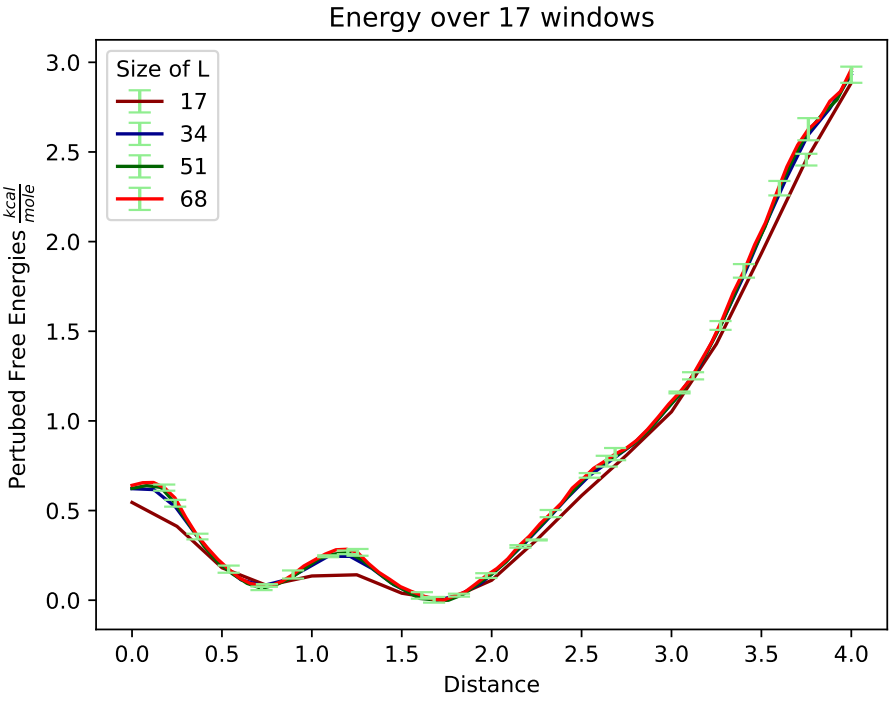
A computational workflow for binding free energies in Python



Master Thesis

Chemical Engineering - Simon Nygaard-Thomsen

Aalborg University

9th of September 2022

AALBORG UNIVERSITY
STUDENT REPORT

10^{th} **semester - ENG**
Chemical Engineering
Frederik Bajers Vej 7H
9000 Aalborg
http://www.engineering.aau.dk

Title:

A computational workflow for binding free energies in Python

Project:

Master Thesis - 60 ECTS

Period of project:

September 2021 - September 2022

Author:
Simon Nygaard-Thomsen

Supervisor:
Casper Steinmann

Number of pages: 37
Appendix: None
Ended: 09-09-2022

Abstract:

The literature show that β-cyclodextrin (β-CD) have good probabilities of forming inclusion complexes in water with ligands having a low solubility themselves, both *in situ* and *in silico*. Through Python, the framework OpenMM was used together with OpenFF (developers of the SMIRNOFF force fields (FF)) to simulate such inclusion complexes in water with Molecular Dynamics (MD). SMIRNOFF uses so-called direct chemical perception, in opposition to older FF like AMBER, to parameterise molecules directly from the chemical graph by the use of SMIRKS. SMIRKS are able to recognise patterns in molecules, making the process more efficient. The goal of the MD simulations was to introduce a biasing potential between the β-CD and a ligand and by measuring the distance between the two molecules in each snapshot, finding the free binding energy of the inclusion complex. Through Umbrella Sampling (US) and analyses by FastMBAR, a software tool for applying the Bennet Acceptance Ratio, it was shown that OpenMM was indeed a viable option in this regard. Through the analyses, it was revealed that a sample size of 2000 snapshots per window in the US and a bin size of 20 per window, both the free energy and standard deviations converged. Comparisons with the literature showed that the method was feasible although some of the free energies of the inclusion complexes were a bit high.

Preface

The purpose of this thesis is to investigate the possibilities within Python, OpenMM, and OpenFF in regard to setup simple simulations and evaluate them with respect to laboratory experiments and similar simulations.

