_{O_{\text{B}}}$ and $\sigma_{O_{\text{U}}}$ corresponding to bound (B) and unbound (U) ligand simulations. In the recent journal publication [194] of the presented LIE approach, each column of A representing a particular descriptor was normalized by subtracting its mean and dividing by its standard deviation. For reasons stated above in association with RBA values and since the LIE model in its original formulation and most extensions and applications do not carry out this step but rather directly fit Ax to y, we omitted normalization as well in the following yielding slightly different results. Nevertheless, x was determined through a least-squares approach using normal equations

$$x = \left(A^\top A\right)^{-1} A^\top y. \tag{6.5}$$

The weights are identical with those obtained through *QR decomposition* which is known to perform better on ill-conditioned matrices. [252] Table 6.3 shows directly fitted weights x for the entire six-parameter model as well as two descriptor subsets yielding a four and two-parameter model. Regarding the number of empirical parameters, the latter model

Table 6.2: Calculated and structural (boolean) descriptor values of matrix A used for fitting coefficients of an empirical six-parameter LIE model. Uncertainties were propagated on the basis of standard deviations associated with bound and unbound ligands.

Ligand	$\Delta\langle E^{\text{elec}}\rangle$	$\Delta\langle E^{\text{vdw}}\rangle$	$\Delta\langle U^{\text{lig}}\rangle$	$T\Delta S^{\text{conf}}$	δ^{benz}	δ^{phen}
17-β-Estradiol (E_2)	2.1 ± 26.6	-50.8 ± 24.6	-7.5 ± 21.5	-4.33	1	1
17-α-Estradiol	-9.7 ± 26.7	-48.4 ± 24.5	2.9 ± 21.9	-3.19	1	1
17-α-Ethinylestradiol	-4.4 ± 26.5	-75.7 ± 24.4	-4.5 ± 22.8	-9.97	1	1
2-Hydroxy-E_2	-41.5 ± 34.2	-35.5 ± 31.6	15.7 ± 24.8	-6.27	1	1
4-Hydroxy-E_2	-50.2 ± 30.3	-33.6 ± 27.9	37.0 ± 23.0	-6.70	1	1
17-β-D-glucuronid-E_2	145.8 ± 94.2	384.3 ± 55.2	-22.1 ± 35.4	-6.13	1	1
3-17-Disulfat-E_2	160.5 ± 54.6	395.1 ± 47.0	19.5 ± 25.9	-1.71	1	0
3-β-D-Glucuronid-E_2	115.7 ± 61.1	355.3 ± 48.8	38.4 ± 35.6	-5.37	1	0
3-β-D-Gluc.-17-sulfat-E_2	180.5 ± 74.0	587.8 ± 54.0	30.3 ± 33.8	-4.20	1	0
16-Epiestriol	-12.8 ± 34.5	-24.4 ± 28.5	29.3 ± 23.6	-5.26	1	1
17-Epiestriol	-0.1 ± 35.2	-28.0 ± 29.8	12.8 ± 26.0	-2.78	1	1
19-Nortestosterone	17.0 ± 24.0	-73.2 ± 22.4	7.8 ± 23.2	-8.64	0	0
1-Nonylphenol	-6.5 ± 20.5	-86.3 ± 19.4	13.8 ± 22.5	-6.94	1	1
5-Androstenediol	3.8 ± 29.3	-70.3 ± 25.4	2.1 ± 24.0	-3.04	0	0
5-α-Dihydrotestosterone	21.7 ± 23.9	-79.6 ± 22.3	4.0 ± 26.1	-8.18	0	0
Benzylbutylphthalate	18.8 ± 18.7	-92.2 ± 19.2	20.8 ± 22.9	-4.22	1	0
β-Zearalenol	-7.0 ± 28.9	-53.3 ± 26.5	72.6 ± 25.0	-4.85	1	1
Bisphenol A	6.0 ± 26.9	-10.7 ± 24.1	2.1 ± 19.8	-5.26	1	1
Clomifene	41.1 ± 19.8	-176.5 ± 21.1	5.5 ± 28.2	-6.88	0	0
Coumestrol	-17.2 ± 30.0	8.6 ± 26.3	-1.6 ± 16.8	-2.52	1	1
Dibutyl phthalate	29.4 ± 21.4	-91.1 ± 18.9	50.3 ± 22.6	-6.07	1	0
Dienestrol	9.6 ± 27.4	-23.1 ± 25.2	8.3 ± 21.3	-2.69	1	1
Estriol	14.7 ± 33.5	-0.4 ± 29.4	-6.6 ± 24.3	-5.39	1	1
Estrone	5.8 ± 23.7	-57.5 ± 22.0	2.5 ± 21.6	-5.72	1	1
Genistein	5.7 ± 29.3	3.2 ± 26.9	2.4 ± 19.8	-3.36	1	1
Hexestrol	15.3 ± 30.4	-42.2 ± 25.3	11.9 ± 22.8	-3.65	1	1
Methoxychlor	21.1 ± 14.3	-131.8 ± 15.2	4.3 ± 20.8	-5.09	1	0
Moxestrol	11.0 ± 26.8	-65.3 ± 25.5	14.0 ± 24.8	-6.69	1	1
Norethindrone	22.3 ± 24.7	-101.6 ± 23.5	9.2 ± 23.8	-8.81	0	0
Norethynodrel	17.7 ± 25.0	-96.7 ± 22.4	4.6 ± 24.6	-10.68	0	0
Tamoxifen	31.0 ± 28.5	-141.1 ± 21.7	-1.6 ± 28.3	-5.99	1	0

resembles the original LIE method[78] which proposes a fixed value of $x_{\text{elec}} = 0.5$ for the electrostatic (elec) term and some fitted weight around $x_{\text{vdw}} = 0.16$ for van der Waals (vdw) contributions. However, fitting only both interaction terms yielded 0.33 and, respectively, 0.74 as optimal coefficients. Their signs are always positive due to the corresponding parameters' positive relation with ΔG in Equation 6.2. That is, small values of both interaction energies on the one hand and ΔG on the other are associated with higher binding affinities. In contrast to any energy contribution, the entropy term is negatively correlated with the Gibbs free energy since a high loss of conformational flexibility upon binding most likely correlates with lower binding affinity. In addition, all entropy differences listed in Table 6.2 obtained negative signs confirming the general loss of conformational entropy during an molecular association process. From that

Table 6.3: Least-squares weights x for a 2, 4, and 6-parameter LIE model fitted directly to experimental ΔG and, in case of the 6D model, after normalization and along with mean and standard deviations (std). Reprinted with major modifications from the original publication of Durmaz et al. 2013.

Descriptor	Normalized fit			Direct fit		
	Mean	std	$x^{(6)'}$	$x^{(6)}$	$x^{(4)}$	$x^{(2)}$
$\Delta \left\langle E^{\text{elec}} \right\rangle$	21.42	49.75	1.55	0.27	0.38	0.33
$\Delta \left\langle E^{\text{vdw}} \right\rangle$	-67.98	16.84	2.02	0.77	0.91	0.74
$\Delta \left\langle U^{\text{lig}} \right\rangle$	12.20	18.70	0.36	0.18	0.27	–
TS^{conf}	12.02	2.05	-0.68	-0.91	-1.27	–
δ^{benz}	0.56	0.28	0.47	6.37	–	–
δ^{phen}	0.40	0.35	-1.60	-16.28	–	–

point of view, all parameter coefficients seem reasonable. We will further engage with the meaning of weight signs in the subsequent chapter where toxicities are estimated relative to some chemical using a LIE model for targets without training sets. The numerical condition κ associated with the squared matrix $A^\top A$ calculated with version 3.8.1 of the Octave function `cond()` on the basis of theoretical observables considerably increases with the model's size. While remaining between 3 and 20 if taking into account energy contributions only (the first three parameters), it quickly exceeds 10^3 and, further, 10^5 when including the entropy difference and, particularly, both boolean chemical descriptors. The results imply that A is very sensitive to small errors in free energies (y) and particularly ill-conditioned regarding δ^{benz} and δ^{phen}. That is, even small errors in y might cause large errors in x. Under such conditions it is generally advisable to use some more stable method such as the QR decomposition [252] which, however, had yielded identical weights x. Rather for illustration purposes, Table 6.3 comprises weights $x^{(6)'}$ along with descriptor-wise mean values and standard deviations according to the normalized approach. [194] The significance of a descriptor is better reflected by its coefficient if it has been normalized in advance. Both interaction energy terms, for example, and the indicator for carbolic acid δ^{phen} achieved the highest weights in $x^{(6)'}$. This observation implies that these descriptors highly correlate with ΔG and that corresponding physical properties might therefore play a central role in binding. Indeed, a structural analysis of the underlying PDB file reveals a particularly complex arrangement of hydrogen bonds partially mediated by a water molecule residing nearby and including several atoms of two polar AAs (Glu353, Arg394) as well as the phenolic hydroxyl group. Once having trained weights x, the best fit

$$\hat{y}_{\text{Fit}} = \Delta G^{\text{Fit}} = Ax$$

Table 6.4: Experimental (Lab) and calculated (Cal) binding affinities ΔG [kJ/mol] along with corresponding absolute deviations d using a 2, 4, and 6-parameter LIE model, MM/PBSA, and Autodock-Vina. The empirical LIE model includes a fitting (Fit) step. Calculated RBA values are exemplarily illustrated for the 6D LIE model (Fit6, Cal6). Squared coefficients of Pearson and Spearman's rank correlation as well as mean absolute deviations in kJ/mol are depicted at the bottom.

Ligand	Relative binding affinity			LIE_6					LIE_4				LIE_2				MMPBSA		Vina	
	RBA_{Lab}	RBA_{Fit6}	RBA_{Cal6}	Lab	Fit	Cal	d_{Fit}	d_{Cal}	Fit	Cal	d_{Fit}	d_{Cal}	Fit	Cal	d_{Fit}	d_{Cal}	Cal	d_{Cal}	Cal	d_{Cal}
17-β-Estradiol (E_2)	100.0000	64.2524	60.9500	-57.6	-56.4	-56.3	1.1	1.3	-54.3	-54.0	3.2	3.6	-47.3	-46.9	10.3	10.7	-126.7	69.1	-44.8	12.8
17-α-Estradiol [a]	58.0000	35.8486	33.9665	-56.2	-54.9	-54.8	1.2	1.4	-52.9	-52.6	3.2	3.6	-47.5	-47.1	8.7	9.1	-131.8	75.7	-40.6	15.5
17-α-Ethinylestradiol [b]	190.0630	197.3233	198.9376	-59.2	-59.3	-59.3	0.1	0.1	-56.9	-56.5	2.3	2.7	-56.0	-55.8	3.2	3.4	-147.6	88.4	-38.5	20.7
2-Hydroxy-E_2 [b]	29.5100	1019.578	1622.849	-54.4	-63.5	-64.7	9.1	10.3	-64.1	-65.3	9.6	10.9	-63.4	-64.5	9.0	10.1	-130.5	76.1	-37.7	16.7
4-Hydroxy-E_2 [b]	66.0690	15.2917	11.0559	-56.5	-52.7	-51.9	3.8	4.6	-50.5	-49.3	6.0	7.2	-57.6	-57.8	1.1	1.3	-130.9	74.4	-42.7	13.8
17-β-D-glucuronid-E_2 [c]	0.0015	0.0044	0.0117	-28.9	-31.7	-34.2	2.8	5.3	-18.5	-14.2	10.4	14.7	-14.0	-11.6	14.9	17.4	20.1	49.1	-34.3	5.4
3-17-Disulfate-E_2 [c]	0.0004	<0.0001	<0.0001	-25.5	-16.0	-11.4	9.5	14.1	-14.6	-9.5	10.9	16.0	-14.2	-9.5	11.3	16.1	225.1	250.6	-29.3	3.8
3-β-D-Glucuronid-E_2 [c]	0.0079	0.0375	0.0566	-33.2	-37.2	-38.3	4.0	5.1	-41.2	-43.1	8.0	9.9	-45.1	-47.2	11.9	13.9	8.3	41.5	-35.2	2.0
3-β-D-Gluc.-17-sulfate-E_2 [c]	0.0001	0.0024	0.0124	-22.0	-30.2	-34.4	8.2	12.4	-29.9	-34.0	8.0	12.0	-31.6	-35.9	9.6	14.0	158.5	180.4	-32.2	10.3
16-Epiestriol [c]	4.9390	4.6452	4.6147	-49.8	-49.7	-49.6	0.2	0.2	-45.1	-44.7	4.7	5.1	-49.1	-49.0	0.7	0.8	-129.3	79.5	-45.6	4.2
17-Epiestriol [c]	39.8800	106.0271	122.2517	-55.2	-57.7	-58.1	2.5	2.9	-55.2	-55.2	0.0	0.0	-50.9	-50.7	4.3	4.5	-134.1	78.9	-44.0	11.2
19-Nortestosterone [a]	0.0100	0.0058	0.0052	-33.8	-32.4	-32.1	1.4	1.7	-35.0	-35.2	1.2	1.4	-38.9	-39.1	5.1	5.2	-134.8	101.0	-45.2	11.4
1-Nonylphenol [b]	0.0032	0.1628	0.2882	-30.9	-41.0	-42.5	10.1	11.6	-35.0	-35.4	4.2	4.5	-39.1	-39.3	8.2	8.4	-122.0	91.1	-28.1	2.8
5-Androstenediol [a]	6.0000	2.3512	1.6277	-50.3	-47.9	-46.9	2.4	3.4	-55.7	-56.4	5.4	6.1	-49.2	-49.1	1.2	1.2	-139.9	89.6	-44.0	6.3
5-α-Dihydrotestosterone [a]	0.0500	0.0035	0.0022	-38.0	-31.1	-29.9	6.9	8.1	-33.7	-33.1	4.3	4.9	-36.5	-36.4	1.5	1.5	-151.5	113.5	-46.1	8.1
Benzylbutylphthalate [c]	<0.0001	0.0001	0.0002	-21.0	-22.8	-23.3	1.8	2.3	-31.3	-31.6	10.3	10.6	-34.1	-34.4	13.1	13.4	-123.8	102.8	-32.7	11.6
β-Zearalenol [a]	16.0000	10.0378	6.4926	-52.8	-51.6	-50.5	1.2	2.3	-44.6	-39.3	8.2	13.5	-57.6	-57.9	4.7	5.0	-147.5	94.6	-36.8	16.0
Bisphenol A [a]	0.0500	0.1399	0.1540	-38.0	-40.6	-40.9	2.7	2.9	-34.7	-34.6	3.2	3.3	-34.2	-34.1	3.8	3.8	-98.9	60.9	-34.8	3.2
Clomifene [a]	25.0000	122.0147	250.6918	-54.0	-58.1	-59.9	4.1	5.9	-64.5	-65.5	10.5	11.5	-60.2	-60.7	6.2	6.7	-185.2	131.2	-29.3	24.7
Coumestrol [b]	0.8900	3.5847	4.2174	-45.4	-49.0	-49.4	3.6	4.0	-46.7	-46.8	1.3	1.4	-40.8	-40.6	4.6	4.8	-96.2	50.8	-38.5	6.9
Dibutyl phthalate [c]	<0.0001	<0.0001	<0.0001	-16.0	-15.3	-14.9	0.8	1.1	-19.9	-21.1	3.9	5.0	-33.0	-33.3	16.9	17.3	-126.3	110.3	-28.1	12.0
Dienestrol [b]	37.1530	0.6515	0.4723	-55.0	-44.6	-43.8	10.4	11.3	-39.3	-38.6	15.7	16.4	-36.6	-36.2	18.4	18.8	-117.5	62.5	-35.2	19.8
Estriol [a]	14.0000	36.2964	39.8539	-52.5	-55.0	-55.2	2.5	2.7	-51.7	-51.7	0.7	0.8	-46.2	-46.0	6.3	6.5	-123.3	70.8	-44.8	7.7
Estrone [b]	7.2400	1.1654	1.0088	-50.8	-46.1	-45.7	4.7	5.1	-41.1	-40.7	9.7	10.0	-40.0	-39.7	10.8	11.1	-130.5	79.7	-46.9	3.9
Genistein [a]	5.0000	1.2314	1.1098	-49.8	-46.2	-46.0	3.6	3.9	-41.7	-41.4	8.1	8.5	-38.0	-37.8	11.8	12.1	-101.6	51.8	-36.8	13.0
Hexestrol [b]	301.9976	33.0693	26.7667	-60.4	-54.7	-54.2	5.7	6.2	-50.6	-49.9	9.8	10.5	-47.5	-47.0	12.9	13.4	-116.3	55.9	-33.9	26.5
Methoxychlor [a]	0.0100	0.0020	0.0012	-33.8	-29.7	-28.3	4.1	5.6	-39.9	-40.1	6.1	6.3	-38.4	-38.5	4.6	4.7	-137.1	103.3	-26.8	7.0
Moxestrol [a]	43.0000	24.8549	23.9183	-55.4	-54.0	-53.9	1.4	1.5	-49.3	-49.0	6.1	6.4	-50.1	-49.9	5.3	5.5	-155.5	100.1	-35.6	19.8
Norethindrone [a]	0.0700	0.2778	0.3764	-38.8	-42.4	-43.2	3.6	4.3	-46.4	-47.2	7.6	8.4	-48.7	-49.0	9.8	10.2	-152.1	113.3	-39.4	0.5
Norethynodrel [b]	0.2137	0.0257	0.0124	-41.7	-36.3	-34.4	5.5	7.3	-39.5	-38.9	2.2	2.8	-44.1	-44.2	2.4	2.5	-149.0	107.3	-35.2	6.5
Tamoxifen [b]	1.6218	1.8644	2.0000	-46.9	-47.3	-47.5	0.4	0.5	-60.7	-62.1	13.8	15.1	-55.0	-55.4	8.1	8.5	-171.5	124.6	-27.6	19.3
Pearson correlation R^2 or mean absolute deviation MD of ΔG:					0.86	0.78	3.8	4.8	0.67	0.58	6.4	7.5	0.51	0.45	7.8	8.4	0.27	92.9	0.22	11.1
Spearman's rank correlation coefficient R_S^2 of ΔG:					0.79	0.76			0.61	0.58			0.54	0.47			0.13		0.18	

[a] Kuiper et al., 1997

[b] Blair et al., 2000

[c] BAM, 2010–2011

of binding free energy differences associated with the training set or some new compound is easily obtained from the inner product of its descriptor values A and coefficients x. If no other compounds with known binding affinity than those included in the training set are available, it is, for the judgement of an empirical model's predictive power, necessary to carry out some cross-validation. That is, no chemical should be involved in the training of weights used for its own assessment. In practice, each compound's binding affinity was predicted using coefficients that had priorly been fitted with respect to the remaining 30 substances by using Equations 6.3 and 6.5. We had already successfully employed this approach commonly denoted as leave-one-out cross-validation (LOOCV) for the validation of HPLC retention times in Chapter 5. The procedure yields an additional list of 31 binding affinities which we can interpret as *predicted* binding energies. Table 6.4 shows Gibbs free energy differences calculated using the MM/PBSA method as well as the popular docking tool Autodock-Vina in addition to three LIE models associated with different parameter sets as demonstrated by Table 6.3. Each of the LIE_m models with dimensions ranging from $m{=}2$ (LIE_2, resembling the original two-parameter LIE model) via $m{=}4$ (LIE_4) to $m{=}6$ (LIE_6, optimal extended model) is accompanied by two sets of theoretical ΔG values one originating from parameter fitting (Fit) and an additional set of predicted values through LOOCV (Cal). In contrast, MM/PBSA and Vina which we have applied for comparison purposes are only associated with predicted (Cal) free energies since no fitting by the user is required. Aside from that, for each of the five models absolute deviations d_{Cal} from experimental (Lab) Gibbs energies are specified, too. Lab ΔG values (listed under the tab LIE_6 in Table 6.4) have been derived (according to Equation 6.4) from relative binding affinities gathered from three different sources (column $\mathrm{RBA}_{\mathrm{Lab}}$) and using the relation between $K_{\mathrm{D}}{=}0.2$ nM and RBA$=100$ of E_2. In order to gain a better intuition for differences in K_{D} values, fitted and cross-validated RBA values related to the optimal six-parameter LIE model are given as well. Using this LIE model (without data normalization as contrary to the reference publication[194]) yielded RBA values ($\mathrm{RBA}_{\mathrm{Fit6}}$ and $\mathrm{RBA}_{\mathrm{Cal6}}$) that in parts strongly deviate from experimental $\mathrm{RBA}_{\mathrm{Lab}}$. The deviation is generally more pronounced in case of cross-validated relative binding affinities ($\mathrm{RBA}_{\mathrm{Cal6}}$) and sometimes more than by one order larger or less than the laboratory result. The largest deviation about a factor of 100 was calculated for 1-nonylphenol, 2-hydroxy-E_2, dienestrol, and the two sulfate derivates of E_2. However, converted into binding free energies associated with columns named Lab, Fit, and Cal of LIE_6, corresponding deviations d_{Fit} and d_{Cal} usually amount to few kJ/mol only. Mean absolute deviation MD (mean absolute error) for these columns are given at the bottom of Table 6.4 along with a squared coefficients of ordinary (Pearson R^2) and Spearman's rank correlation (R_{S}^2) for

fitted and cross-validated energies. Both correlation coefficients are extremely correlated regarding all models. Highest correlations amounting to R^2_{Fit}=0.86 and R^2_{Cal}=0.78 and sketched in Figure 6.5 are associated with the two theoretical coefficients for fitted (Fit) and cross-validated (Cal) free energies of LIE_6. Corresponding mean absolute errors account for 3.8 and 4.8 kJ/mol. These results are only negligibly poorer than those gained from normalized input data but still remarkable in light of the fully automatic approach to binding mode (Chapter 4) and affinity prediction. Neglecting the two structural parameters according to LIE_4 yielded considerably less correlation, R^2_{Fit}=0.67 for fitted and R^2_{Cal}=0.58 associated with cross-validated (predicted) energies. However, the correlations are still significantly better than those of the original two-parameter LIE model having achieved R^2_{Cal}=0.45 which considerably increased to R^2_{Cal}=0.53 if additionally taking into account the ligand's strain energy only. In particular, all LIE models investigated here perform substantially better than MM/PBSA and Vina which attained R^2_{Cal}=0.27 and R^2_{Cal}=0.22. For MM/PBSA calculations a Gromacs implementation by Kumari et al. denoted as g_mmpbsa[189] was applied to the same set of 60 complex trajectories as had been used for LIE models, though, the mode associated with the lowest overall MM/PBSA binding energy ΔG instead of lowest interaction energy was further evaluated. In contrast, feeding MM/PBSA with favorable binding modes according to interaction energies yielded even worse results (R^2_{Cal}=0.1). One should, however, note that all MM/PBSA calculations have been carried out on the basis of host–guest complexes rather than using the unbound ligand in addition. The state of the art docking tool AutoDock-Vina[98] was utilized with largely analogous settings (60 modes, same partial charges, grid box centered on the geometric center of the co-crystallized ligand E_2 and large enough to capture all ligands of the training set) and with a high exhaustiveness value of 100. These results clearly confirm the superiority of average-based MD methods on the basis of systematically chosen initial binding modes over single step methods (docking) with randomly proposed and evaluated binding modes even though the computational effort increases substantially.[49,57,90,91] Moreover, considering thermodynamic end state methods only, any LIE equation performs considerably better than the continuum-solvation model thereby confirming recent studies.[77]

Regarding the evaluation of our LIE_m model's predictive power, we have been relying on the LOOCV method. In related literature, the (squared) coefficient of cross-validation we have termed R^2_{Cal} due to Table 6.4 is usually denoted as q^2_{Loo}. Some scientists regard the use of it with suspicion if the model has not been applied to further test sets.[253] Nevertheless, meeting a couple of criteria substantially increases its reliability: the size of the training set (here amounting to 31) is recommended to be a manifold of the number of descriptors (six in the presented model) and the set itself should vary

strongly in both chemical scaffolds as well as binding affinities in order to be capable for substantially differing compounds. For the given set of ligands, it ranged over nearly 10^7 magnitudes. Apart from a high value of the coefficient R^2_{Cal} of cross-validation, a reliable predictive model is characterized by a regression line that hardly diverges from the bisecting line (regarding both slope and intercept) since the two plotting axes represent the same entity ΔG as depicted in Figure 6.5.[253] By these means and particularly considering the automated beforehand prediction of an optimal binding mode, the overall performance of this six parameter model is much more than satisfying, though, for isolated chemicals, strongly erroneous predictions were made. For instance, the relative binding affinity of one of the most flexible ligands, 1-Nonylphenol, equipped with eight freely rotatable bonds as well as a hydroxyphenyl functionality was estimated about a hundred times higher than what is known from laboratory experiments. Having omitted 1-Nonylphenol would have increased the cross-validated LIE_6 correlation up to R^2_{Cal}=0.81. Instead, the original two-parameter LIE model (LIE_2) incorporating only van der Waals and electric interaction energies yielded less than R^2_{Cal}=0.5. A substantial increase to R^2_{Cal}=0.67 was achieved by simply adding the hydroxyphenyl indicator.[194] The most reasonable cause coming into question for both poor prediction of 1-nonylphenol as well as the significance of the hydroxyphenolic indicator is

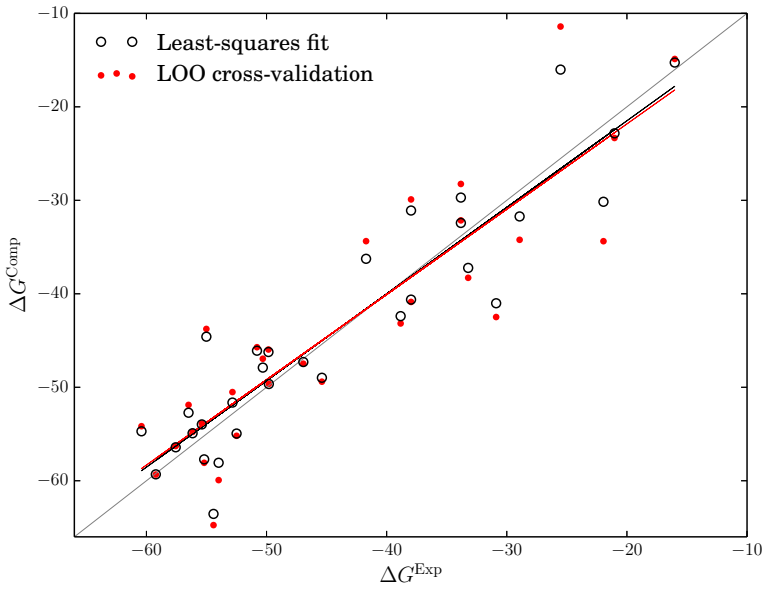


Figure 6.5: Correlation of both fitted and cross-validated versus experimental binding free energies (given in [kJ/mol]) associated with 31 chemical compounds and ERα using a six-parameter empirical linear interaction energy-based approach. Corresponding regression lines and bisecting lines are sketched as well.

the absence of a water molecule acting as bridge for several hydrogen bonds with the target-specific phenolic hydroxy group and two AAs during MD simulation. Since this notable physical effect is not properly considered, the boolean hydroxyphenolic indicator becomes overestimated regarding its weight. In turn, the binding affinity of compounds (such as assumed for 1-nonylphenol) which most likely do not reveal the typical complex hydrogen bonding network as known in case of all steroid-based ligands equipped with a respective hydroxy group is overvalued. As already stated, other LIE or QSAR type predictive models for the estimation of binding affinities to ERα either yielded poor cross-validation coefficients[242] or included the manual selection of an initial binding mode (or according to crystal structures).[84,109,243]

6.5 Evaluation of MD Settings and Parameters

The above presented linear model incorporates non-physical parameters related to structural properties. Thus, the model cannot be considered as entirely physical relying on statistical averages only. If one prefers a purely physical approach such as

$$\Delta G^{\text{comp}} = x_1 \Delta \left\langle E^{\text{elec}} \right\rangle + x_2 \Delta \left\langle E^{\text{vdW}} \right\rangle + x_3 \Delta \left\langle U^{\text{lig}} \right\rangle + x_4 T \Delta S^{\text{conf}} \tag{6.6}$$

after having omitted the two additional structural descriptors as incorporated by Equation 6.2, the model's predictive power decreases significantly to R^2_{Cal}=0.58 which is still an acceptable value. As already stated, the considerable decrease is most likely caused by the absence of a water molecule at the binding site usually forming several hydrogen bonds. Consequently, its effect is compensated by the phenolic hydroxy indicator revealing an increased significance. On the basis of two independent MD runs (indicated by solid and dashed lines, respectively) of the entire set of compounds and according to the purely thermodynamic model represented by Equation 6.6 we want to briefly investigate the effect of few MD parameters on R^2_{Cal}. Interestingly, as illustrated by Figure 6.6, correlation coefficients of the two independent runs (solid and dashed lines, respectively) starting from the same initial state develop very similarly. This is well reflected by a squared coefficient of 0.8 regarding the correlation of the two sets of R^2_{Cal} values. The average deviation between any two corresponding values was only 0.026. What one may conclude from this observation is that, possibly due to steric constraints, the dynamics of a host–guest system starting from a favorable binding mode and observed for less than 1 ns is likely to repeat regarding host–guest interaction. Further, the influence of the evaluated MD time range was moderate in case of MD simulations *without* position restraints (indicated by the string "noPR") in combination with an optimal binding mode selection according to Coulomb interactions only (indicated by

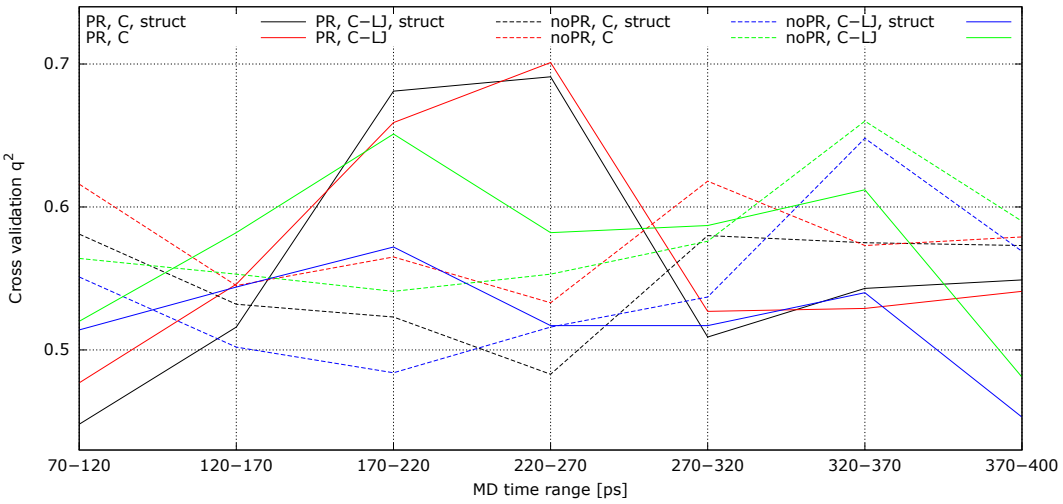


Figure 6.6: Running squared coefficients of leave-one-out cross-correlation depending on MD settings and calculated upon two independent runs (represented by solid and dashed lines) associated with the same initial modes: C_α atoms either positionally restrained (PR) or not (noPR), preferential binding mode selected either according to Coulomb (C) or, in addition, van der Waals (C-LJ) interaction energy terms.

the letter "C") as carried out for the linear models represented by Equations 6.2 and 6.6 (blue lines). A similar trend within a similarly narrow range of R^2_{Cal} values corresponding to red lines was observed for optimal binding modes chosen according to Coulomb *and* van der Waals interactions (indicated by "C-LJ") out of 60 systems per ligand simulated *with* PR ("PR"). Interestingly, the plots of both strategies (PR/C-LJ and noPR/C) starting at relatively high values exhibit an initial decay followed by a rise during the second half of the abscissa. In contrast, a contrary trend was observed if having combined "PR" with "C" (black lines) or, respectively, "noPR" with "C-LJ" (green lines). These two approaches are characterized by an initial rise followed by a decay during the second half of the abscissa. Moreover, R^2_{Cal} values of the latter two approaches are spread over a significantly larger range (up to approximately 0.25) showing large changes whereas values in case of the former two strategies (red and blue lines) range within approximately 0.15. In light of these results, there seems to be some correlation between the analytic strategies. If planning to apply PR to an MD production run, its seems, during the first 150 ps, more appropriate to determine optimal binding modes according to the sum of Coulomb and van der Waals interaction energies, whereas during the subsequent approximately 150 ps, it seems better to neglect van der Waals forces upon binding mode selection. If, in contrast, the application of PRs is not planned, the effect of the set of interaction energy terms used for binding mode selection turns to the opposite. However, further investigations including much more simulations would be necessary in order to figure out whether the relationship described here is based rather

on coincidence or causality. It should be noted that the application of PRs on certain particles as implemented by the Gromacs software does not entirely prevent their mobility. Rather, their masses are increased by a factor amounting to 1000 per default such that corresponding particles are substantially less affected by interatomic forces. Figure 6.7 illustrates the effect of PRs associated with C_α atoms on the time-dependent RMS deviation of all ER_α atoms (averaged over the entire set of 31 ligands as well as all binding modes). Qualitatively, both RMSD curves (with and without PRs) reveal the same progress following the root function, though, the red curve associated with unrestrained systems grows significantly faster. In turn, RMSD values of the restrained system are delayed in time.

Nevertheless, for each of the four investigated analytical approaches (regarding the presence of PRs in combination with particular energy contributions chosen for the binding mode identification), R^2_{Cal} of the LIE_4 more or less consistently exceeds 0.45 in contrast to the original LIE approach itself. The latter yielded respective squared coefficients ranging between 0.29 and 0.48 (not shown in Figure 6.6) that are always below corresponding LIE_4 values. Using a state of the art docking tool, we have confirmed that average-based MD methods on the basis of systematically chosen initial poses are superior over random single step methods. Nevertheless, for the purpose of saving computational time, we considered the question whether binding pose estimation solely on the basis of the lowest interaction energy out of 60 local energy minima (used as initial structure for MD runs) yields similar results as the one incorporating MD trajectory averages presented above. However, it turned out that for most ligands these optimal binding modes did not well agree with the average-based approach as illustrated by Figure 6.8 in case of the E_2–ERα complex. It shows, on the one hand, that each of the four subsequent time ranges (colored lines) of 100 ps length yielded another favorable mode (lowest of 60 interaction energies) as represented by a circle that

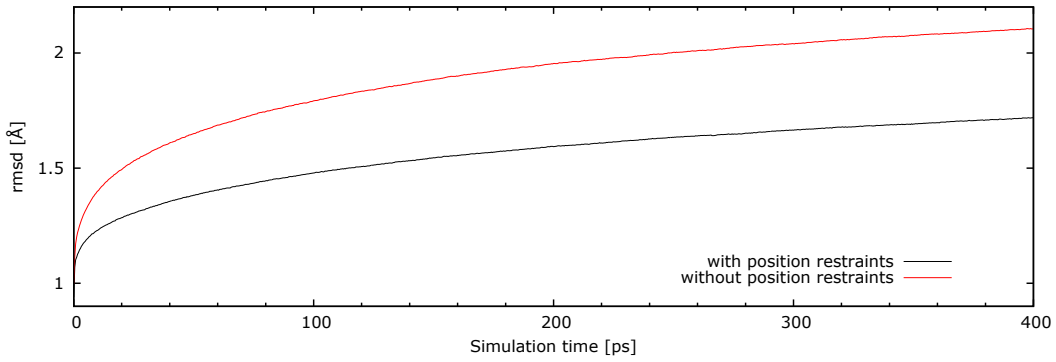


Figure 6.7: Root mean square deviation of protein backbone during MD with and without position restraints.

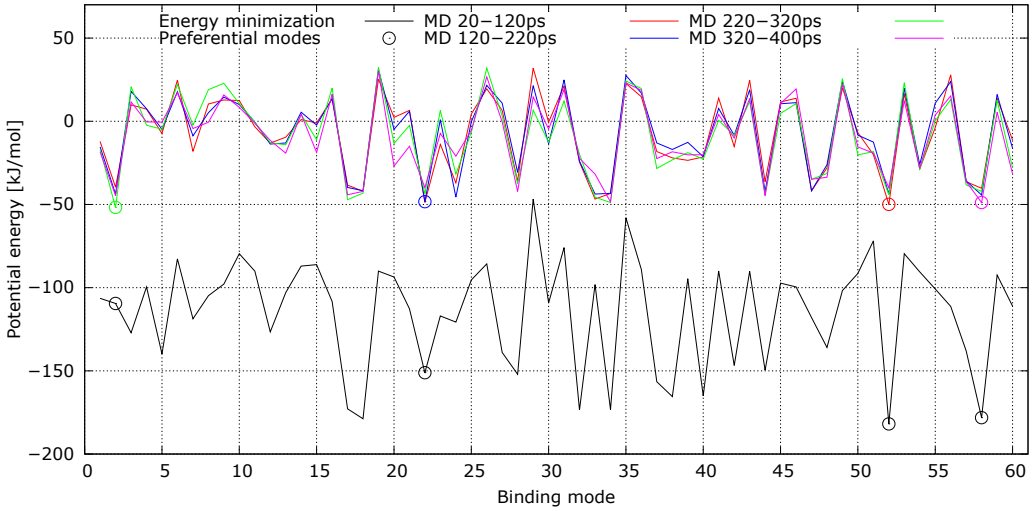


Figure 6.8: Comparison of E_2–ERα interaction energies regarding 60 MD simulations starting from uniformly distributed initial binding modes. **Black plot**: local energy minima, **colored plots**: time-averaged energies associated with four different MD time ranges, **circles**: optimal binding modes according to MD time ranges.

is colored in accordance with the respective time range. Already this observation indicates difficulties with the selection of a single representative mode. Considering on the other hand minimal energies of respective modes obtained from the preceding energy minimization procedure reveals significant discrepancies. That is two, namely mode 2 (green circle, range 220-320 ps) and mode 22 (blue circle, range 120-220 ps), of the four binding modes suggested by averages of four consecutive MD time ranges were far away from the lowest minimal energy binding structure which is associated with mode 53. However, the latter mode related to the lowest MD energy average of the first time range 20-120 ps (red circle) as well as mode 58 (magenta, range 320-420 ps) correspond to the two lowest energy minima among the 60 binding modes. Nevertheless, the order of local energy minima does not necessarily (and sufficiently) correlate with the order of lowest MD averages. On the basis of a set of optimal binding modes selected according to energy minima, for no combination of parameters as addressed by Figure 6.6, a R^2_{Cal} value larger than 0.3 was observed.