I would like to thank my supervisor Casper Steinmann, with whom I have had weekly meetings for almost a year, who have shown an earnest interest in this thesis and the work required, and is an altogether great teacher

Reading guide

All used literature will be listed in the bibliography in the back of the report. Throughout the report literature will be referred to according to the Harvard method, [Surname, Year]. This leads to the bibliography where literature as articles and books are indicated with; Author, title, edition and publisher. Literature such as websites will be indicated with author, title and date.

Figures and tables are numbered, respective to the chapter, which means first figure in chapter 3 will be numbered Fig. 3.1. Captions will be below figures and above tables.

Constants	
$\mathbf{a}$	Acceleration
A	Helmholz Free Energy
$\mathbf{A}$	Reduced energy matrix
C	Ratio between two states in relation to BAR
$\Delta\Delta G$	Relative change in Free Energy
ΔG	Gibbs Free Energy
E	Energy
ϵ	Dielectric constant
$\mathbf{F}$	Force
f	Fermi Function
H	Host
K	Force constant
K_{eq}	Equilibrium constant
k_b	Boltzmann constant
L	Ligand
λ	Window or coordinate state
m	Mass
n_λ	Number of coordinate states at λ
ϕ	Dihedral angle
Q	Partial charge
$\mathbf{r}$	Position
R	Distance between two particles or molecules
R_0	Reaction coordinate
ΔS	Entropy
σ	Standard deviation
t	Time
T	Temperature
θ	Angle
U	Energy potential or internal energy
$\mathbf{v}$	Velocity
V	Energy Potential
ζ	Friction coefficient

Abbreviations	
AI	Artificial Intelligence
BAR	Bennett Acceptance Ratio
β-CD	β-cyclodextrin
CD	Cyclodextrin
CM	Classical Mechanics
CoM	Center of Mass
ITC	Isometrical Calorimetry
LJ	Leonard Jones potential
MD	Molecular Dynamics
NMR	Nuclear Magnetic Resonance
NPT	Constant mole, pressure and temperature
NVT	Constant mole, volume and temperature
PES	Potential Energy Surface
PMF	Potential Mean Force
QM	Quantum Mechanics
US	Umbrella Sampling
VdW	Van der Waals
WHAM	Weighted Histogram Analysis Method

Elements	
Br	Bromine
C	Carbon
Cl	Chlor
H	Hydrogen
N	Nitrogen
Na	Natrium
O	Oxygen
P	Phosphor
S	Sulphur

Contents

Introduction 1

Cyclodextrins (CD) are formed of six, seven, or eight α-1,4-glycosidic covalently linked glucopyranose units and have the common names $\alpha-, \beta-$, and γ-cyclodextrin (see Figure 1.1a) (Larsen et al., 2005). There exists bigger CDs, however most studies focuses on the listed three (Larsen, 2002). CDs have a conical shape with the secondary alcohols (OH2 and OH2) at the wide end and the primary alcohol (OH6) at the narrow and the size of the cavity increases with sugar molecules (Sabadini et al., 2006). When forming a ring, O5, H3, and H5 will point inwards resulting in an exterior which is more polar than the cavity (see Figure 1.1b). These properties of cyclodextrins make it possible to dissolve an otherwise water insoluble compound in water by letting it form an inclusion complex with CD, which in turn makes it useful within the fields of pharmaceutical, food, and the chemical industry (Larsen et al., 2005; Uekama et al., 1998; Davis and Brewster, 2004; Fukahori et al., 2006).