6.6 Concluding remarks

This chapter dealt with the development of an empirical linear model for the prediction of ERα host–guest binding affinities based on classical MD simulations. As both methods analyze the same trajectory data, an LIE-based affinity model constitutes the

consequent next step of the preceding binding mode estimation described in Section 4.3. Model development and parameter fitting were carried out on the basis of 31 highly diverse ligands associated with reaction constants ranging over 10^7 magnitudes. Due to the large range of affinities and chemical scaffolds and since the size of the training set exceeded the number of model parameters by its fivefold, overfitting must be considered minimal. The strategy basically extending the original one/two-parameter LIE method incorporates two additional physical and two structural parameters apart from interaction energy terms. Both squared coefficients for the fitted data and the more meaningful leave-one-out cross-validation of predicted energies were elevated up to values around 0.8 in case of normalized as well as unnormalized data which is remarkable in light of a fully automated process. In this regard, it is superior to most other predictive models for the estimation of binding affinities to ERα which suffer either from poor cross-validation coefficients, a manual selection of initial binding modes, or a less divers set of scaffolds. Using the original LIE parameter set (Coulomb and Lennard-Jones potential) yielded a squared LOOCV coefficient significantly less than 0.5. However, on the given set of compounds and poses all LIE models performed substantially better than MM/PBSA as a representative of the continuum-type of end state methods.

One of the additional physical model parameters is related to conformational entropies. For a rough estimation of a ligand's entropy loss upon target binding a Monte-Carlo approach to the variance of atomic coordinates was implemented that had been published few years before. Considering conformational entropies significantly increased the original LIE model's predictive power. Choosing the coordinates' standard deviation on the basis of both bound and (the mean of) 60 unbound modes as an RMSD cutoff ensures a high discrimination of their conformational entropies. In contrast to the original entropy model, no internal but Cartesian coordinates had been used such that, beside conformational changes, translational and rotational degrees of freedom were captured by this entropy method as well. The second additional physical descriptor significantly increasing the model's predictive power is related to ligand strain energy that is the potential energy uptake upon binding. The introduction of two indicators for the presence of a benzene and, particularly, hydroxyphenyl functionality had further improved the model. Indeed, the functional group associated with the latter descriptor is involved in several hydrogen bonds partially mediated by a water molecule nearby such that its absence during simulation is compensated by the significant phenolic hydroxy group.

Neglecting both structural descriptors and considering only thermodynamic parameters still yields significantly higher coefficients of cross-validation exceeding 0.5. Interestingly, a comparison of running coefficient values of two entirely independent MD

runs reveals some kind of robustness of the presented average-based model insofar that their trends emerge very similarly with an absolute average deviation of R^2_{Cal} values less than 0.03. Considering the first 300 ps of host–guest MD trajectories, there seems to be some relationship between two MD parameters: the presence of position restraints on the one side and the choice of an initial binding mode on the other. The trend of runs lacking PRs but associated with binding modes selected according to the Coulomb potential is analogue to the trend of simulations with PRs and orientations chosen on the basis of both Coulomb and LJ potentials (C/LJ). Both trends start with an initial decay followed by an increase. An opposing trend (growth first, then decrease) was observed in case of an inverse setting combination, namely for PRs combined with C/LJ binding modes as well as an for MD runs without PRs but binding modes on the basis of Coulomb interactions. In addition, cross-validated correlation coefficients of the latter pair of approaches are spread over a significantly larger range. Whether this observed relationship is based on coincidence or not requires further and systematic investigations using many copies of independent MD trajectories.

Much worse results have been achieved using methods relying on a single state of the molecular complex. Taking into account only geometries associated with the lowest minimal energy out of 60 local energy minimizations per ligand, mostly lead to binding mode propositions different from MD average results and an insufficient correlation with experimentally determined binding affinities. Having chosen analogous settings (number of modes, identical partial charges, position of the grid box), the AutoDock-Vina docking software resulted in a squared correlation coefficient less than 0.25 which is acceptable neither for drug design (primarily avoid false positives) nor for toxicity estimations (primarily avoid false negatives). These results highly conform with other critical evaluations of molecular docking techniques. Of course, the computational effort of less than one minute per ligand for 60 binding modes on six modern CPU cores is negligible compared to approximately 20 minutes for 60 MD trajectories using about 1500 CPU cores. As a consequence, for a reliable prediction we strongly recommend any free energy of binding calculation on the basis of several as heavily as systematically distinct complex geometries, although this is accompanied by massive calculations. By the way, due to a permanent progress in terms of computer hardware, parallelism, and software development we can expect that the time required by an MD simulation as described above will drop from 20 to one or two minutes within the next ten years.

As already pointed out in Section 4.3, the comparison of predicted ligand orientations with crystallographic data retrieved from the Protein database pdb.org revealed remarkably reliable binding mode predictions. Hence, this chapter serves as an additional evaluation for that strategy since high correlations between experimental and

predicted binding affinities as achieved with the six-parameter model as well imply that binding modes were chosen correctly. All told, a promising hands-off approach to the prediction of highly accurate binding modes and affinities for molecular host–guest systems has been presented that will be capable for high-throughput screenings in a couple of years.[194] Similar to easy to handle docking tools, the algorithm does not require other manual operations than the definition of a spatial vector specifying a particular target's binding site and an arbitrary set of coordinates representing the drug-sized compound.

7 Risk assessment on sulfamethoxazole transformation products

We became acquainted in the previous chapter with an accurate predictive model useful for estimating absolute binding affinities to ERα. However, the purely empirical approach requires training data that is not inevitably available for arbitrary target molecules. Nevertheless, for reasons stated above one might prefer an extended LIE model over other binding free energy or docking methods. In this short chapter we will therefore elaborate a simple strategy enabling us to use the entirely physics-based four-parameter LIE_4 model (Equation 6.6) presented in the previous chapter without having trained any weight coefficients in advance. As a consequence, rather than calculating absolute binding free energies the presented method was used to prioritize chemical compounds with respect to their protein binding affinities in relation to some reference substance. All calculations have been carried out on the basis of a set of 29 documented *transformation products* (TPs) arising from biotic or abiotic degradation of the antibiotic sulfamethoxazole (SMZ) which served as reference.

7.1 Introduction

Over the past century, considerable amounts of various anthropogenic compounds including drugs, synthetic materials, and other chemicals have been released to the environment via the route of waste and waste water. Here, many of them accumulate due to their persistent character, probably after metabolization through biotic or biotic processes. They have been measured in many ecological compartments particularly including drinking water.[254–256] Certainly, industrial products have to undergo an environmental risk assessment in order to obtain approval, but the risk arising from metabolites is typically not accounted for during this process, although, the endocrine potential of wastewater[257] as well as growth-inhibiting effect of drinking water[258] have already been reported on the basis of *in vivo* tests. However, in recent years, first and higher order TPs as well increasingly attracted human and ecotoxicologists'

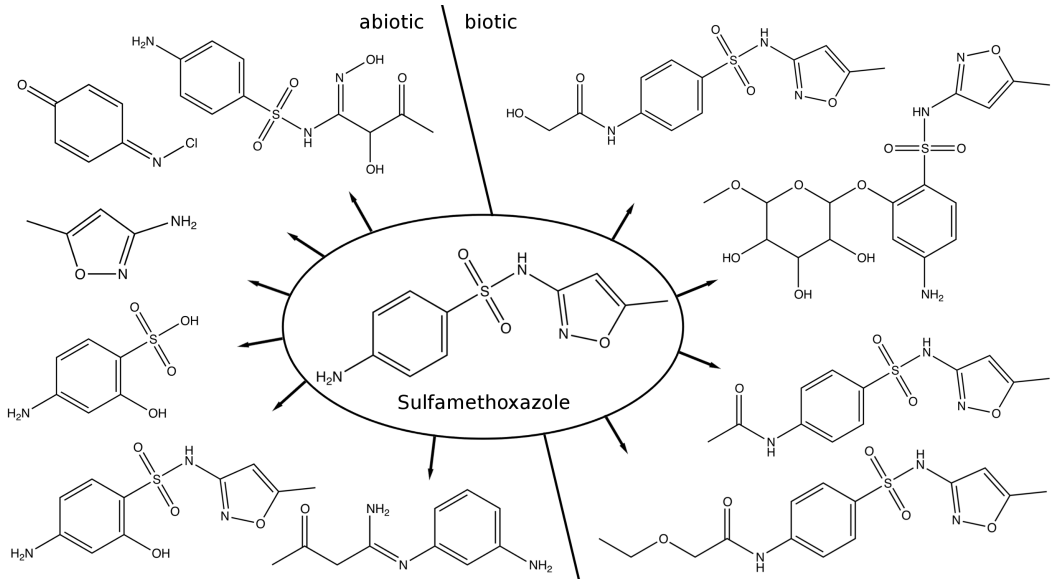


Figure 7.1: The antibiotic sulfamethoxazole and several documented transformation poducts arising from biotic or abiotic degradation.

attention.[259] Dozens of degradation products have been detected for some chemicals such as SMZ[260–263] or carbamazepine.[264–268] Some of them arise from elimination of the parental substance in sewage treatment plants or from biodegradation in humans. Anyways, an ensuing potential increase in the substance's environmental risk cannot be entirely ruled out.[258,259] Often, the amounts detected in the environment are too small for the sake of conventional laboratory screenings such that, as pointed out earlier, computer-aided risk assessment provides the only access to toxicity estimation for such newly discovered metabolites. Furthermore, the applicability of conventional methods is technically limited by the vast number of possible interactions between TPs and biological target structures (receptors, enzymes, ion channels etc.) coming into question. The task's complexity further increases if including TPs whose formation was predicted by *in silico* methods.[259]

In the following we will use an LIE-based method for the toxicological prioritization of TPs derived from the antibiotic SMZ (center molecule in Figure 7.1). SMZ belonging to the class of sulfonamides is directed towards Gram-negative and Gram-positive microorganisms and made use of since the middle of the 20th century. The suppressive effect on bacterial synthesis of folic acid is due to its structural analogy to *para*-aminobenzoic acid (PABA) resulting in a competitive inhibition of PABA binding to the enzyme dihydropteroate synthase (DHPS) which is not expressed in humans. As a consequence, DHPS looses its ability to synthesize the folic acid precursor dihydropteroic acid by covalently connecting PABA with dihydropteroate diphosphate. Since folic acid

Table 7.1: SMILES notation and CAS number of SMZ (000) and 29 transformation products (001-029) documented in various publications. N/A: not available.

Nr.	transformation product	SMILES notation	CAS number
000	Sulfamethoxazole (SMZ)	CC1=CC(NS(=O)(=O)C2=CC=C(N)C=C2)=NO1	723-46-6
001[a]	5'-Hydroxy-SMZ	C(O)C1=CC(NS(=O)(=O)C2=CC=C(N)C=C2)=NO1	34245-10-8
002[b]	N4-Hydroxylamino-SMZ	CC1=CC(NS(=O)(=O)C2=CC=C(NO)C=C2)=NO1	114438-33-4
003[a]	N4-Acetyl-SMZ	CC1=CC(NS(=O)(=O)C2=CC=C(N(C(=O)C))C=C2)=NO1	21312-10-7
004[a]	N4-Acetyl-5-hydroxy-SMZ	C(O)C1=CC(NS(=O)(=O)C2=CC=C(N(C(=O)C))C=C2)=NO1	75144-40-0
005[c]	N4-Ethoxyacetyl-SMZ	CC1=CC(NS(=O)(=O)C2=CC=C(NC(=O)COCC)C=C2)=NO1	21662-79-3
006[b]	N4-Nitroso-SMZ	CC1=CC(NS(=O)(=O)C2=CC=C(N(=O))C=C2)=NO1	131549-85-4
007[c]	N4-Glycolyl-SMZ	CC1=CC(NS(=O)(=O)C2=CC=C(NC(=O)CO)C=C2)=NO1	51729-63-6
008[a]	N1-Acetyl-SMZ	CC1=CC(NS(=O)(=O)C2=CC=C(N)C=C2)=NO1	18607-98-2
009[a]	SMZ N1-glucuronide	CC1=CC(N(C2C(O)C(O)C(O)C(OC)O2)S(=O)(=O)C2=CC=C(N)C=C2)=NO1	14365-52-7
010[d]	SMZ-2'-glucuronide	CC1=CC(NS(=O)(=O)C2=CC=C(N)C=C2)N(C3C(O)C(O)C(O)C(OC)O3)O1	16854-34-5
011[e]	SMZ-2-glucuronide	CC1=CC(NS(=O)(=O)C2=C(OC3C(O)C(O)C(O)C(OC)O3)C=C(N)C=C2)=NO1	37393-49-0
012[f]	N4-chloro-SMZ	CC1=CC(NS(=O)(=O)C2=CC=C(NCl)C=C2)=NO1	151928-89-1
013[f]	o-chloro-SMZ	CC1=CC(NS(=O)(=O)C2=CC=C(N)C(Cl)=C2)=NO1	151928-90-4
014[f]	N-chloro-p-benzoquinoneimine	C1(=NCl)C=CC(=O)C=C1	637-61-6
015[g]	Benzenesulfonamide	CC1=CN=C(NS(=O)(=O)C2=CC=C(N)C=C2)O1	51821-47-7
016[g]	Butanimidamide	CC(=O)C=C(N)NS(=O)(=O)C2=CC=C(N)C=C2	210241-73-9
017[g]	$C_{10}H_{13}N_3O_4S$	CC(=O)C(O)C(=N)NS(=O)(=O)C2=CC=C(N)C=C2	N/A
018[g]	$C_{10}H_{13}N_3O$	CC(=O)CC(N)=NC1=CC(N)=CC=C1	N/A
019[h]	2-Hydroxy-SMZ (Int.1)	CC1=CC(NS(=O)(=O)C2=C(O)C=C(N)C=C2)=NO1	N/A
020[h]	3-Hydroxy-SMZ (Int.1)	CC1=CC(NS(=O)(=O)C2=CC(O)=C(N)C=C2)=NO1	N/A
021[h]	$C_{10}H_{13}N_3O_5S$ (Int.2)	CC1(O)C(O)C(NS(=O)(=O)C2=CC=C(N)C=C2)=NO1	N/A
022[h]	3-Isoxazolamine (Int.3)	CC1=CC(N)=NO1	1072-67-9
023[h]	$C_{10}H_{13}N_3O_5S$	CC(=O)C(O)C(=NO)NS(=O)(=O)C2=CC=C(N)C=C2	N/A
024[h]	$C_{10}H_{12}N_2O_5S$	CC(=O)C(O)C(=O)NS(=O)(=O)C2=CC=C(N)C=C2	N/A
025[h]	$C_6H_7NO_3S$	OS(=O)(=O)C2=CC=C(N)C=C2	N/A
026[h]	$C_4H_8N_2O_3$	CC1(O)C(O)C(N)=NO1	N/A
027[h]	2-Benzenesulfonic-acid	OS(=O)(=O)C1=C(O)C=C(N)C=C1	146117-42-2
028[h]	3-Benzenesulfonic-acid	OS(=O)(=O)C1=CC(O)=C(N)C=C1	53819-11-7
029[f]	Azo-SMZ	CC1=CC(NS(=O)(=O)C2=CC=C(N=NC3=CC=C(S(=O)(=O)NC4=NOC(C)=C4)C=C3)C=C2)=NO1	97254-40-5

[a] Vree, 1995

[b] Sanderson, 2007

[c] Kaplan, 1973

[d] Ueda, 1967

[e] Ueda, 1972

[f] Dodd, 2004

[g] Mohatt, 2011

[h] Hu, 2007

is essential for the reproduction of microorganisms, only bacterial strains capable of an alternative route to its synthesis would resist SMZ.[269,270] A graphical representation of DHPS in complex with a substrate as well as a product molecule is available in the introduction (Figure 1.7). Its sub-therapeutic occurrence not only in drinking water during the past decades might have caused the increased bacterial resistance associated with Escherichia coli (E. coli) and Staphyllococcus aureus (S. aureus) strains observed by the end of the last century.[271] Anyhow, as a generally accepted rule it is necessary to minimize SMZ exposition to the environment in order to mostly avoid additional increase and development of resistances. Certainly, this holds true for TPs revealing the same bacteriostatic effect. Thus, from a toxicological point of view, it is not sufficient to eliminate only parental substances such as SMZ in sewage treatment plants but also potentially risky TPs. An obvious approach to toxicity assessment entails estimating the free energy of binding to DHPS using *in silico* methods. Substances yielding a theoretically higher binding affinity than SMZ are expected to cause the same inhibiting effect on DHPS. Hence, this chapter aims at a qualitative estimation of TP binding affinities in relation to SMZ. According to the calculations, all metabolites were categorized and prioritized for both toxicological reasons and follow-up laboratory experiments often including the costly development of chemical synthesis protocols. Table 7.1 shows a list of SMZ TPs gathered from numerous publications[260–263,272–275] and investigated throughout this chapter.

7.2 Data preparation and force field simulations

Apart from few deviations described below, target and small molecules underwent the same molecular modeling procedure as described for ERα and estrogens in Chapter 6. Crystallographic structure files of five DHPS proteins originating from different species and listed in Table 7.2 were retrieved from PDB. In order to have knowledge about the substrate binding site, we restricted ourselves only on those PDB entries that included either some co-crystallized substrate or product molecule. After having removed all but protein atoms of the first complete chain (if more than one was available), the enzymes were parametrized in accordance with the AMBER99sb force field that is particularly convenient for explicitly solvated biological systems. SMZ as well as its TPs were, in contrast, sketched, cleaned in 3D, and exported as PDB files using MarvinSketch v5.5. The AmberTool Antechamber was used for the assignment of physical parameters and partial atomic charges according to GAFF and, respectively, the AM1BCC method.

A structural alignment of the five selected enzymes on the basis of a superposition of

Table 7.2: Bacterial origin and PDB id of crystallographic DHPS structures retrieved from the Protein Data Bank and used as targets for molecular dynamics binding calculations of SMZ and its transformation products. The last column is related to the presence of SMZ resistance where +/− indicates that resistant as well as sensitive strains of the corresponding microorganism exist.

Protein source	PDB id	SMZ resistance
Staphylococcus aureus	1AD4 [269]	+/−
Streptococcus pneumoniae	2VEG [276]	−
Bacillus anthracis	3TYE [277]	+
Escherichia coli	1AJ2 [278]	−
Yersinia pestis	3TYZ [277]	+/−

their substrates' pteridine ring system is depicted in Figure 7.2. Each of the two vicinal DHPS binding sites are occupied by a substrate molecule/analog. The one on the right (background) appearing vertically aligned comprises the deeply buried pteridine system at the top and a less buried diphosphate group (associated with violet phosphor atoms) at the bottom attached to it. This co-crystallized substrate stems from one of the underlying PDB entries. At the PABA binding site on the cavity's left (in the foreground) we find a biotic SMZ TP denoted as N4-acetyl-SMZ where it was positioned in place of PABA by means of an alignment of their benzene rings. All other metabolites (including SMZ) were placed following the same rules and resulting in an initial binding mode per compound. Ensuing from this initial mode, 59 further poses were generated according to the icosahedron-based procedure described in Section 4.3 in order to capture the most favorable host-guest binding mode. It should be noted, that all simulations were carried out without the dihydropteroate diphosphate substrate since its presence is not required for the sake of inhibition. Prior to MD simulations, the complex was put in a 10 nm simulation box and explicitly solvated with the tip4pew water model. Formal charges were neutralized by adding a corresponding number of sodium or chlorine ions, respectively. Apart from these 60 bound configurations, one unbound system was generated consisting of the metabolite (along with counterions if charged) and solvent molecules only.

Just like in the previous chapter, a multistage MD simulation was performed, though, using Gromacs in the version 4.6.5. First, the complex underwent 7000 steepest descent energy minimization steps if the maximum force acting on any atom had not ended up below 300 kJ mol^{-1} before. Host and guest molecules were positionally restrained during a subsequent 200 ps equilibration phase in order to achieve a physical temperature-specific distribution of the liquid phase. The simulation temperature was

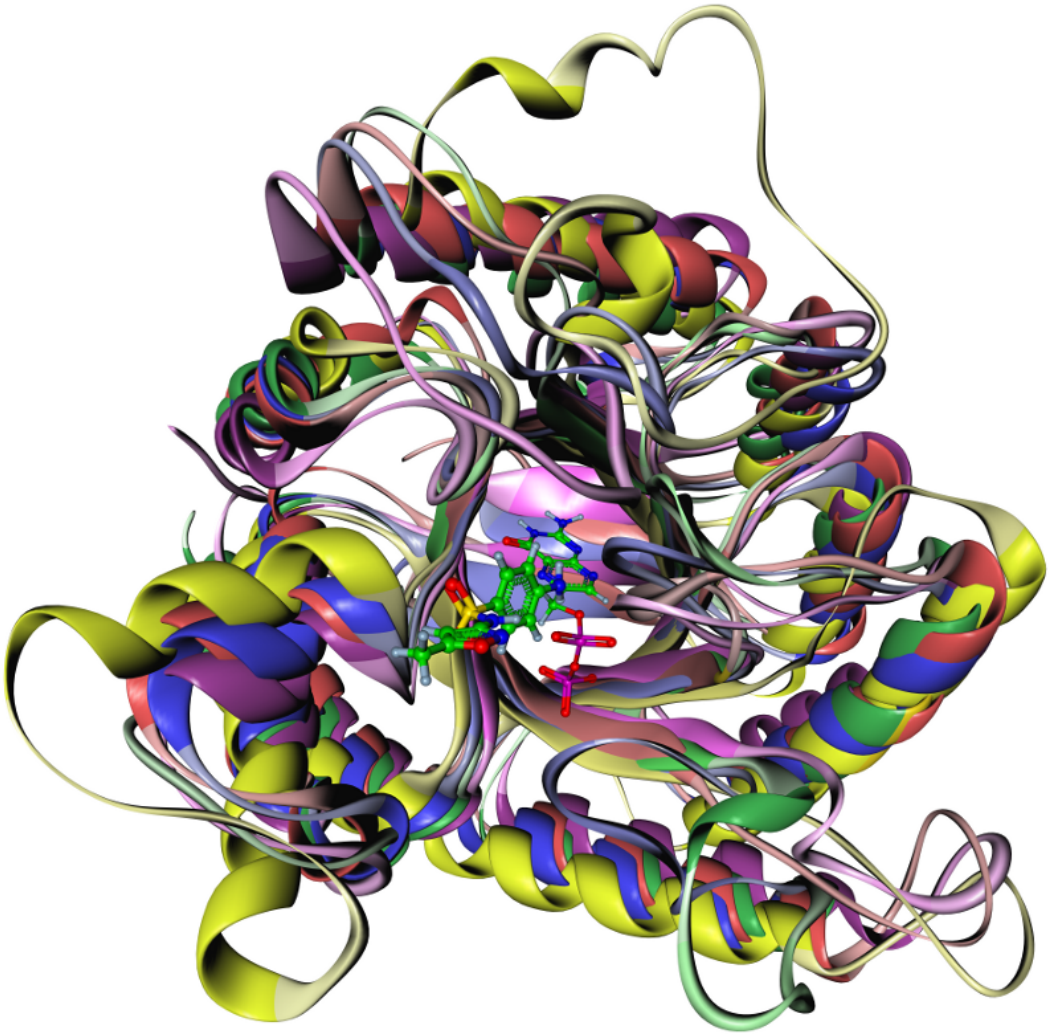


Figure 7.2: Secondary structure depiction of a structural alignment of five isoforms of the enzyme DHPS originating from different bacterial species: **S. aureus** (green), **S. pneumoniae** (yellow), **B. anthracis** (red), **E. coli** (violett), and **Y. pestis** (blue). The binding site shows N4-acetyl-SMZ on the left in place of PABA and, considerably deeper inside DHPS, the natural substrate dihydropteroate diphosphate on the right. Atom colors: **carbon** (green), **oxygen** (red), **nitrogen** (blue), **sulphur** (yellow), **phosphor** (violet), and **hydrogen** (grey). Both ligands are partially overlapping with respect to the amino nitrogen of N4-acetyl-SMZ.

coupled to 293 K by stochastically rescaling atomic velocities. Finally, another 200 ps simulation was carried out as a production run without position restraints but with constraints on all bonds according to the LINCS approach and allowing to increase the discretized time step size from 1 to 2 fs. During this final run, the pressure was coupled weakly using Berendsen's algorithm. Regarding the calculation of interaction energies, cutoff values for Coulomb and Lennard Jones potentials were set to 10 and 14 Å, respectively, where the former was determined according to the smooth particle mesh Ewald summation.

7.3 Prioritization of transformation products

Following the approach described in previous chapters, the favorable one ξ out of 60 binding modes associated with different initial binding modes of each compound was determined on the basis of respective MD time-averages (indicated by angular brackets) of both interaction energy terms

$$\xi = \underset{i \in [1,60]}{\operatorname{argmin}} \left[\Delta \left\langle E^{\text{elec}} \right\rangle_i + \Delta \left\langle E^{\text{vdw}} \right\rangle_i \right].$$

An estimation of TP binding affinities to DHPS isomers relative to the parental compound SMZ was carried out on the basis of four thermodynamic quantities derived from MD trajectories ξ: protein–ligand interaction energies due to electric, E^{elec}, and van der Waals forces, E^{vdw}, as well as the ligand molecule's potential (strain) energy, U, and conformational entropy (loss) upon binding, S. Each descriptor X was expressed as a difference $\Delta X = X^{\xi} - X^{\text{unbound}}$ between the unbound and the optimal bound system ξ. In case of the unbound simulation, the set of atoms surrounding the ligand mainly consists of water whereas the bound system particularly includes an additional target molecule. As illustrated by the two previous chapters, regarding both the binding mode identification as well as the choice of descriptors, the applied strategy had shaped up as robust and significant.

The basic idea behind the prioritization of TPs using a non-parameterized LIE model is illustrated by Table 6.3 in Chapter 6. The sign of a weight coefficient x_i tells us whether the binding affinity is negatively or positively related to the corresponding descriptor. Considering, for instance, only thermodynamic parameters with coefficients $x^{(4)}$ associated with the purely physical 4-parameter LIE model (LIE_4), we notice that all terms except for the entropy carry a positive sign. Applied to the respective free energy model in Equation 6.6

$$\Delta G^{\xi} = x_1 \Delta \left\langle E^{\text{elec}} \right\rangle + x_2 \Delta \left\langle E^{\text{vdW}} \right\rangle + x_3 \Delta \left\langle E^{\text{pot}} \right\rangle + x_4 T S^{\text{conf}}$$

which is what was actually used for the prioritization of TPs, this observation implies that if one of those three energy differences ("bound" energy minus "unbound" energy) increases, the free energy difference ΔG between bound and unbound grows, too. As we already know, states associated with low energies are more favorable than high-energy states according to thermodynamics. In contrast and as expected, the coefficient of the conformational entropy parameter yielded a negative sign in Table 6.3 indicating that high values associated with a high loss of conformational flexibility correlate with lower binding affinities. The observation copes well with the thermodynamic formulation of

Table 7.3: Code for the toxicological prioritization of SMZ transformation products for further assessment according to calculated energies and conformational entropies in relation to SMZ. A plus sign indicates higher binding probability than SMZ according to the corresponding energy contribution.

Toxicological priority	$\Delta \langle E^{\text{elec}} \rangle$	$\Delta \langle E^{\text{vdw}} \rangle$	$\Delta \langle U \rangle$	$T\Delta S$
None	$-$	$\pm$	$\pm$	$\pm$
	$\pm$	$-$	$\pm$	$\pm$
Low	$+$	$+$	$-$	$-$
Medium	$+$	$+$	$+$	$-$
	$+$	$+$	$-$	$+$
High	$+$	$+$	$+$	$+$

the Gibbs free energy comprising enthalpic and entropic contributions (Equation 2.4). Thus, if some degradation product yields lower energy differences *and* a higher entropy difference (lower entropy loss) than SMZ, it will most likely bind with a higher probability than SMZ itself. This situation is demonstrated by the bottom line of Table 7.3 representing chemical compounds associated with highest priorities (red). In general, a $+$ symbol in a particular column indicates a higher binding probability compared to SMZ in terms of the corresponding parameter. Another category corresponding to the lowest priority (yellow) is characterized by both interaction energy terms in favor of a higher TP affinity ($+$) and both ligand-based parameters in favor of SMZ ($-$). If, in addition, exactly one of the two latter descriptors, $\Delta \langle U \rangle$ or $T\Delta S$, militate for TP binding, we arrive at medium priority (orange). Due to two reasons, all priority levels require both interaction energies to favor TP binding: Apart from providing the basis for the original LIE model they have achieved, which is of much greater import, the

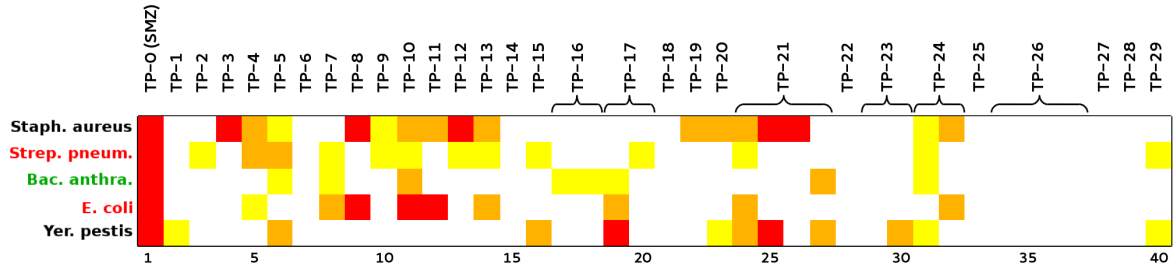


Figure 7.3: Sensitivity matrix for the prioritization of 29 SMZ transformation products (40 corresponding columns if considering stereoisomerism) with respect to their predicted binding affinity to DHPS originating from five bacterial species (rows) that are either SMZ resistant (green), sensitive (red), or strain-dependent (black). Matrix colors: red (highest priority), orange (medium), yellow (lowes), white (no priority since at least one interaction energy term in favor of higher SMZ binding probability).

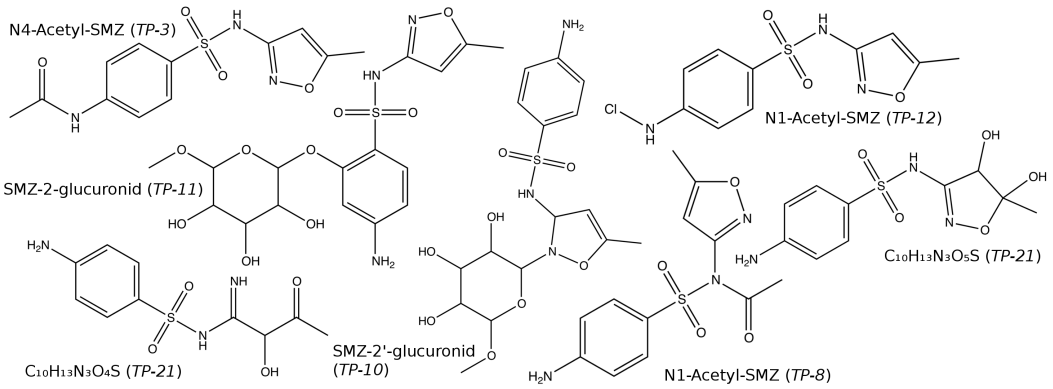


Figure 7.4: Highest priority degradation products of SMZ regarding DHPS binding.

by far highest weights (1.55 and 2.02) of all physical parameters regarding the normalized fitting in Table 6.3. With every additional ligand-based descriptor favoring some TP (associated with a plus symbol), its binding probability and, consequently, priority for a risk assessment increases. Hence, if at least one interaction energy term had favored the reference compound SMZ for binding, the TP was categorized as harmless (green) and no further investigation of bacteriostatic activities in regard to DHPS is recommended.

In practice and depending on the four observable values, all TPs characterized by both interaction energies lower than those of SMZ were considered potentially harmful and distributed over the three categories listed in Table 7.3. This was done with respect to each DHPS target as depicted in Figure 7.3. Across all enzyme isoforms, the sensitivity matrix reveals only seven TPs falling into the category associated with the highest priority. Figure 7.4 shows the chemical structures of these compounds which are recommended for risk assessment with higher priority than any other SMZ TP. For two of them high binding probabilities with respect to DHPS of two different species have been calculated, namely, S. aureus and E. coli in case of N1-acetyl-SMZ (TP-8) and S. aureus and Y. pestis in case of a stereoisomer of a compound with the chemical formula $C_{10}H_{13}N_3O_4S$ (TP-21). These three species are the only ones theoretically affected by metabolites of the highest category. And since each of them either is sensitive to SMZ (E. coli) or comprises resistant as well as sensitive strains (S. aureus and Y. pestis), it is of particular interest to minimize the exposition of such potential binders to the environment in order to reduce the formation of resistances to a minimum. A well-known example for such critical species are multi-resistant strains of S. aureus often found in hospitals and grocery stores. The treatment of increasing resistances to many antibiotics is becoming more and more challenging.[269] According to the presented results, S. aureus indeed seems sensitive to more SMZ metabolites than any other species inves-

tigated in this chapter. Interestingly, DHPS of the only entirely SMZ-resistant species B. anthracis is characterized by the lowest number of high and medium priority (orange) binders. For the sake of completeness, another nine and five compounds fell into the second and, respectively, lowest category considering at least one DHPS isomer where TPs already categorized with higher priorities have been excluded. A comprehensive evaluation of the computational results is currently not possible for lack of laboratory experiments.

7.4 Concluding remarks

Due to an increasing number of novel transformation products a risk assessment and prioritization by means of *in silico* methods becomes more an more indispensable. The limited resources of time-consuming and expensive laboratory experiments including the development of a synthesis protocol as well as *in vitro* and *in vivo* tests for toxicity quantification can then be restricted on metabolites that are most likely to affect critical biological targets. Due to structural similarities between parent compounds and their low order degradation products, one is typically interested in the question whether TPs are able to bind to the same targets as known for the parent compound.

In light of these considerations we developed an MD average-based model for prioritizing degradation products of the antibiotic SMZ with respect to their binding affinity to its bacterial target DHPS. SMZ prevents this enzyme from forming a precursor of folic acid that is essential for folic acid synthesis. Certainly, microorganisms tend to develop resistances against antibiotics when repeatedly exposed to them in sub-therapeutic trace amounts. For that reason, it is of great importance to minimize environmental contamination with potentially bioactive anthropogenic substances. The task was to identify TPs of SMZ that are likely to mimic the DHPS-suppressive characteristic of SMZ.

Due to its high accuracy at moderate computational costs we again decided in favor of the entirely physical four-parameter LIE model presented in the previous chapter. From 60 MD time series associated with different initial binding modes as described earlier the favorable binding mode was chosen according to the minimal sum of both interaction energy terms. Lacking training data with known binding affinities to DHPS, no parameter coefficients could be calculated which would have been necessary for predicting absolute binding affinities. However, since we were primarily interested in the prioritization of TPs relative to their parent compound SMZ rather than in absolute affinities, we easily refrained from weight coefficients. The central idea was to first com-

pare a TP's observable values (the content of matrix A) the entirely physical LIE_4 model comprises of with those determined for SMZ. A categorization of each TP was dependent on which and how many of those parameters were in favor of TP binding rather than of SMZ. That is, whenever one of its energy differences was less or entropy difference was larger than that of SMZ. For all categories of potentially harmful degradation products, both interaction energy differences had to favor TP binding. If only one of these two contributions that particularly correlate with binding free energies had been in favor of SMZ, the TP was classified as harmless.

Accordingly, seven TPs of the highest category (highest risk) have been identified particularly affecting DHPS of the (partially) sensitive species S. aureus, E. coli, and Y. pestis. Some of those highest priority compounds were predicted to bind to several DHPS isoforms, but interestingly none of them seems to bind to the only resistant species under investigation, B. anthracis. According to our calculations, four high-priority metabolites are directed towards S. aureus of which critical multi-resistant strains are known that are hard to suppress with common antibiotics. Hence, one is greatly interested in an as wide as possible elimination of SMZ as well as critical analogs already in properly upgraded sewage plants before they reach ground water. From this point of view, degradation products of the highest priority are strongly recommended for toxicological risk assessment through thorough laboratory tests. Currently, the lack of analogous laboratory results makes the extensive evaluation of computational results difficult.