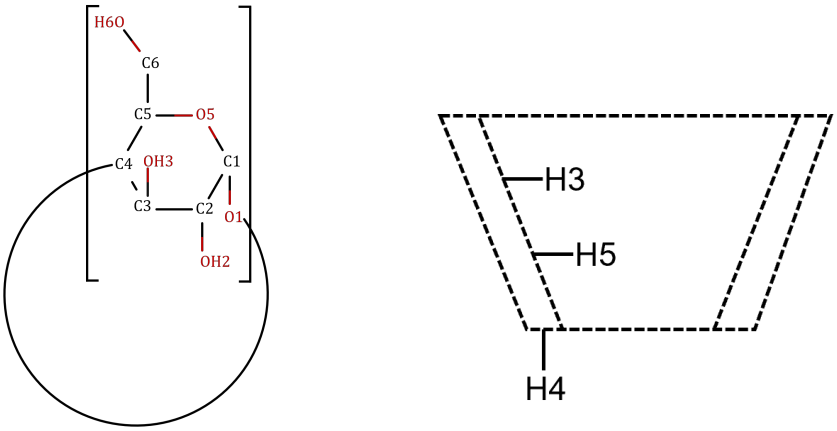


Figure 1.1. Different depictions of cyloclodextrins. *(a)* Cyclodextrin with n α-1,4-glycosidic linkages where n is the number of α-pyranose. The subcased numbers relate to all atoms connected to the specific carbon e.g. C6 is connected to H61, H62, and O6 which has H60 as the last atom. *(b)* Sketch of cyclodextrin showing the position of H3, H5 (inside the CD), and H4 (pointing downwards from the primary edge). (Larsen et al., 2005; Sabadini et al., 2006)

When solvated in water the alcohol groups will interact with the water in bulk, resulting in the cavity being slightly less polar than the outside. The space in the cavity will be taken up by 3-4 water molecules but these readily leave if a less polar molecule is present and of course close enough (Szejtli, 1998; Larsen et al., 2005). This phenomenon can be explained by the state functions (the change in Gibbs free energy ΔG, enthalpy ΔH, and entropy ΔS) and the inherent need of any compound or system to be in as a low an energy state

as possible. While any system strives to have as high an entropy as possible the need of having a low energy can overcome this energy barrier. If the water molecules are removed from the cavity in favour of a less polar molecule, the entropy and enthalpy will decrease, meaning that the reaction is driven by enthalpy. The decrease in entropy is due to the system becoming less chaotic as more intermolecular bonds are able to form. The increase in number of intermolecular bonds also decreases the enthalpy as the system becomes less strained. The "freed" water molecules will become solvated in the rest of the water, while the less polar molecule will form a complex with the CD (Szejtli, 1998).

Larsen et al. (2005) investigated $\alpha-, \beta-$, and γ-CD forming inclusion complexes with prednisolone and 6α-methyl-prednisolone in water. Isothermal Titration Calorimetry (ITC) resulted in inclusion stability constants of $3479(139)\,\mathrm{M}^{-1}$ and $1022(30)\,\mathrm{M}^{-1}$ for prednisoline and 6α-methyl-prednisolone, respectively. Nuclear Magnetic Resonanse (NMR) showed that the more polar end of the compounds would stick out of the CDs and into the solution while the non-polar end would be inside the cavity. When complexing with γ-CD the association constant K would be higher than that of β-CD, however the solubility of the inclusion complex would decrease significantly at higher concentrations, making for a poor host. The association and solubility constants of the inclusion complex formed with α-CD were much lower than that with β-CD

Even though the analyses mentioned above show reliable results, it can be difficult to predict beforehand if a compound make for a good inclusion complex or not. Through computer simulations this problem can, to some extent, be overcome. The quality and reliability of computer chemistry have improved immensely over the last decades and have gone from testing a few specifically chosen compounds to screen a huge number of viable candidates through appliance of Artificial Inteligence (AI) (Steinmann and Jensen, 2021). The next section will briefly introduce the basics of the simulation approach.

1.1 Mechanics of a chemical system

All chemical systems consists of atoms which can be split in two main particle groups; nuclei and electrons. In Quantum Mechanics (QM) both of the particle types are allowed to contribute whereas in Classical Mechanics (CM) only the nuclei are considered. At velocity slower than the speed of light Newtons second law (Equation (1.1)) is used to describe the classical mechanics i.e. the movement of the atoms and molecules: (Jensen, 2017)

$$\mathbf{F} = m \cdot \mathbf{a}, \tag{1.1}$$

where $\mathbf{F}$ is the force, the mass, m, is constant and $\mathbf{a}$ is the acceleration. A starting condition of CM is the position of the center of mass of the particles. The force is equivalent to the derivative of the potential V with respect to the position $\mathbf{r}$, which in turn is the same as the mass multiplied with the second derivative of the acceleration of the positions with respect to time t. (Equation (1.2)). (Jensen, 2017)

$$-\frac{\partial V}{\partial \mathbf{r}} = m \cdot \mathbf{a}, \quad \mathbf{a} = \frac{\partial^2 \mathbf{r}}{\partial t} \tag{1.2}$$