8 Conclusion and Outlook

The accurate quantification of biological host–guest binding affinities plays an increasingly important role in various scientific fields. Among other applications, toxicologists and drug developers are often interested in binding constants for association/dissociation reactions of novel chemical compounds and biological target molecules. And indeed systematic investigations on transformation products reveal a growing number of novel chemical structures arising from degradation of anthropogenic substances which need to undergo a risk assessment as well. Besides well-established *in vivo* and *in vitro* laboratory experiments, one regularly employs *in silico* methods nowadays that are usually less demanding regarding both time and cost. Despite continuously exponential technological and algorithmic advances over the past decades, molecular simulations still pose a big computational challenge owed to the mathematical complexity of biological macromolecular systems often consisting of hundreds of thousands of atoms. This particularly holds for accurate binding affinity methods based on the time-average over an ensemble of states. As a general observation, the reliability of theoretical results is largely correlated with the demand for computing resources. In Chapter 2 we became acquainted with various fast as well as costly approaches to the estimation of (relative) binding affinities and met some general ideas for the creation of reasonable binding poses. Due to its exceeding complexity, a thorough exploration of some macromolecular conformational space by means of MD simulations is impossible within an acceptable time period. One is therefore strongly reliant on one or more energetically favorable representatives of a host–guest complex in order to obtain plausible free energies. The major goal of this thesis concerns the development of an automatized strategy for an accurate prediction of host–guest binding modes and affinities. Accordingly, the subgoals of this thesis can roughly be divided into two categories dealing with two different aspects of the calculation of host–guest binding affinities derived from classical force field trajectories. On the one side we engaged with the decomposition of conformational space followed by the selection of representative molecular geometries serving as input for free energy calculations. In particular, we focussed on a more systematic generation of binding poses covering a broad range of possible molecular interaction modes and, thereby, rendering the dreaded trapping problem obsolete. On the other side and ensuing from these

input geometries, we developed binding affinity models for a couple of (bio)molecular systems based on the linear interaction energy (LIE) method.

Algorithms and results dealing with conformational space discretization for the purpose of binding affinity estimation were presented in Chapter 4. All methods mainly address the trapping problem typically accompanying MD simulations of complex macromolecular systems. Since there is no way to sample the entire conformational space of large systems within reasonable time, it is all the more necessary to particularly carefully select initial geometries representing energetically most favorable states associated with largest statistical weights. Following a simple two-stage procedure, the first part of that chapter proposed the determination of global minima structures of small molecules such as the six major HBCD stereoisomers. HBCDs are characterized by cyclic structures associated with large energetic barriers and two (topological and stereoisomeric) symmetry levels. Ensuing from an HMC sampling at an artificially high temperature of 1500 K for the sake of easily overcoming even high energetic barriers, each geometry was locally minimized and the one associated with the lowest energy was chosen as global energy minimum conformation. Regarding HBCD stereoisomers, all symmetry levels were perfectly reflected by these global energy minima indicating that five Markov chains with at least 50 k steps (better 100 k) each are sufficient for most drug-sized molecules consisting of up to 50 atoms in order to achieve an extensive sampling of the conformational space. At least two global minimum geometries of the three HBCD diastereomers match respective crystal structures very well.

If more than one representative of a molecule (e. g. a small subset of an HMC sampling) is desired for further analysis, the novel linearly scaling *maxdist* cluster algorithm provides an as fast as simple way to identify multiple heavily distinct (related to torsional distances) conformations widely representing the entire conformational space. Insofar, it differs from existing cluster algorithms such as k-means which are particularly designed to identify cluster centers located in densely populated areas. Due to its successive selection of representatives, two stopping criteria are conceivable with *maxdist* depending on the user's preferences and yielding the same sequence of centers: a predefined number of clusters and/or a maximum allowed reference distance $d_{\max}$ of any frame to its nearest representative. However, regarding the trapping issue associated with MD simulations, one would rather decide in favor of a maximum torsional distance criterion approximately amounting to $\pi/3$ in order to achieve considerably overlapping ensembles of states if using those representatives as initial geometries for samplings. Such independent samplings are in turn useful for free energy reweighting and the investigation of molecular kinetics. The expected number of representatives can be roughly estimated through a hyperbolic function of $d_{\max}$ whose power is specified by

the number of torsional degrees of freedom minus some value of α. α was interpreted as a measure for the deviation from an ideal distribution due to molecular constraints since a large value decreases the number of centers necessary to meet the maximum distance requirement. A way to a rigorous calculation of α was not in the focus of this thesis and remains a task for the future. On the basis of pentane and HBCD, the number of centers predicted by the hyperbolic model achieved remarkable coefficients of the correlation with the actually clustered number.

A last space-decomposing strategy intended to remedy the trapping issue is related to simulations of host–guest complexes. In contrast to state-of-the-art docking tools as the only current access to the automatized generation of binding poses, the presented approach systematically decomposes the space of binding modes into 60 more or less uniformly distributed rotational poses of the global minimum geometry which serve as initial conformations for MD runs. Again, the underlying idea is to cover the most of the space of binding modes for the purpose of free energy estimations, thereby striving after some overlap between the distributions of two neighboring poses. Using exactly 60 modes in accordance with the icosahedron's order of symmetry guarantees that the rotational overall distance between any two neighboring poses remains close to $\pi/3$ and that no significant domain of the binding mode space is slipped. Thus, compared to ordinary docking tools, the presented algorithm exhibits two major differences: first, the optimal binding mode is chosen on the basis of time-averages rather than on a single frame. Secondly, proposed poses are, in contrast to docking, constructed *without* knowledge about the ligand binding site such that unphysical atomic collisions between host and guest are most likely to occur depending on the shape and size of both cavity and ligand. However, since one is usually interested in the most favorable binding mode according to energy averages, unphysical states associated with extraordinarily high energies quickly become irrelevant. Using crystallographic structure files of the hormone receptors ERα and EcR available at PDB and in complex with seven different ligands yielded excellent predictions. Considering only the optimal binding mode proposal for each ligand, more than 50 % of them achieved RMS deviations from corresponding reference structures mostly far below one Å. Only the largest and most bulky ligand tamoxifen yielded an RMSD value significantly higher than 1.5 Å that is regarded as a common limit for excellent predictions. In spite of its computational demand regarding numerous MD simulations per host–guest combination, the promising approach seems reasonable for binding pose prediction in particular since the produced conformational data serves as basis for binding affinity calculations as well. However, the computational demand can be significantly reduced if the number of initial binding modes is narrowed in advance regarding symmetric and physically intractable poses.

A first application of the space discretization strategies developed in Chapter 4 was described in Chapter 5 where an LIE-based model as originally designed for protein–ligand systems has been transferred to high-performance liquid chromatographic (HPLC) separation of highly similar HBCD stereoisomers on a chiral stationary phase denoted as β-pmCD. Using this barrel-shaped matrix which exhibits molecular cavities large enough for drug-sized molecules, we mainly focussed on the prediction of the elution order and corresponding retention times of HBCD isomers. Instead of simulating a solvent gradient comprising the two solvents water and acetonitrile (ACN) used for HPLC separation, all calculations were carried out once in each solvent. Average host–guest center of mass distances, steady state characteristics as well as consistency analysis clearly indicate that HBCD prefers to reside in ACN rather than in water. The observations are in very good agreement with HPLC results where indeed elevated concentrations of water at the beginning enhance β-pmCD–HBCD interaction and separation whereas an increasing fraction of ACN favors the elution. Accordingly, the correct elution order was nearly perfectly reproduced with pure water on the basis of electric and Lennard-Jones interactions, whereas ACN simulations failed with any combination of energy contributions. The elution order turned out to be best approximated by time-averages rather than a single lowest-energy frame of each stereoisomer. The results also reveal that averaged energies associated with the optimal binding mode out of 60 are slightly superior to the Boltzmann-weighted sum of all 60 averages which in turn implies that the space of binding modes was sufficiently decomposed and explored using 60 initial poses. Their predicted overall elution order deviates from the correct one to the smallest possible extent, that is, the two isomers associated with the smallest energy difference were swapped. Enantiomeric separation was correctly predicted by each of the three strategies (regarding water as solvent). Moreover, the results are consistent with respect to the time range under consideration. The presented empirical linear model is useful whenever the experimental assignment of peaks to chemical compounds is impossible or for the optimization of suitable stationary phases for a given mixture of compounds.

A similar approach was undertaken in Chapter 6 in the course of the development of a novel estrogen receptor α (ERα) model that complements the optimal binding mode identification presented in Section 4.3. The original two-parameter LIE model comprising only interaction energy terms was extended by two additional physical and two structural parameters resulting in an LIE-QSAR hybrid. Fitted and, particularly, predicted binding free energies of the resulting six-dimensional model (LIE_6) yields excellent correlations with experimental binding affinities that are remarkable in the light of the fully automated process. The LIE_6 model must be considered robust noting high coefficients of cross-validation and due to the set of 31, a multiple of the number of

model descriptors, highly diverse ligands associated binding constants ranging over as many as 10^7 magnitudes. It is superior to other predictive models for the estimation of binding affinities to ERα which suffer either from poor cross-validation coefficients, a manual selection of initial binding modes, or a less divers set of scaffolds. Having extended the original LIE descriptor set by two additional thermodynamic parameters, the ligand's uptake on strain energy and entropy loss upon association, significantly increased the model's predictive power. Taking in addition into account two ERα-specific structural parameters addressing the presence of a benzene and particularly hydroxyphenyl functionality further improved the model. Since the latter group is involved in several hydrogen bonds partially mediated by a water molecule in close vicinity, it is safe to assume that this parameter compensates the absence of that water molecule during MD simulations. Both extended models (with and without structural descriptors) yielded an outstanding agreement with lab results particularly outperforming the original LIE method, MM/PBSA, AutoDock-Vina, and an own single state approach based on 60 minimum energy conformations. Insofar, these observations serve as an additional evaluation for the binding mode estimation strategy since high correlations between experimental and predicted binding affinities as achieved with the six-parameter model as well imply that binding modes were chosen correctly.

For the special case of TPs arising from biotic or abiotic degradation of anthropogenic substances and increasingly detected in various environmental compartments, the entirely thermodynamic LIE_4 model was adapted to their toxicological risk assessment in relation to the parent molecule. Their inherent structural similarity with the parent compound suggests a similar mode of undesirable biological interaction. Thus, in order to figure out whether subtherapeutic amounts of certain TPs of the antibiotic SMZ are likely to cause bacterial resistances against SMZ, qualitative binding affinities of a set of 30 documented TPs of the antibiotic SMZ to its biological target DHPS have been determined. A convenient training set of DHPS ligands for the estimation of LIE weights was missing. The weights' signs, however, can be deduced from thermodynamic relationships. Thus, depending on which and how many LIE descriptors attested a TP binding probability higher than in case of SMZ itself, all TPs were divided into three groups with increasing priority. Across five DHPS isoforms associated with different bacterial species, seven high affinity TPs have been identified that fall into the highest category (all parameters attest TP binding). Interestingly, all affected DHPS isoforms originate from species comprising SMZ-sensitive strains such as S. aureus against which four metabolites are directed and of which critical multi-resistant strains are known that are hard to suppress. These highest priority TPs, that are likely to mimic the DHPS-suppressive characteristic of SMZ and therefore exert selection pressure on respective

strains, are primarily suggested for toxicological risk assessment such that critical candidates can be early eliminated before spreading in the environment.

From a toxicological point of view and during late drug development stages, as reliable as possible binding free energies are required, but accurate computational results are only possible at the expense of time. Nevertheless, *in silico* methods are generally faster and less expensive than laboratory experiments, in particular, if binding assays or synthesis protocols need to be developed. Moreover, in light of the continuous exponential progress in computer hardware and algorithms, MD simulations as described here will soon become a matter of a minute. Accordingly, we strongly recommend any binding free energy calculation on the basis of several heavily distinct complex geometries. In combination with the presented ERα model, a fully automated approach was developed that requires no more manual operations than a spatial vector specifying the binding site of a target and an arbitrary geometry of a small molecule under assessment.

Bibliography

[1] K. A. Dill and S. Bromberg. *Molecular Driving Forces: Statistical Thermodynamics in Chemistry and Biology*. Garland Science, New York, USA, 1st edition, 2002.

[2] H. Chanson. *Applied Hydrodynamics: An Introduction to Ideal and Real Fluid Flows*. CRC Press, Leiden, The Netherlands, 1st edition, 2009.

[3] P. Kapranov, J. Drenkow, J. Cheng, J. Long G. Helt, S. Dike, and T. R. Gingeras. Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. *Genome Res.*, 15:987–997, 2005.

[4] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell*. Garland Science, New York, USA, 7th edition, 2009.

[5] J. M. Berg, J. L. Berg, and L. Stryer. *Stryer Biochemie*. Spektrum Akademischer Verlag, Heidelberg, Germany, 2007.

[6] G. D. Dalkas, D. Vlachakis, D. Tsagkrasoulis, and A. Kastania S. Kossida. State-of-the-art technology in modern computer-aided drug design. *Brief. Bioinform.*, 14:745–752, 2013.

[7] S. M. Paul, D. S. Mytelka, C. T. Dunwiddie, C. C. Persinger, B. H. Munos, S. R. Lindborg, and A. L. Schacht. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat. Rev. Drug Discovery*, 9:203–214, 2010.

[8] J. P. Hughes, S. Rees, S. B. Kalindjian, and K. L. Philpott. Principles of early drug discovery. *Br. J. Pharmacol.*, 162:1239–1249, 2011.

[9] A. Suenaga, N. Okimoto, Y. Hirano, and K. Fukui. An efficient computational method for calculating ligand binding affinities. *PLoS One*, 7:e42846, 2012.

[10] Umweltbundesamt. Bewertung der Anwesenheit teil- oder nicht bewertbarer Stoffe im Trinkwasser aus gesundheitlicher Sicht. *Bundesgesundheitsbl. - Gesundheitsforsch. - Gesundheitsschutz*, 46:249–251, 2003.

[11] M. S. Marlow, J. Dogan, K. K. Frederick, K. G. Valentine, and A. J. Wand. The

role of conformational entropy in molecular recognition by calmodulin. *Nat. Chem. Biol.*, 6:352–358, 2010.

[12] B. O. Brandsdal, F. Österberg, M. Almlöf, I. Feierberg, V. B. Luzhkov, and J. Åqvist. Free energy calculations and ligand binding. *Adv. Protein. Chem.*, 66:123–158, 2003.

[13] P. Tosco and T. Balle. A 3D-QSAR-driven approach to binding mode and affinity prediction. *J. Chem. Inf. Model.*, 52:302–307, 2012.

[14] J. Simons. An experimental chemist's guide to *ab initio* quantum chemistry. *J. Phys. Chem.)*, 95:1017–1029, 1991.

[15] A. R. Leach. *Molecular Modelling: Principles and Applications*. Prentice Hall, Dorchester, Great Britain, 2nd edition, 2001.

[16] B. J. Alder and T. E. Wainwright. Phase transition for a hard sphere system. *J. Chem. Phys.*, 27:1208–1209, 1957.

[17] B. J. Alder and T. E. Wainwright. Studies in molecular dynamics. I. general method. *J. Chem. Phys.*, 31:459–466, 1959.

[18] N. Metropolis and S. Ulam. The Monte Carlo method. *J. Am. Stat. Assoc.*, 44:335–341, 1949.

[19] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087–1092, 1953.

[20] A. Rahman. Correlations in the motion of atoms in liquid argon. *Phys. Rev.*, 136:A405–A411, 1964.

[21] F. H. Stillinger and A. Rahman. Improved simulation of liquid water by molecular dynamics. *J. Chem. Phys.*, 60:1545–1557, 1974.

[22] J. A. McCammon, B. R. Gelin, and M. Karplus. Dynamics of folded proteins. *Nature*, 267:585–590, 1977.

[23] M. Karplus. Molecular dynamics of biological macromolecules: A brief history and perspective. *Biopolymers*, 68:350–358, 2003.

[24] D. Frenkel and B. Smit. *Understanding Molecular Simulation – From Algorithms to Applications*, volume 1 of *Computational Science Series*. Academic Press, 2002.

[25] D. K. Beveridge and F. M. DiCapua. Free energy via molecular simulation:

Applications to chemical and biomolecular systems. *Annu. Rev. Biophys. Biophys. Chem.*, 18:431–492, 1989.

[26] P. Kollman. Free energy calculations: Applications to chemical and biochemical phenomena. *Chem. Rev.*, 93:2395–2417, 1993.

[27] T. Rodinger and R. Pomès. Enhancing the accuracy, the efficiency and the scope of free energy simulations. *Curr. Opin. Struct. Biol.*, 15:164–170, 2005.

[28] J. H. Peters and B. L. de Groot. Ubiquitin dynamics in complexes reveal molecular recognition mechanisms beyond induced fit and conformational selection. *PLoS Comput Biol.*, 8, 2012.

[29] B. Ma, S. Kumar, C.-J. Tsai, and R. Nussinov. Folding funnels and binding mechanisms. *Protein Eng.*, 12:713–720, 1999.

[30] E. Breitmaier and G. Jung. *Organische Chemie*. Georg Thieme Verlag, Stuttgart, Germany, 5th edition, 2005.

[31] J. Maupetit, P. Derreumaux, and P. Tuffery. PEP-FOLD: an online resource for de novo peptide structure prediction. *Nucleic Acids Res.*, 37(Web Server issue):W498–503, 2009.

[32] T. Nugent and D. T. Jones. Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis. *Proc. Natl. Acad. Sci. U.S.A.*, 109:E1540–1547, 2012.

[33] Y. Yang, E. Faraggi, H. Zhao, and Y. Zhou. Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of the query and corresponding native properties of templates. *Bioinformatics*, 27:2076–2082, 2011.

[34] Y. Zhang and J. Skolnick. The protein structure prediction problem could be solved using the current PDB library. *Proc. Natl. Acad. Sci. U.S.A.*, 102:1029–1034, 2005.

[35] D. S. Marks, L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina, and C. Sander. Protein 3D structure computed from evolutionary sequence variation. *PLoS One*, 6:e28766, 2011.

[36] G. Helles. A comparative study of the reported performance of *ab initio* protein structure prediction algorithms. *J. R. Soc. Interface*, 5:387–396, 2008.

[37] A. Kryshtafovych and K. Fidelis. Protein structure prediction and model quality

assessment. *Drug Discovery Today*, 14:386–393, 2009.

[38] J. Jung and W. Lee. Structure-based functional discovery of proteins: Structural proteomics. *J. Biochem. Mol. Biol.*, 37:38–34, 2004.

[39] A. Brünger and M. Nilges. Computational challenges for macromolecular structure determination by X-ray crystallography and solution NMR-spectroscopy. *Q. Rev. Biophys.*, 26:49–125, 1993.

[40] M. Nilges. Structure calculation from NMR data. *Curr. Opin. Struct. Biol.*, 6:617–623, 1996.

[41] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Res.*, 28:235–242, 2000.

[42] I. P. Crawford, J. Slock, D. P. Stahly, E. W. Six, and C. Y. Han. An apparent Bacillus subtilis folic acid biosynthetic operon containing pab, an amphibolic trpg gene, a third gene required for synthesis of *para*-aminobenzoic acid, and the dihydropteroate synthase gene. *J. Bacteriol.*, 172:7211–7226, 1990.

[43] F. J. Schmitz, M. Perdikouli, A. Beeck, J. Verhoef, and A. C. Fluit. Resistance to trimethoprim-sulfamethoxazole and modifications in genes coding for dihydrofolate reductase and dihydropteroate synthase in european Streptococcus pneumoniae isolates. *J. Antimicrob. Chemother.*, 48:931–942, 2001.

[44] C. Chipot, M. S. Shell, and A. Pohorille. In C. Chipot and A. Pohorille, editors, *Free Energy Calculations – Theory and Applications in Chemistry and Biology*, volume 86 of *Chemical Physics*, chapter 1 Introduction. Springer-Verlag, Berlin/Heidelberg, Germany, 2007.

[45] S. Genheden and U. Ryde. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opin. Drug Discov.*, 10:449–461, 2015.

[46] R. Clausius. Ueber die bewegende Kraft der Wärme und die Gesetze, welche sich daraus ableiten lassen. *Pogg. Ann.*, 79:368–397, 1850.

[47] W. Göpel and H. D. Wiemdörfer. *Statistische Thermodynamik*. Spektrum Akademischer Verlag, Heidelberg/Berlin, Germany, 2000.

[48] H. Bisswanger. *Enzyme Kinetics: Principles and Methods*. Wiley-VCH, Weinheim, Germany, 2008.

[49] N. Foloppe and R. Hubbard. Towards predictive ligand design with free-energy

based computational methods? *Curr. Med. Chem.*, 13:3583–3608, 2006.

[50] T. Fliessbach. *Statistische Physik*, volume 4 of *Spektrum Lehrbuch*. Spektrum Akademischer Verlag, 2006.

[51] M. E. Tuckerman. *Statistical Mechanics: Theory and Molecular Simulation*. Oxford University Press, New York, 2010.

[52] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London, UK, 1996.

[53] P. H. Hünenberger. Thermostat algorithms for molecular dynamics simulations. In C. Holm and K. Kremer, editors, *Advanced Computer Simulation*, volume 173 of *Advances in Polymer Science*, pages 105–149. Springer-Verlag Berlin Heidelberg, 2005.

[54] H. S. Leff. The Boltzmann reservoir: A model constant-temperature environment. *Am. J. Phys.*, 68:521–524, 2000.

[55] R. Srinivasan. *Importance sampling - Applications in communications and detection*. Springer-Verlag, Berlin, Germany, 2002.

[56] C. Chipot and A. Pohorille. In C. Chipot and A. Pohorille, editors, *Free Energy Calculations – Theory and Applications in Chemistry and Biology*, volume 86 of *Chemical Physics*, chapter 2 Calculating Free Energy Differences Using Perturbation Theory. Springer-Verlag, Berlin/Heidelberg, Germany, 2007.

[57] K. A. Sharp. In H. Gohlke, editor, *Protein-Ligand Interactions*, volume 53 of *Methods and Principles in Medicinal Chemistry*, chapter 1 Statistical Thermodynamics of Binding and Molecular Recognition Models. Wiley-VCH, Weinheim, Germany, 2012.

[58] S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman, and J. M. Rosenberg. The weighted histogram analysis method for free-energy calculations on biomolecules. i. the method. *J. Comput. Chem.*, 13:1011–1021, 1992.

[59] R. W. Zwanzig. Hightemperature equation of state by a perturbation method. i. nonpolar gases. *J. Chem. Phys.*, 22:1420–1426, 1954.

[60] H.-J. Woo and B. Roux. Calculation of absolute protein–ligand binding free energy from computer simulations. *PNAS*, 102:6825–6830, 2005.

[61] S. g. Kang, P. Das, S. J. McGrane, A. J. Martin, T. Huynh, A. K. Royyuru, A. J. Taylor, P. G. Jones, and R. Zhou. Molecular recognition of metabotropic

glutamate receptor type 1 (mGluR1): Synergistic understanding with free energy perturbation and linear response modeling. *J. Phys. Chem. B*, 118:6393–6404, 2014.

[62] X. Ge and B. Roux. Absolute binding free energy calculations of sparsomycin analogs to the bacterial ribosome. *J. Phys. Chem. B*, 118:9525–9539, 2010.

[63] R. S. Rathore, M. Sumakanth, M. S. Reddy, P. Reddanna, A. A. Rao, M. D. Erion, and M. R. Reddy. Advances in binding free energies calculations: QM/MM-based free energy perturbation method for drug design. *Curr. Pharm. Des.*, 19:4674–4686, 2013.

[64] J. G. Kirkwood. Statistical mechanics of fluid mixtures. *J. Chem. Phys.*, 3:300–313, 1935.

[65] E. Darve. In C. Chipot and A. Pohorille, editors, *Free Energy Calculations – Theory and Applications in Chemistry and Biology*, volume 86 of *Chemical Physics*, chapter 4 Thermodynamic Integration Using Constrained and Unconstrained Dynamics. Springer-Verlag, Berlin/Heidelberg, Germany, 2007.

[66] A. de Ruiter, S. Boresch, and C. Oostenbrink. Comparison of thermodynamic integration and Bennett acceptance ratio for calculating relative protein–ligand binding free energies. *J. Comput. Chem.*, 34:1024–1034, 2013.

[67] C. H. Bennet. Efficient estimation of free energy differences from monte carlo data. *J. Comput. Phys.*, 22:245–268, 1976.

[68] A. M. Ferrenberg and R. H. Swendsen. Optimized Monte Carlo data analysis. *Phys. Rev. Lett.*, 63:1195–1198, 1989.

[69] G. M. Torrie and J. P. Valleau. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.*, 23:187–199, 1977.

[70] B. Roux. The calculation of the potential of mean force using computer simulations. *Comput. Phys. Commun.*, 91:275–282, 1995.

[71] P. A. Kollman, I. Massova, C. Reyes, B. Kuhn, S. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, O. Donini, P. Cieplak, J. Srinivasan, D. A. Case, and T. E. Cheatham. Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. *Acc. Chem. Res.*, 33:889–897, 2000.

[72] I. Andricioaei and M. Karplus. On the calculation of entropy from covariance

matrices of the atomic fluctuations. *J. Chem. Phys.*, 115:6289–6292, 2001.

[73] B. R. Brooks, D. Janezic, and M. Karplus. Harmonic analysis of large systems. *J. Comput. Chem.*, 16:1522–1553, 1995.

[74] D. Sitkoff, K. A. Sharp, and B. Honig. Accurate calculation of hydration free energies using macroscopic solvent models. *J. Phys. Chem.*, 98:1978–1988, 1994.

[75] C.-Y. Yang, H. Sun, J. Chen, Z. Nikolovska-Coleska, and S. Wang. Importance of ligand reorganization free energy in protein–ligand binding-affinity prediction. *J. Am. Chem. Soc.*, 131:13709–13721, 2009.

[76] W. C. Still, A. Tempczyk, R. C. Hawley, and T. Hendrickson. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.*, 112:6127–6129, 1990.

[77] T. Hou, J. Wang, Y. Li, and W. Wang. 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, 2011.

[78] J. Åqvist, C. Medina, and J. E. Samuelsson. A new method for predicting binding affinity in computer-aided drug design. *Protein Eng.*, 7:385–391, 1994.

[79] A. Ben-Naim and Y. Marcus. Solvation thermodynamics of nonionic solutes. *J. Chem. Phys.*, 81:2016–2027, 1984.

[80] T. Hansson and J. Åqvist. Estimation of binding free energies for HIV proteinase inhibitors by molecular dynamics simulations. *Protein Eng.*, 8:1137–1144, 1995.

[81] X. Jia, J. Zeng, J. Z. H. Zhang, and Y. Mei. Accessing the applicability of polarized protein-specific charge in linear interaction energy analysis. *J. Comput. Chem.*, 35:737–747, 2014.

[82] C. R. Vosmeer, R. Pool, M. F. Van Stee, L. Peric-Hassler, N. P. Vermeulen, and D. P. Geerke. Towards automated binding affinity prediction using an iterative linear interaction energy approach. *Int. J. Mol. Sci.*, 15:798–816, 2014.

[83] M. Nervall, P. Hanspers, J. Carlsson, L. Boukharta, and J. Åqvist. Predicting binding modes from free energy calculations. *J. Med. Chem.*, 51:2657–2667, 2008.

[84] M. M. H. van Lipzig, A. M. ter Laak, A. Jongejan, N. P. E. Vermeulen, M. Wamelink, D. Geerke, and J. H. N. Meermann. Prediction of ligand binding

affinity and orientation of xenoestrogens to the estrogen receptor by molecular dynamics simulations and the linear interaction energy method. *J. Med. Chem.*, 47:1018–1030, 2004.

[85] E. Stjernschantz, J. Marelius, C. Medina, M. Jacobsson, N. P. E. Vermeulen, and C. Oostenbrink. Are automated molecular dynamics simulations and binding free energy calculations realistic tools in lead optimization? An evaluation of the linear interaction energy (LIE) method. *J. Chem. Inf. Model.*, 46:1972–1983, 2006.

[86] R. Dias and W. F. de Azevedo Jr. Molecular docking algorithms. *Curr. Drug Targets*, 9:1040–1047, 2008.

[87] D. B. Kitchen, H. Decornez, J. R. Furr, and J. Bajorath. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov.*, 3:935–949, 2004.

[88] H. Hwang, T. Vreven, B. G. Pierce, J.-H. Hung, and Z. Weng. Performance of ZDOCK and ZRANK in CAPRI rounds 13-19. *Proteins*, 78:3104–3110, 2010.

[89] X.-Y. Meng, H.-X. Zhang, M. Mezel, and M. Cul. Molecular docking: A powerful approach for structure-based drug discovery. *Curr. Comput. Aided Drug Des.*, 7:146–157, 2011.

[90] G. L. Warren, C. W. Andrews, A.-M. Capelli, B. Clarke, J. LaLonde, M. H. Lambert, M. Lindvall, N. Nevins, S. F. Semus, S. Senger, G. Tedesco, I. D. Wall, J. M. Woolven C. E. Peishoff, and M. S. Head. A critical assessment of docking programs and scoring functions. *J. Med. Chem*, 49:5912–5931, 2006.

[91] J. Michel and J. W. Essex. Prediction of protein–ligand binding affinity by free energy simulations: assumptions, pitfalls and expectations. *J. Comput. Aided Mol. Des.*, 24:639–658, 2010.

[92] I. D. Kuntz, J. M. Blaney, S. J. Oatley, R. Langridge, and T. E. Ferrin. A geometric approach to macromolecule–ligand interactions. *J. Mol. Biol.*, 161:269–288, 1982.

[93] D. T. Moustakas, P. T. Lang, S. Pegg, E. Pettersen, I. D. Kuntz, N. Brooijmans, and R. C. Rizzo. Development and validation of a modular, extensible docking program: DOCK 5. *J. Comput. Aided Mol. Des.*, 20:601–619, 2006.

[94] M. Rarey, B. Kramer, T. Lengauer, and G. Klebe. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.*, 261:470–489, 1996.

[95] W. Welch, J. Ruppert, and A. N. Jain. Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites. *Chemistry & Biology*, 3:449–462, 1996.

[96] G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew, and A. J. Olson. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comp. Chem.*, 19:1639–1662, 1998.

[97] M. L. Verdonk, J. C. Cole, M. J. Hartshorn, C. W. Murray, and R. D. Taylor. Improved protein–ligand docking using GOLD. *Proteins*, 58:609–623, 2003.

[98] O. Trott and A. J. Olson. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. *J. Comp. Chem.*, 31:455–461, 2010.

[99] H.-J. Böhm. The development of a simple empirical scoring function to estimate the binding constant for a protein–ligand complex of known three-dimensional structure. *J. Comput. Aided Mol. Des.*, 8:243–256, 1994.

[100] R. Wang, Y. Lu, X. Fang, and S. Wang. An extensive test of 14 scoring functions using the PDBbind refined set of 800 protein-ligand complexes. *J. Chem. Inf. Comput. Sci.*, 44:2114–2125, 2004.

[101] I. Muegge and Y. C. Martin. A general and fast scoring function for protein–ligand interactions: A simplified potential approach. *J. Med. Chem.*, 42:791–804, 1999.

[102] R. Kim and J. Skolnick. Assessment of programs for ligand binding affinity prediction. *J. Comput. Chem.*, 29:1316–1331, 2008.

[103] S. J. Free and J. Wilson. A mathematical contribution to structure–activity studies. *J. Med. Chem.*, 7:395–399, 1964.

[104] C. Hansch and T. Fujita. ρ–σ–π Analysis. A method for the correlation of biological activity and chemical structure. *J. Am. Chem. Soc.*, 86:1616–1626, 1964.

[105] R. Perkins, H. Fang, W. Tong, and W. J. Welsh. Quantitative structure-activity relationship methods: perspectives on drug discovery and toxicology. *Environ. Toxicol. Chem.*, 22:1666–1679, 2003.

[106] W. J. Welsh, W. Tong, and P. G. Georgopoulos. In S. Ekins, editor, *Computational Toxicology: Risk Assessment for Pharmaceutical and Environmental Chemicals*, Technologies for the Pharmaceutical Industry, chapter 6 Toxicoinformatics:

An Introduction. John Wiley and Sons, Hoboken, New Jersey, 2007.

[107] A. K. Ghose and G. Crippen. Atomic physicochemical parameters for three-dimensional structure-directed quantitative structureactivity relationships I. Partition coefficients as a measure of hydrophobicity. *J. Med. Chem.*, 7:565–578, 1986.

[108] W. Tong, D. R. Lowis, R. Perkins, Y. Chen, W. J. Welsh, D. W. Goddette, T. W. Heritage, and D. M. Sheehan. Evaluation of quantitative structure-activity relationship methods for large-scale prediction of chemicals binding to the estrogen receptor. *J. Chem. Inf. Comput. Sci.*, 38:669–677, 1998.

[109] L. M. Shi, H. Fang, W. Tong, J. Wu, R. Perkins, R. M. Blair, W. S. Branham, S. L. Dial, C. L. Moland, and D. M. Sheehan. QSAR models using a large diverse set of estrogens. *J. Chem. Inf. Comput. Sci.*, 41:186–195, 2001.

[110] R. D. Cramer and B. Wendt. Template CoMFA: the 3D-QSAR grail? *J. Chem. Inf. Model.*, 54:660–671, 2014.

[111] L. B. Salum and A. D. Andricopulo. Fragment-based QSAR: perspectives in drug design. *Mol. Divers.*, 13:277–285, 2009.

[112] J. W. Ponder and D. A. Case. Force fields for protein simulations. *Adv. Protein Chem.*, 66:27–85, 2003.

[113] W. Lichtenthaler. Hundert Jahre Schlüssel-Schloß-Prinzip: Was führte Emil Fischer zu dieser Analogie? *Angew. Chem.*, 106:2456–2467, 1994.

[114] J. R. H. Tame. Scoring functions – the first 100 years. *J. Comput. Aided Mol. Des.*, 19:445–451, 2005.

[115] H. Gohlke, M. Hendlich, and G. Klebe. Knowledge-based scoring function to predict protein–ligand interactions. *J. Mol. Biol.*, 295:337–356, 2000.

[116] C. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard. A direct approach to conformational dynamics based on hybrid Monte Carlo. *J. Comput. Phys.*, 151:146–168, 1999.

[117] M. Weber. Conformation-based transition state theory. Technical report, Konrad-Zuse-Zentrum für Informationstechnik Berlin, 2007.

[118] M. Tabor. *Chaos and Integrability in Nonlinear Dynamics: An Introduction*. John Wiley and Sons, New York, USA, 2002.

[119] M. Šuvakov and V. Dmitrašinović. Three Classes of Newtonian Three-Body

Planar Periodic Orbits. *Phys. Rev. Lett.*, 110:114301, 2013.

[120] R. W. Hockney. The potential calculations and some applications. *Methods Comput. Phys.*, 9:136–211, 1970.

[121] H. S. Leff. Computer "experiments" on classical fluids. II. equilibrium correlation functions. *Phys. Rev*, 165:201–214, 1968.

[122] X. Zhang and K.-L. Han. High-order symplectic integration in quasi-classical trajectory simulation: Case study for $O(^1D) + H_2$. *Int. J. Quantum Chem.*, 106:1815–1819, 2005.

[123] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case. Development and testing of a general amber force field. *J. Comput. Chem.*, 25:1157–1174, 2004.

[124] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz Jr., D. M. Ferguson, D. C. Sepllmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.*, 117:5179–5197, 1995.

[125] K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov, and A. D. MacKerell Jr. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.*, 31:671–690, 2009.

[126] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.*, 118:11225–11236, 1996.

[127] T. A. Halgren. Merck Molecular Force Field. I-V. *J. Comp. Chem.*, 17:490–641, 1996.

[128] W. L. Jorgensen and J. Tirado-Rives. The OPLS force field for proteins. energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.*, 110:1657–1666, 1988.

[129] W. R. P. Scott, P. H. Hünenberger, I. G. Tironi, A. E. Mark, S. R. Billeter, J. Fennen, A. E. Torda, T. Huber, P. Krüger, and W. F. van Gusteren. The GROMOS biomolecular simulation program package. *J. Phys. Chem.*, 103:3596–3607, 1999.

[130] S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, and A. H. de Vries.

The MARTINI force field: Coarse grained model for biomolecular simulations. *J. Phys. Chem. B*, 111:7812–7824, 2007.

[131] W. Xie and J. Gao. Design of a next generation force field: The X-POL potential. *J. Chem. Theory Comput.*, 3:1890–1900, 2007.

[132] W. Xie, J. Pu, A. D. MacKerell Jr., and J. Gao. Development of a polarizable intermolecular potential function (PIPF) for liquid amides and alkanes. *J. Chem. Theory Comput.*, 3:1878–1889, 2007.

[133] G. Kaminski and W. L. Jorgensen. Performance of the AMBER94, MMFF94, and OPLS-AA force fields for modeling organic liquids. *J. Phys. Chem.*, 100:18010–18013, 1996.

[134] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins Struct. Funct. Bioinf.*, 65:712–725, 2006.

[135] P. Gans. *Vibrating Molecules*. Chapman and Hall, New York, 1971.

[136] S. Barlow, A. L. Rohl, S. Shi, C. M. Freeman, and D. O'Hare. Molecular mechanics study of oligomeric models for poly(ferrocenylsilanes) using the extensible systematic forcefield (ESFF). *J. Am. Chem. Soc.*, 118:7578–7592, 1996.

[137] A. K. Rappe, C. J. Casewit, K. S. Colwell, W. A. Goddard III, and W. M. Skiff. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc.*, 114:10024–10035, 1992.

[138] S. L. Mayo, B. D. Olafson, and W. A. Goddard. DREIDING: a generic force field for molecular simulations. *J. Phys. Chem.*, 94:8897–8909, 1990.

[139] J. P. Ryckaert and A. Bellemans. Molecular dynamics of liquid alkanes. *Far. Disc. Chem. Soc.*, 66:95–106, 1978.