1.1.1 Time dependent systems

To propagate the system in time, Equation (1.2) will has to be solved numerically. At a given time t_i the system has the position of $\mathbf{r}_i$. As time propagates one time step Δt later the new positions $\mathbf{r}_{i+1}$ can be calculated from the velocity, acceleration, etc. as a Taylor expansion with respect to time. The simplest form of dynamics is given as the Verlet algorithm (Equation (1.3)) (Jensen, 2017):

$$\mathbf{r}_{i+1} = (2\mathbf{r}_i - \mathbf{r}_{i-1}) + \mathbf{a}(\Delta t)^2 + \dots \tag{1.3}$$

Beyond the third order in time step the Verlet algorithm is no longer true and the velocity is not calculated explicitly. To overcome this the Velocity Verlet algorithm is used instead which combines the Verlet algorithm with a leap frog algorithm. The latter (Equation (1.4)) introduces a half-step to the time step, however out of phase which is corrected when combined into the velocity Verlet algorithm (Equations (1.5) and (1.6)) Jensen (2017)

$$\mathbf{v}_{i+\frac{1}{2}} = \mathbf{v}_{i-\frac{1}{2}} + \mathbf{a}_i \Delta t \tag{1.4}$$

$$\mathbf{r}_{i+1} = \mathbf{r}_i + \mathbf{v}_i \Delta t + \frac{1}{2}\mathbf{a}_i \Delta t^2 \tag{1.5}$$

$$\mathbf{v}_i = \mathbf{v}_i + \frac{1}{2}(\mathbf{a}_i + \mathbf{a}_{i+1})\Delta t \tag{1.6}$$

The Langevin equation is another method of propagating a system in time (Equation (1.7)). One main difference from the velocity Verlet algorithm is that it introduces the friction coefficient ζ, which is used to model to the surrounding molecules so that only the average interactions are included. ζ is proportional to the velocity of the atoms and $\mathbf{F}_{random}$ which averages to zero. While ζ removes energy to the system, the random component $\mathbf{F}_{random}$ add energy and is associated with a temperature. $\mathbf{F}_{intra}$ accounts for the intra molecular forces and in some cases the non-bonded forces. (Jensen, 2017)

$$m\frac{d^2\mathbf{r}}{dt^2} = -\zeta\frac{d\mathbf{r}}{dt} + \mathbf{F}_{intra} - \mathbf{F}_{random} \tag{1.7}$$

Zhang et al. (2019) recently proposed a so called "middle scheme" to the Langevin equation (by introducing the leap-frog algorithm). They showed that this method yield more realistic energies of the solvent than the Langevin method. As the focus of this project is the guest-host interactions in water it makes sense to apply the Langevin Middle method.

1.1.2 Importance of the time step

When running any simulation the time step is an important parameter to control. The largest value of Δt has to be one magnitude smaller than the fastest process in the system. This corresponds to the lightest particle, which for most systems is the one of hydrogen which vibrates with a frequency of $3000\,\mathrm{cm}^{-1} \approx 10^{14}\mathrm{s}^{-1}$, which again corresponds to a time step of one femtosecond (10^{-15}s). (Jensen, 2017)

1.2 Force Fields

A big part of any computer simulation is to make it as cheap as possible while keeping the quality of the calculations high enough. For this purpose Force Fields are used. Force fields

have the advantage of adhering to the laws of classical mechanics while the parameters of the bonded forces (bonds, angles and torsion) and non-bonded (van der Waals and electrostatic) forces between atoms and molecules are predefined. By implementing the force field parameters in a simulation, the calculations performed by CM becomes cheaper as opposed to finding them directly with CM and/or QM. The force field parameters are often derived from QM and/or CM or from experimental data. When applying force fields the energy of each step is found by the terms of Equation (1.8) and propagated in time by Equation (1.2). (Jensen, 2017)

$$E_{FF} = \underbrace{E_{bond} + E_{angle} + E_{torsion}}_{bonded} + \underbrace{E_{vdW} + E_{elec}}_{non-bonded} \tag{1.8}$$

Figure 1.2 show the internal forces, i.e. the first three terms of Equation (1.8). The stretching of the bond between two atoms, the angle bend of three atoms and the torsion term which describe the dihedral angle of four atoms. Beyond the fourth atom (No bond in Figure 1.2) the atoms can no longer "feel" each other much like the chemical shift of NMR and vibrations in Infrared pectroscopy (IR). This is both due to the distance between the atoms and that the atoms in between shadow the forces.