[140] J. E. Lennard-Jones. On the determination of molecular fields. *Proc. R. Soc. Lond. A*, 106:463–477, 1924.

[141] A. Jakalian, B. L. Bush, D. B. Jack, and Ch. I. Bayly. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. method. *J. Comput. Chem.*, 21:132–146, 2000.

[142] H. Heinz and U. W. Suter. Atomic charges for classical simulations of polar systems. *J. Phys. Chem. B*, 108:18341–18352, 2004.

[143] B. Heinz and A. Nicholls. Classical electrostatics in biology and chemistry. *Sci-*

ence, 268:1144–1149, 1995.

[144] R. S. Mullikan. Electronic population analysis on LCAO-MO molecular wave functions. i. *J. Chem. Phys.*, 23:1833–1840, 1955.

[145] R. F. W. Bader, P. J. MacDougall, and C. D. H. Lau. Bonded and nonbonded charge concentrations and their relation to molecular geometry and reactivity. *J. Am. Chem. Soc.*, 106:1594–1605, 1984.

[146] C. A. Reynolds, J. W. Essex, and W. G. Richards. Atomic charges for variable molecular conformations. *J. Am. Chem. Soc.*, 114:9075–9079, 1992.

[147] J. Gasteiger and M. Marsili. A new model for calculating atomic charges in molecules. *Tetrahedron Lett.*, 34:3181–3184, 1978.

[148] A. K. Rappé and W. A. Goddard III. Charge equilibration for molecular dynamics simulations. *J. Phys. Chem.*, 95:3358–3363, 1991.

[149] J. Gasteiger and M. Marsili. Iterative partial equalization of orbital electronegativity - a rapid access to atomic charges. *Tetrahedron*, 36:3219–3228, 1980.

[150] L. F. Kuyper, R. N. Hunter, D. A. Ashton, K. M. Merz Jr., and P. A. Kollman. Free energy calculations on the relative solvation free energies of benzene, anisole, and 1,2,3-trimethoxybenzene: Theoretical and experimental analysis of aromatic methoxy solvation. *J. Phys. Chem.*, 95:6661–6666, 1991.

[151] C. I. Bayly, P. Cieplak, W. D. Cornell, and P. A. Kollman. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: The RESP model. *J. Phys. Chem.*, 97:10269–10280, 1993.

[152] W. D. Cornell, P. Cieplak, C. I. Bayly, and P. A. Kollman. Application of RESP charges to calculate conformational energies, hydrogen bond energies, and free energies of solvation. *J. Am. Chem. Soc.*, 115:9620–9631, 1993.

[153] A. Jakalian, D. B. Jack, and Ch. I. Bayly. Fast, efficient generation of high-quality atomic charges. AM1-BCC Model: II. parameterization and validation. *J. Comput. Chem.*, 23:1623–1641, 2002.

[154] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer-Verlag, New York, USA, 2nd edition, 2006.

[155] J. A. Nelder and R. Mead. A simplex method for function minimization. *Comput. J.*, 7:308–313, 1965.

[156] J. C. Lagarias, J. A. Reeds, M. H. Wright, and P. E. Wright. Convergence prop-

erties of the Nelder-Mead simplex method in low dimensions. *SIAM J. Optim.*, 9:112–147, 1998.

[157] M. R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *J. Res. Nat. Inst. Stand. Technol.*, 49:409–436, 1952.

[158] R. Fletcher and C. M. Reeves. Function minimization by conjugate gradients. *Computer Journal*, 7:149–154, 1964.

[159] R. P. Brent. *Algorithms for Minimization without Derivatives*. Prentice-Hall, NewJersey, USA, 1973.

[160] Martin Riedmiller and Heinrich Braun. A direct adaptive method for faster backpropagation learning: The RPROP algorithm. In *Proceedings of the IEEE International Conference on Neural Networks*, pages 586–591, San Francisco, CA, 1993.

[161] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.

[162] S. M. Le Grand and K. M. Merz Jr. The application of the genetic algorithm to the minimization of potential energy functions. *J. Global Optim.*, 3:49–66, 1993.

[163] D. J. Earl and M. W. Deem. Parallel tempering: Theory, applications, and new perspectives. *Phys. Chem. Chem. Phys.*, 7:3910–3916, 2005.

[164] V. Durmaz, R. Becker, and M. Weber. How to simulate affinities for host-guest systems lacking binding mode information: Application in the liquid chromatographic separation of hexabromocyclododecane stereoisomers. *J. Mol. Model.*, 18:2399–2408, 2012.

[165] J. Candy and W. Rozmus. A symplectic integration algorithm for separable Hamiltonian functions. *J. Comput. Phys.*, 92:230–256, 1991.

[166] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C*. Cambridge University Press, New York, NY, USA, 2 edition, 1992.

[167] E. Hairer, C. Lubich, and G. Wanner. Geometric numerical integration illustrated by the Störmer–Verlet method. *Acta Numerica*, 12:339–450, 2003.

[168] H. C. Andersen. Molecular dynamics simulations at constant pressure and/or temperature. *J. Chem. Phys.*, 72:2384–, 1980.

[169] G. Bussi, D. Donadio, and M. Parrinello. Canonical sampling through velocity

rescaling. *J. Chem. Phys.*, 126:014101, 2007.

[170] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gusteren, A. DiNicola, and J. R. Haak. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, 81:3684–3690, 1984.

[171] M. J. Abraham, D. van der Spoel, E. Lindahl, B. Hess, and the GROMACS development team. *GROMACS User Manual version 5.0.4*, 2014.

[172] W. G. Hoover. Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A*, 31:1695–1997, 1985.

[173] G. J. Martyna, M. L. Klein, and M. Tuckerman. Nosé–Hoover chains: The canonical ensemble via continuous dynamics. *J. Chem. Phys.*, 97:2635–2643, 1992.

[174] T. Schlick. *Molecular Modeling and Simulation: An Interdisciplinary Guide*. Springer-Verlag New York, Secaucus, New Jersey, USA, 2002.

[175] M. Parrinello and A. Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.*, 52:7182–7190, 1981.

[176] P. Eastman and V. J. Pande. CCMA: A robust, parallelizable constraint method for molecular simulations. *J. Chem. Theory Comput.*, 6:434–437, 2010.

[177] J. P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen. Numerical integration of the Cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *J. Comput. Phys.*, 23:327–341, 1977.

[178] M. Yoneda, H. J. C. Berendsen, and K. Hirasawa. A noniterative matrix method for constraint molecular-dynamics simulations. *Mol. Simulat.*, 13:395–405, 1994.

[179] H. C. Andersen. Rattle: A "velocity" version of the Shake algorithm for molecular dynamics calculations. *J. Comput. Phys.*, 52:24–34, 1983.

[180] B. J. Leimkuhler and R. D. Skeel. Symplectic numerical integrators in constrained Hamiltonian systems. *J. Comput. Phys.*, 112:117–125, 1994.

[181] S. Miyamoto and P. A. Kollman. SETTLE: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comput. Chem.*, 13:952–962, 1992.

[182] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije. LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.*, 18:1463–1472,

1997.

[183] A. Hastings, B. J. Pendelton, Y. Chen, and B. Robson. Hybrid Monte Carlo simulations theory and initial comparison with molecular dynamics. *Biopolymers*, 33:1307–1315, 1993.

[184] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

[185] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer Texts in Statistics. Springer-Verlag New York, Secaucus, New Jersey, USA, 2005.

[186] J. Wang, W. Wang, P. A. Kollman, and D. A. Case. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graphics Modell.*, 25:247–260, 2006.

[187] D. Stalling, M. Westerhoff, and H.-C. Hege. Amira: a highly interactive system for visual data analysis. In *The Visualization Handbook*, pages 749–767. Elsevier, 2005.

[188] A. Guerler, S. Moll, M. Weber, H. Meyer, and F. Cordes. Selection and flexible optimization of binding modes from conformation ensembles. *BioSystems*, 92:42–48, 2008.

[189] R. Kumari and R. Kumar. g_mmpbsa – A GROMACS tool for high-throughput MM-PBSA calculations. *J. Chem. Inf. Model.*, 54:1951–1962, 2014.

[190] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.*, 4:435–447, 2008.

[191] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison. Open babel: An open chemical toolbox. *J. Cheminf.*, 3:33, 2011.

[192] W. Humphrey, A. Dalke, and K. Schulten. VMD – Visual Molecular Dynamics. *J. Mol. Graphics*, 14:33–38, 1996.

[193] M. Weber, R. Becker, V. Durmaz, and R. Köppen. Classical hybrid Monte-Carlo simulation of the interconversion of hexabromocyclododecane stereoisomers. *Mol. Simulat.*, 34:727–736, 2008.

[194] V. Durmaz, S. Schmidt, P. Sabri, C. Piechotta, and M. Weber. A hands-off linear interaction energy approach to binding mode and affinity estimation of estrogens. *J. Chem. Inf. Model.*, 53:2681–2688, 2013.

[195] H. J. Barda, D. C. Sanders, and T. J. Benya. Bromine compounds. In S. Hawkins and G. Schultz, editors, *Ullmann's Encyclopedia of Industrial Chemistry*, pages 405–429. Wiley-VCH, Weinheim, Germany, 1985.

[196] C. A. de Witt. An overview of brominated flame retardants in the environment. *Chemosphere*, 46:583–624, 2002.

[197] A. Covaci, A. S. C. Gerecke, R. J. Law S. Voorspoels, M. Kohler, N. V. Heeb, H. Leslie, C. R. Allchin, and J. DeBoer. Hexabromocyclododecanes (HBCDs) in the environment and humans: a review. *Environ. Sci. Techno.*, 40:3680–3688, 2006.

[198] G. T. Tomy, K. Pleskach, T. Oswald, T. Halldorson, P. A. Helm, G. Macinnis, and C. H. Marvin. Enantioselective bioaccumulation of hexabromocyclododecane and congener-specific accumulation of brominated diphenyl ethers in an eastern Canadian Arctic marine food web. *Environ. Sci. Technol.*, 42:3634–3639, 2008.

[199] R. Köppen, R. Becker, S. Esslinger, and I. Nehls. Enantiomer-specific analysis of hexabromocyclododecane in fish from Etnefjorden (Norway). *Chemosphere*, 80:1241–1245, 2010.

[200] S. Esslinger, R. Becker, C. Jung, C. Schröter-Kermani, W. Bremser, and I. Nehls. Temporal trend (1988–2008) of hexabromocyclododecane enantiomers in herring gull eggs from the german coastal region. *Chemosphere*, 83:161–167, 2011.

[201] Retrieved on February 02, 2015 at `https://treaties.un.org/doc/Publication/CN/2013/CN.934.2013-Eng.pdf`.

[202] Y. Sugita and Y. Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Letters*, 314:141–151, 1999.

[203] S. P. Brooks and A. Gelman. General methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Stat.*, 7(4):434–455, 1998.

[204] A. Gelman and D. Rubin. Inference from iterative simulation using multiple sequences. *Statist. Sci.*, 7:457–511, 1992.

[205] N. V. Heeb, W. B. Schweizer, P. Mattrel, R. Haag, A. C. Gerecke, M. Kohler, P. Schmid M. Zennegg, and M. Wolfensberger. Solid-state conformations and absolute configurations of (+) and (−) α-, β-, and γ-hexabromocyclododecanes (HBCDs). *Chemosphere*, 68:940–950, 2007.

[206] A. Vattani. *k*-means requires exponentially many iterations even in the plane.

Discrete Comput. Geom., 45:596–616, 2011.

[207] A. Salam and M. S. Deleuze. High-level theoretical study of the conformational equilibrium of *n*-pentane. *J. Chem. Phys.*, 116:1296–1302, 2002.

[208] A. Bujotzek, O. Schütt, A. Nielsen, K. Fackeldey, and M. Weber. ZIBgridfree: Efficient conformational analysis by partition-of-unity coupling. *J. Math. Chem.*, 52:781–804, 2014.

[209] X. Li, J. Huang, P. Yi, R. A. Bambara, R. Hilf, and M. Muyan. Single-chain estrogen receptors (ERs) reveal that the ERalpha/beta heterodimer emulates functions of the ERalpha dimer in genomic estrogen signaling pathways. *Mol. Cell. Biol.*, 24:7681–7694, 2004.

[210] J. F. Couse, J. Lindzey, K. Grandien, J. A. Gustafsson, and K. S. Korach. Tissue distribution and quantitative analysis of estrogen receptor-alpha (ERalpha) and estrogen receptor-beta (ERbeta) messenger ribonucleic acid in the wild-type and ERalpha-knockout mouse. *Endocrinology*, 138:4613–4621, 1997.

[211] B. J. Deroo and K. S. Korach. Estrogen receptors and human disease. *J. Clin. Invest.*, 116:561–567, 2006.

[212] C. J. Fabian and B. F. Kimler. Selective estrogen-receptor modulators for primary prevention of breast cancer. *J. Clin. Oncol.*, 23:1644–1655, 2005.

[213] S. Oesterreich and N. E. Davidson. The search for ESR1 mutations in breast cancer. *Nature Genetics*, 45:1415–1416, 2013.

[214] J. A. McLachlan, K. S. Korach, R. R. Newbold, and G. H. Degen. Diethylstilbestrol and other estrogens in the environment. *Fundam. Appl. Toxicol.*, 4:686–691, 1984.

[215] A. Warnmark, E. Treuter, J.-A. Gustafsson, R. E. Hubbard, A. M. Brzozowski, and A. C. W. Pike. Interaction of transcriptional intermediary factor 2 nuclear receptor box peptides with the coactivator binding site of estrogen receptor alpha. *J. Biol. Chem.*, 277:21862–21868, 2002.

[216] W. L. Jorgensen, J. Chandrasekhar, and J. D. Madura. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79:926–935, 1983.

[217] E. J. Sorin and V. S. Pande. Exploring the helix-coil transition via all-atom equilibrium ensemble simulations. *Biophys. J.*, 88:2472–2493, 2005.

[218] H. J. C. Berendsen, D. van der Spoel, and R. van Drunen. GROMACS: A message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun*, 91:43–56, 1995.

[219] D. van der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen. GROMACS: Fast, flexible, and free. *J. Comput. Chem.*, 26:1701–1718, 2005.

[220] U. Essmann, L. Perera, and M. L. Berkowitz. A smooth particle mesh Ewald method. *J. Chem. Phys.*, 103:8577–8592, 1995.

[221] I. M. Billas, T. Iwema, J. M. Garnier, A. Mitschler, N. Rochel, and D. Moras. Structural adaptability in the ligand-binding pocket of the ecdysone hormone receptor. *Nature*, 426:91–96, 2003.

[222] H. Engelhardt, editor. *Practice of High Performance Liquid Chromatography*. Springer-Verlag, Berlin, Germany, 1986.

[223] G. Zhang, Q. Sun, Y. Hou, Z. Hong, J. Zhang, L. Zhao, H. Zhang, and Y. Chai. New mathematic model for predicting chiral separation using molecular docking: Mechanism of chiral recognition of triadimenol analogues. *J. Sep. Sci.*, 32:2401–2407, 2009.

[224] N. Issaraseriruk, A. Shitangkoon, and T. Aree. Molecular docking study for the prediction of enantiodifferentiation of chiral styrene oxides by octakis(2,3-di-*O*-acetyl-6-*O*-*O*-butyldimethylsilyl)-gamma-cyclodextrin. *J. Mol. Graphics Modell.*, 28:506–512, 2010.

[225] J. J. P. Stewart. Optimization of parameters for semiempirical methods. *J. Comput. Chem.*, 10:209–220, 1989.

[226] A. Pérez-Garrido, A. M. Helguera, M. N. D.S. Cordeiro, and A. G. Escudero. QSPR modelling with the topological substructural molecular design approach: β-cyclodextrin complexation. *J. Pharm. Sci.*, 98:4557–4576, 2009.

[227] E. Alvira, J. A. Mayoral, and J. I. García. Enantiodiscrimination of equol in β-cyclodextrin: an experimental and computational study. *J. Pharm. Sci.*, 60:103–113, 2008.

[228] K. B. Lipkowitz, G. Pearl, B. Coner, and M. A. Peterson. Explanation of where and how enantioselective binding takes place on permethylated β-cyclodextrin, a chiral stationary phase used in gas chromatography. *J. Am. Chem. Soc.*, 119:600–610, 1997.

[229] R. Köppen, R. Becker, F. Emmerling, C. Jung, and I. Nehls. Enantioselective preparative HPLC separation of the HBCD–stereoisomers from the technical product and their absolute structure elucidation using X-ray crystallography. *Chirality*, 19:214–222, 2007.

[230] F. H. Allen. The cambridge structural database: a quarter of a million crystal structures and rising. *Acta Crystallogr., Sect. B: Struct. Sci.*, 58:380–388, 2002.

[231] K. Harata, K. Uekama, T. Imai, F. Hirayama, and M. Otagiri. Crystal structures of heptakis(2,3,6-tri-*O*-methyl)-β-cyclodextrin complexes with (*r*)- and (*s*)-flurbiprofen. *J. Inclusion Phenomena*, 6:443–460, 1988.

[232] A. M. Nikitin and A. P. Lyubartsev. New six-site acetonitrile model for simulations of liquid acetonitrile and its aqueous mixtures. *J. Comput. Chem.*, 28:2020–2026, 2007.

[233] U. Essmann, L. Perera, and M. L. Berkowitz. A smooth particle mesh ewald method. *J. Chem. Phys.*, 103:8577–8592, 1995.

[234] J. Åquist, C. Medina, and J.-E. Samuelsson. A new method for predicting binding affinity in computer-aided drug design. *Protein Eng.*, 60:103–113, 2008.

[235] J.-H. Shi, Z. J. Ding, and Y. Hu. Experimental and theoretical studies on the enantioseparation and chiral recognition of mandelate and cyclohexylmandelate on permethylated β-cyclodextrin chiral stationary phase. *Chromatographia*, 74:319–325, 2011.

[236] N. T. McGachy, N. Grinberg, and N. Variankaval. Thermodynamic study of *n*-trifluoroacetyl-*o*-alkyl nipecotic acid ester enantiomers on diluted permethylated β-cyclodextrin stationary phase. *J. Chromatogr. A*, 1064:193–204, 2005.

[237] H. Kim, K. Jeong, S. Lee, and S. Jung. Molecular modeling of the chiral recognition of propranolol enantiomers by a β-cyclodextrin. *Bull. Korean Chem. Soc.*, 24:95–98, 2003.

[238] L. Perić-Hassler, E. Stjernschantz, C. Oostenbrink, and D. P. Geerke. CYP 2D6 binding affinity predictions using multiple ligand and protein conformations. *Int. J. Mol. Sci.*, 14:24514–24530, 2013.

[239] Q. Liu, C. K. Kwoh, and J. Li. Binding affinity prediction for protein–ligand complexes based on β contacts and B factor. *J. Chem. Inf. Model.*, 53:3076–3085, 2013.

[240] M. Griebel, S. Knapek, and G. Zumbusch. *Numerical Simulation in Molecular*

Dynamics. Springer, Heidelberg, Germany, 2007.

[241] T. Colborn. Developmental effects of endocrine-disrupting chemicals in wildlife and humans. *Environ. Health Perspect*, 101:378–384, 1993.

[242] Z. Wang, Y. Li, C. Ai, and Y. Wang. In silico prediction of estrogen receptor subtype binding affinity and selectivity using statistical methods and molecular docking with 2-arylnaphthalenes and 2-arylquinolines. *Int. J. Mol. Sci.*, 11:3434–3458, 2010.

[243] A. Khandelwal and S. Balaz. Improved estimation of ligand-macromolecule binding affinities by linear response approach using a combination of multimode MD simulation and QM/MM methods. *J. Comput. Aided Mol. Des.*, 21:131–137, 2007.

[244] P. Vasanthanathan, L. Olsen, F. S. Jørgensen, N. P. E. Vermeulen, and C. Oostenbrink. Computational prediction of binding affinity for CYP1A2–ligand complexes using empirical free energy calculations. *Drug Metab. Dispos.*, 38:1347–1354, 2010.

[245] A. K. Shiau, D. Barstad, P. M. Loria, L. Cheng, P. J. Kushner, D. A. Agard, and G. L. Greene. The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen. *Cell*, 95:927–937, 1998.

[246] G. G. J. M. Kuiper, B. Carlsson, K. Grandien, E. Enmark, J. Häggblad, S. Nilsson, and J.-Å. Gustafsson. Comparison of the ligand binding specificity and transcript tissue distribution of estrogen receptors α and β. *Endocrinology*, 138:863–870, 1997.

[247] R. M. Blair, H. Fang, W. S. Branham, B. S. Hass, S. L. Dial, C. L. Mol, W. Tong, L. Shi, R. Perkins, and D. M. Sheehan. The estrogen receptor relative binding affinities of 188 natural and xenochemicals: Structural diversity of ligands. *Tox. Sci.*, 54:138–153, 2000.

[248] K. Sharp. Entropy–enthalpy compensation: Fact or artifact? *Protein Sci.*, 10:661–667, 2001.

[249] T. A. Özal and N. F. A. van der Vegt. Confusing cause and effect: Energy-entropy compensation in the preferential solvation of a nonpolar solute in dimethyl sulfoxide/water mixtures. *J. Phys. Chem. B*, 110:12104–12112, 2006.

[250] M. Weber and K. Andrae. A simple method for the estimation of entropy differences. *MATCH Commun. Math. Comput. Chem.*, 63:319–332, 2010.

[251] M. Klimm, A. Bujotzek, and M. Weber. Direct reweighting strategies in conformation dynamics. *MATCH Commun. Math. Comput. Chem.*, 65:333–346, 2011.

[252] P. Deuflhard and A. Hohmann. *Numerical Analysis*. Walter de Gruyter, Berlin, New York, 1995.

[253] S. Agatonovic-Kustrin, M. Alexander, D. W. Morton, and J. V. Turner. Beware of q^2! *J. Mol. Graphics Modell.*, 20:269–276, 2002.

[254] T. Heberer. Tracking persistent pharmaceutical residues from municipal sewage to drinking water. *J. Hydrol.*, 266:175–189, 2002.

[255] P. E. Stackelberg, E. T. Furlong, M. T. Meyer, S. D. Zaugg, A. K. Henderson, and D. B. Reissman. Persistence of pharmaceutical compounds and other organic wastewater contaminants in a conventional drinking-water-treatment plant. *Sci. Total Environ.*, 329:99–113, 2004.

[256] M. Benotti, R. A. Trenholm, B. J. Vanderford, J. C. Holady, B. D. Stanford, and S. A. Snyder. Pharmaceuticals and endocrine disrupting compounds in U.S. drinking water. *Environ. Sci. Technol.*, 43:597–608, 2009.

[257] K. O. Kusk, T. Krüger, M. Long, C. Taxvig, A. E. Lykkesfeldt, H. Frederiksen, A. M. Andersson, H. R. Andersen, K. M. Hansen, C. Nellemann, and E. C. Bonefeld-Jørgensen. Endocrine potency of wastewater: contents of endocrine disrupting chemicals and effects measured by in vivo and in vitro assays. *Environ. Toxicol. Chem.*, 30:413–426, 2011.

[258] L. Schlüter-Vorberg, C. Prasse, T. A.Ternes, H. Mückter, and A. Coors. Toxification by transformation in conventional and advanced wastewater treatment: the antiviral drug acyclovir. *Environ. Sci. Technol. Lett.*, 2:342–346, 2015.

[259] B. I. Escher and K. Fenner. Recent advances in environmental risk assessment of transformation products. *Environ. Sci. Technol.*, 45:3835–3847, 2011.

[260] M. C. Dodd and C.-H. Huang. Transformation of the antibacterial agent sulfamethoxazole in reactions with chlorine: Kinetics, mechanisms, and pathways. *Environ. Sci. Technol.*, 38:5607–5615, 2004.

[261] L. Hu, P. M. Flanders, P. L. Miller, and T. J. Strathmann. Oxidation of sulfamethoxazole and related antimicrobial agents by TiO_2 photocatalysis. *Water Res.*, 41:2612–2626, 2007.

[262] T. B. Vree, A. J. A. M. van der Ven, P. P. Koopmans, E. W. J. van Ewijk-

Beneken Kolmer, and C. P. W. G. M. Verwey-van Wissen. Pharmacokinetics of sulfamethoxazole with its hydroxy metabolites and N_4-acetyl-, N_1-glucuronide conjugates in healthy human volunteers. *Clin. Drug Invest.*, 9:43–45, 1995.

[263] J. L. Mohatt, L. Hu, K. T. Finneran, and T. J. Strathmann. Microbially mediated abiotic transformation of the antimicrobial agent sulfamethoxazole under iron-reducing soil conditions. *Environ. Sci. Technol.*, 45:4793–4801, 2011.

[264] J. L. Maggs, M. Pirmohamed, M. R. Kitteringham, and B. K. Park. Characterization of the metabolites of carbamazepine in patient urine by liquid chromatography/mass spectrometry. *Drug Metab. Dispos.*, 25:275–280, 1997.

[265] D. C. McDowell, M. M. Huber, M. Wagner, U. von Gunten, and T. A. Ternes. Ozonation of carbamazepine in drinking water: Identification and kinetic study of major oxidation products. *Environ. Sci. Technol.*, 39:8014–8022, 2005.

[266] R. E. Pearce, G. R. Vakkalagadda, and J. S. Leeder. Pathways of carbamazepine bioactivation in vitro I. characterization of human cytochromes P450 responsible for the formation of 2- and 3-hydroxylated metabolites. *Drug Metab. Dispos.*, 30:1170–1179, 2002.

[267] W. Lu and J. P. Uetrecht. Peroxidase-mediated bioactivation of hydroxylated metabolites of carbamazepine and phenytoin. *Drug Metab. Dispos.*, 36:1624–1636, 2008.

[268] H. Breton, M. Cociglio, F. Bressolle, H. Peyriere, and J. P. Blayac D. Hillaire-Buys. Liquid chromatography–electrospray mass spectrometry determination of carbamazepine, oxcarbazepine and eight of their metabolites in human plasma. *J. Chromatogr. B*, 828:80–90, 2005.

[269] I. C. Hampele, A. D'Arcy, G. E. Dale, D. Kostrewa, J. Nielsen, C. Oefner, M. G. Page, H. J. Schönfeld, D. Stüber, and R. L. Then. Structure and function of the dihydropteroate synthase from Staphylococcus aureus. *J. Mol. Biol.*, 25:21–30, 1997.

[270] W. A. Petri. Sulfonamides, trimethoprim-sulfamethoxazole, quinolones, and agents for urinary tract infections. In L. L. Brunton, B. A. Chabner, and B. C. Knollmann, editors, *Goodman and Gilman's - The Pharmacological Basis of Therapeutics*, pages 1443–1476, New York, 2011. McGraw-Hill.

[271] P. Veyssier and A. Bryskier. Dihydrofolate reductase inhibitors, nitroheterocycles (furans), 8-hydroxyquinolines. In A. Bryskier, editor, *Antimicrobial Agents*, pages

941–963, Washington DC, 2005. ASM Press.

[272] J. P. Sanderson, D. J. Naisbitt, J. Farrell, C. A. Ashby, M. J. Tucker, M. J. Rieder, M. Pirmohamed, S. E. Clarke, and B. K. Park. Sulfamethoxazole and its metabolite nitroso sulfamethoxazole stimulate dendritic cell costimulatory signaling. *J. Immunol.*, 178:5533–5342, 2007.

[273] S. A. Kaplan, R. E. Weinfeld, C. W. Abruzzo, K. McFaden, M. L. Jack, and L. Weissman. Pharmacokinetic profile of trimethoprim-sulfamethoxazole in man. *J. Infect. Dis.*, 128:547–555, 1973.

[274] M. Ueda, N. Murakami, H. Atsumura, and K. Huruki. Studies on metabolism of drugs. VII. On the metabolite of sulfamonomethoxine in human. *Yakugaku Zasshi*, 87:455–457, 1967.

[275] Y. Tsurui and T. Koizumi. Studies on metabolism of drugs. XII. Quantitative separation of metabolites in human and rabbit urine after oral administration of sulfamonomethoxine and sulfamethomidine. *Chem Pharm Bull (Tokyo)*, 20:2042–2046, 1972.

[276] C. Levy, D. Minnis, and J. P. Derrick. Dihydropteroate synthase from Streptococcus pneumoniae: structure, ligand recognition and mechanism of sulfonamide resistance. *Biochem. J.*, 412:379–388, 2008.

[277] M. K. Yun, Y. Wu, Z. Li, Y. Zhao, M. B. Waddell, A. M. Ferreira, R. E. Lee, D. Bashford, and S. W. White. Catalysis and sulfa drug resistance in dihydropteroate synthase. *Science*, 335:1110–1114, 2012.

[278] A. Achari, D. O. Somers, J. N. Champness, P. K. Bryant, J. Rosemond, and D. K. Stammers. Crystal structure of the anti-bacterial sulfonamide drug target dihydropteroate synthase. *Nat. Struct. Biol.*, 4:490–497, 1997.

Appendix A

List of Abbreviations

β-pmCD	Permethylated β-cyclodextrin
AA	Amino acid
ACN	Acetonitrile
ADMET	Absorption, distribution, metabolism, elimination, and toxicity
AMBER	Assisted model building with energy refinement
BAM	Federal Institute for Materials Research and Testing (*Bundesanstalt für Materialforschung*)
BAR	Bennett acceptance ratio
CIP	Cahn-Ingold-Prelog
CoMFA	Comparative Molecular Field Analysis
CSD	Cambridge Structural Database
DHPS	Dihydropteroate synthase
E_2	17-β-Estradiol
EcR	Ecdysone receptor
EDC	Endocrine disrupting chemical
EP	Endothiapepsin
ERα	Estrogen receptor α
ESP	Eletrostatic potential
FEP	Free energy perturbation
FF	Force field
FLOPS	Floating point operations per second
GAFF	General AMBER force field

GBMR	Gradient-based minimization routine
HBCD	Hexabromocyclododecane
HMC	Hybrid Monte Carlo
HPLC	High-performance liquid chromatography
LBD	Ligand binding domain
LIE	Linear interaction energy
LJ	Lennard-Jones
LOOCV	Leave-one-out cross-validation
LRA	Linear response approximation
MC	Monte-Carlo
MCMC	Markov chain Monte Carlo
MD	Molecular dynamics
MM	Molecular mechanics
MM/GBSA	Molecular mechanics generalized Born surface area
MM/PBSA	Molecular mechanics Poisson-Boltzmann surface area
MMFF	Merck molecular force field
NMR	Nuclear magnetic resonance
PABA	*para*-Aminobenzoic acid
PDB	Protein Data Base
PES	Potential energy surface
PLS	Partial least squares
PMF	Potential of mean force
POP	Persistent organic pollutant
PR	Position restraints
PSRF	Potential scale reduction factor
QM	Quantum mechanics
QSAR	Quantitative structure–activity relationship
QSPR	Quantitative structure–property relationship
RBA	Relative binding affinity
RESP	Restrained eletrostatic potential
RMS	Root mean square
RMSD	Root mean square deviation

SA	Simulated annealing
SASA	Surface accessible surface area
SMZ	Sulfamethoxazole
TI	Thermodynamic integration
US	Umbrella sampling
VDW	Van-der-Waals
WHAM	Weighted histogram analysis method

Appendix B

List of Symbols

β	Inverse product of Boltzmann constant and temperature
γ	Coupling parameter
δ	Delta function
μ	Chemical potential
ξ	Normalization factor of phase space partition function (Section 2.2)
ξ	Reaction coordinate (Section 2.3)
π	Boltzmann factor
σ^2	Statistical variance
ϕ	ϕ angle of amino acids
ψ	ψ angle of amino acids
Ω	Degeneracy of microstates
∇	Nabla operator
d	Dimension
h	Planck constant
k_B	Boltzmann constant
p	Conjugate momenta
p	Pressure
q	Generalized coordinates
r	Cartesian coordinates
r_{ij}	Distance between atoms i and j
v	Velocity
v_{ij}	Bond vector between two atoms i and j
x	Point (coordinates and momentum) in phase space (Chapter 2)
x	Weight coefficients
z_i	Charge of atom i
A	Helmholtz energy
E	Energy

G	Gibbs energy
G°	Standard Gibbs energy
H	Enthalpy
H	Hesse matrix (Section 3.4)
$\mathcal{H}$	Hamiltonian
K	Kinetic energy
K_a	Chemical association constant
K_d	Chemical dissociation constant
L	Ligand
$\mathcal{L}$	Lagrangian
M	Diagonal matrix of atom masses
N	Number of particles
N_f	Number of degrees of freedom
O	Observable
P	Probability
P	Protein
Q	Heat
Q	Phase space partition function
R	Gas constant
S	Entropy
T	Temperature
U	Internal energy (Section 2.1)
U	Potential energy
V	Volume
W	Work
Z	Configurational space partition function

Appendix C

List of Figures

Appendix D

List of Tables

Appendix E

Index

N

P

Q

R

S

T

U

V

W

Acknowledgements

Hereby I want to express my thanks to everyone having contributed to the completion of this thesis. Above all, I am indebted to my supervisor Marcus Weber for his continuous support and encouragement during all ups and downs. I would also like to thank Paul Wrede for additional supervision of this thesis and his always helpful hints. Apart from many other ZIB members and alumni, I want to thank my former ZIB colleagues Bernd Kallies, Frank Cordes, and, in particular, my longtime room mate (and I dare to say friend) Alexander Bujotzek for their always open ear and friendly support in any technical and scientific as well as personal matter. Special thanks go to a couple of cooperation partners: Roland Becker, Christian Piechotta, and Sebastian Schmidt from the Federal Institute for Materials Research and Testing (Bundesanstalt für Materialforschung, BAM) for providing me with biochemical insights, and Harald Mückter from the Ludwig Maximilian University of Munich for fruitful scientific discussions as well as stimulating social events beyond research. Regarding financial support I would like to thank BAM for several interesting projects, the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) for funding the SFB-765 joint project, and the German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung, BMBF) for funding both the TransRisk and BB3R joint projects. Finally, I want to thank my better half Marie for her patience and loving support during these sometimes stressful months.

Eidesstattliche Erklärung

Hiermit erkläre ich alle Hilfsmittel und Hilfen angeben und auf dieser Grundlage die Arbeit selbständig verfasst zu haben. Zudem versichere ich, dass die vorliegende Arbeit in keinem früheren Promotionsverfahren eingereicht worden ist.

Ort, Datum Unterschrift

Zusammenfassung

Insbesondere für die toxikologische Risikobewertung und den pharmakologischen Wirkstoffentwurf gewinnt die rechnergestützte Vorhersage exakter Bindungsaffinitäten für Protein-Ligand-Systeme nach wie vor zunehmend an Bedeutung. Der Einsatz sogenannter *in-silico*-Methoden schont nicht nur zeitliche wie finanzielle Resourcen, sondern bietet zudem oft die einzige Möglichkeit zur Bewertung neuartiger Substanzen, die beispielsweise aus der Metabolisierung anthropogener Chemikalien hervorgegangen sind. Dem kontinuierlichen technologischen wie algorithmischen Fortschritt der vergangenen Jahrzehnte zum Trotz stellt die Bestimmung freier Bindungsenergien mithilfe molekülдynamischer Berechnungen aufgrund der hohen mathematischen Komplexität biologischer Wirt-Gast-Komplexe eine gewaltige Herausforderung dar. Die Ziele dieser Dissertation lassen sich in zwei Kategorien einteilen. Auf der einen Seite wurden algorithmische Strategien zur weiträumigen Abtastung und Zerlegung des Konformationsraums sowie zur automatisierten Auswahl repräsentativer Konformere bzw. Bindungsmodi entwickelt. Angesichts der für molekülдynamische Simulationen typischen langen Verweilzeit in metastabilen Konformationen lag das Augenmerk auf einer möglichst umfassenden Repräsentation physikalisch zugänglicher Bereiche des Konformationsraums. Ausgehend von diesen Eingangsgeometrien wurden auf der anderen Seite für eine Reihe von (bio)molekularen Systemen empirische Vorhersagemodelle entwickelt, die im Kern auf dem Verfahren der linearen Interaktionsenergie (LIE) beruhen. Dabei handelte es sich hauptsächlich um eine Abschätzung flüssigchromatographischer Retentionszeiten sowie der Elutionsreihenfolge von Stereoisomeren des Flammschutzmittels Hexabromocyclododecan (HBCD) und ein neues Modell der Bindungsaffinität zum humanen Estrogenrezeptor-α (ER-α) basierend auf einem LIE-QSAR-Hybriden. In einer letzten Anwendung diente eine nichtparametrisierte Abwandlung eines rein physikalischen ER-α-Modells zur toxikologischen Priorisierung von Transformationsprodukten des Antibiotikums Sulfamethoxazol im Rahmen einer Risikobewertung bezüglich ihrer Bindungswahrscheinlichkeit an das bakterielle Enzym Dihydropteroat-Synthase. In ihrer Gesamtheit beschreibt diese Dissertation eine neuartige sowie vollständig automatisierte Prozedur zur Bestimmung von Bindungsgeometrien und -affinitäten, die sich mit der räumlichen Angabe der Bindestelle und einer beliebigen Ligandengeometrie begnügt. Im Vergleich mit etablierten Dockingroutinen bzw. thermodynamischen Methoden wurden deutlich verlässlichere Resultate erzielt, was nicht zuletzt den systematischen Raumzerlegungstrategien geschuldet ist. Sowohl bei den Retentionszeiten von HBCD als auch den Bindungsaffinitäten an ERα betrug die quadrierte Korrelation mit Laborwerten mehr als 0,8. Etwa 85 % (100 %) der vorhergesagten Bindingsmodi von Protein-Ligand-Komplexen wichen um weniger als 1,53 Å (2,05 Å) von Kristallstrukturen ab.