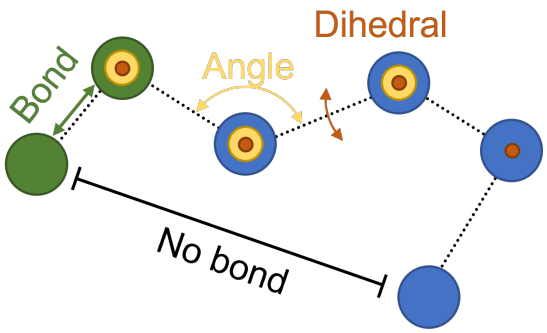


Figure 1.2. The different forces interacting in a molecule. The colour signifies which force is applied and the arrows show the direction. Bond stretch (two atoms): green, angle bend (three atoms): yellow, and dihedral angle (four atoms): red. The *No bond* signifies, that there are no direct bond between the atoms, meaning that they only interact through intramolecular forces. Jensen (2017).

1.2.1 Intramolecular forces

The bond and angle terms (Equations (1.9) and (1.10)) can in most cases be described with a harmonic potential which increases as the quadratic when displaced from the minimum (or equilibrium) Equation (1.9) describes how the bond between two atoms stretches where $k^{\mathbf{r},\mathbf{r}_0}$ is the force constant between the atoms $\mathbf{r}$ and $\mathbf{r}_0$.

$$E_{bond} = \frac{1}{2}k^{\mathbf{r},\mathbf{r}_0}(\mathbf{r} - \mathbf{r}_0)^2 \tag{1.9}$$

$$E_{angle} = \frac{1}{2}k^{\theta,\theta_0}(\theta - \theta_0)^2 \tag{1.10}$$

Equation (1.10) describes how the angle of three atoms bends around the middle one (see the yellow in Figure 1.2) where k^{θ,θ_0} is the force constant of the angle bend, and θ and θ_0 represent two sets of angles.

$$E_{torsion} = \sum_i V_i[1 \pm \cos n_i\phi_i] \tag{1.11}$$

The torsional term (Equation (1.11)) accounts for the dihedral angle i.e. the orange part of Figure 1.2. V_i is the energy barrier, which has to be overcome to rotate around the middle bond, ϕ is the angle (again, revolving around the middle bond) and n describes how many periods/possible the dihedral angle has. It is very much different from E_{bond} and E_{angle} as it can revolve around the second bond (orange arrow) more or less freely due to a low energy barrier (Jensen, 2009, 2017). Even though the energy barrier is low, the molecule will often only have a few positions in which there is an equilibrium (Jensen, 2017).

1.2.2 Intermolecular forces

The last two terms of Equation (1.8) describe how molecules interact with each other. The energy of van der Waals (shown below a as Leonard Jones potential (Equation (1.12))) describes the repulsion two between non bonded particles, specifically non polar particles or particles with non-polar areas. At a low distance, the first term of Equation (1.12) increases the energy of E_{vdW} due to the steric repulsion between two particles. This effect also prevents said particles from collapsing on top of each other. The attraction term (second term of Equation (1.12)) applies when there is a charge polarisation between two particles, creating an induce dipole-dipole interaction.

$$E_{LJ}(R) = \frac{C_1}{R^{12}} - \frac{C_2}{R^6}, \tag{1.12}$$

where $E_{LJ}(R)$ is the van der Waals energy as a function of the distance R, the repulsion depends on the first term $\frac{1}{R^{12}}$, the attraction term is described by $\frac{1}{R^6}$ while C_1 and C_2 are constants.

The electrostatic energy depends on the partial charges of the atoms and how the electrons are distributed in the molecules. In force fields there are no electrons, instead each atom is assigned the partial charges Q^A and Q^B of atom A and B, respectively (Equation (1.13)).

$$E_{el} = \frac{Q^A Q^B}{\varepsilon R^{AB}}, , \tag{1.13}$$

where ε is the dielectric constant, and R^{AB} is the distance between atoms A and B.

1.3 Selecting a force field

The purpose of force fields is to make simulations cost less in regards to computer time while at the same time retain some level of accuracy. Some force fields are parameterised specifically for inorganic systems while other are parameterised towards organic systems. Some are designed to handle many interactions at a cost in the level of accuracy while others are designed to handle relatively few interactions at a high level of accuracy.

AMBER is a force field designed towards organic chemistry with a main focus on proteins i.e. many interactions (Jensen, 2017; Case et al., 2014). The GLYCAM06 force field is developed on the basis of AMBER but specialised towards mono- and oligosaccharides and small molecules (Kirschner et al., 2007). When handling cyclodextrins, GLYCAM06 is a good match, as it is written and parameterised specifically for this type of molecule. While

both AMBER and GLYCAM06 produce reliable results, they have a steep learning curve and requires a lot of initial programming before simple simulations can be set up. A part of the challenge arises when trying to implement some ligand unknown to GLYCAM06, as this has to be parameterised before running a simulation.

GLYCAM06 contains six different elements (C, H, N, O, S and P) each defined in the force field with different parameters according to element and interaction. How an element behaves depends on its neighbours, which is why each element (except phosphate) has two or more (up to ten) different atom types of which there is 30 in total. Carbon alone accounts for ten different atom types of which six are different kinds of sp3 hybridised aliphatic carbons. (Kirschner et al., 2007) By limiting the number of atom types to 30, parameters too similar can be avoided and the list of possible interactions are kept relatively short. This, however is a problem with General AMBER Force Field (GAFF) as it contains and accounts for a large number of atoms and possible combinations.

Sage, a recent addition to the many different force fields have been developed by the Open Force Field Consortium (2022). Sage is the second generation force field of SMIRNOFF99Frosst which was made to be a minimalistic AMBER-compatible general force field for small molecules, specifically for drug-like molecules. SMIRNOFF99Frosst is comparable to the physical properties of GAFF but is much less complex in its use, as it uses the SMARTS direct chemical perception that SMIRNOFF makes possible. (Mobley et al., 2018)

1.3.1 Chemical perception

There are two methods belonging to the term chemical perception: **indirect** chemical perception and **direct** chemical perception. Both methods are based on the input of a chemical graph depicting a molecule and the end result are the same. The path, however is very much different.

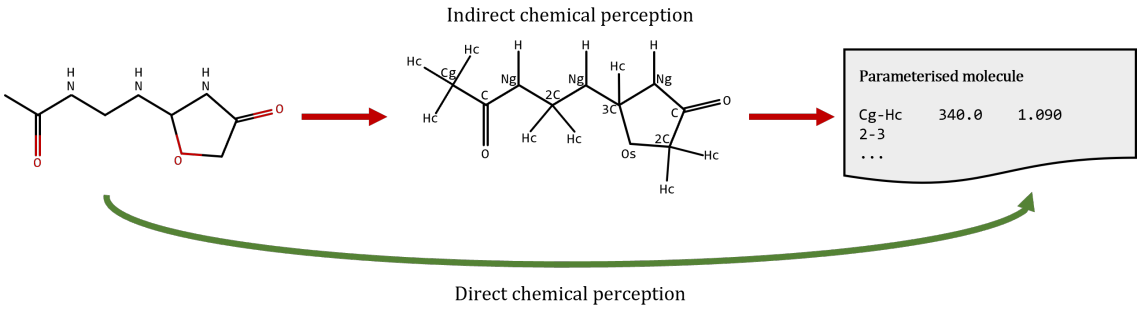


Figure 1.3. Two different paths of parametrisising with indirect chemical perception (top) and direct chemical percepttion (bottom). The atom types are based on GLYCAM06-j and so is the parameters.

The indirect method, which older programs like AMBER, GAFF, GLYCAM06, and others use, is basically just reading of lists (Figure 1.3). An input is given, specifying two to four atoms and how they are connected. The connection applies to the bond, angle, torsion and Van der Waals radius. All this is specified by the use of short keywords consisting of 1 to 3 letters which then hold all the information. This means that relatively big force fields like

GAFF uses several thousand lines of code, often made through human chemical intuition and experience. (Mobley et al., 2018) This sometimes leads to errors in the parametrisising of the final calculations, e.g. if a planar molecule is forced to bend due to bonds, which are **not** double bonds but perceived as such (Wang et al., 2006; Mobley et al., 2018).

Direct chemical perception means that the molecule is parameterised on the fly rather than by the preset indirect chemical perception. By this method, the amount of data needed to be read and stored is drastically lowered. With direct chemical perception the pattern of the chemical graph is recognised (Figure 1.3). By using the pattern to assign the valence, bond order and chemical environment the need for atom types are removed.

1.3.2 Applying SMIRKS

While the SMIRKS language is normally represented by the symbol H, C, N, and O, in OpenFF the symbols are replaced by hashtag followed by the atom number `#1, #6, #7,` and `#8` respectively (Mobley et al., 2018). In SMIRKS a carbon connected to four atoms would look like this `#6X4` while a carbon with only bonds to three other atoms (i.e. one bond being a double bond) would be represented by `#6X3`.

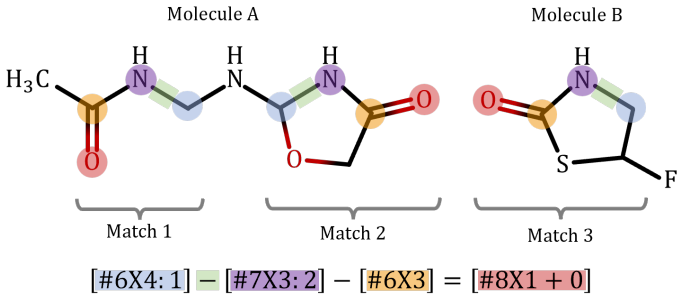


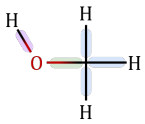
Figure 1.4. SMIRKS representation of two molecules with three matching patterns. (Mobley et al., 2018)

SMIRKS is a smart choice due to it's ability to recognise patterns. Figure 1.4 shows the chemical graph of two molecules with the SMIRKS representation. There is a total of three matches (same order of appearance of atoms) in the two molecules, making it relatively easy to find the correct parameters and reuse those, `#6X4:1` (blue) has a single (-) labelled (green) bond to the trivalent nitrogen `#7X3:2` (purple). Each of these are indexed (`:1` and `:2`) for special treatment, in this case to assign bond stretch parameters. The amide is single bound to a trivalent alpha carbon `#6X3` (orange) which has one bond to a carbon and one double bond (=) to the beta-oxygen `#8X1+0` (red) which is neutral, denoted by the `+0`. In molecules some functional groups have a greater influence on the molecule than others. In Figure 1.5a below, an excerpt of the XML-representation of methanol is shown. For each force type related to the molecule in question SMIRNOFF loops over the chemical graph and finds all the SMIRKS assosiated with the molecule.

Figure 1.5a shows the SMIRKS of a harmonic bond force (Equation (1.9)). In the case of methanol carbon bonds to three different hydrogens with the same properties (blue). The `:1` and `:2` denotes the indices of the atoms. Carbon also has a single bond to a hydroxyl group (green). While this in reality covers two bonds (C – O – H), it is instead treated as a

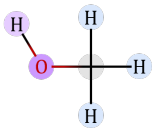
C−OH due to the change induced into oxygen by hydrogen. The last bond (light purple) describes the bond of the hydroxyl group in its own right. In the case of methanol, the bond force is relatively straight forward.

```
<HarmonicBondForce length_unit="angstroms" k_unit="kilocalories_per_mole/angstrom**2">
 <Bond smirks="[#6X4:1]-[#1:2]" length="1.090" k="680.0"/>
 <Bond smirks="[#6X4:1]-[8&amp;X2&amp;H1:2]" length="1.410" k="640.0"/>
 <Bond smirks="[#8X2:1]-[#1:2]" length="0.960" k="1106.0"/>
</HarmonicBondForce>
```



(a)

```
<NonbondedForce coulomb14scale="0.833333" lj14scale="0.5" sigma_unit="angstroms"
epsilon_unit="kilocalories_per_mole">
 <Atom smirks="[#1:1]" rmin_half="1.4870" epsilon="0.0157"/>
 <Atom smirks="[$(#1)-[#6]-[#7,#8,#9,#16,#17,#35]):1]" rmin_half="1.3870" epsilon="0.0157"/>
 <Atom smirks="[#1$(*-[#8]):1]" rmin_half="0.0000" epsilon="0.0000"/>
 <Atom smirks="[#6:1]" rmin_half="1.9080" epsilon="0.1094"/>
 <Atom smirks="[#8:1]" rmin_half="1.6837" epsilon="0.1700"/>
 <Atom smirks="[#8X2+0$(*-[#1]):1]" rmin_half="1.7210" epsilon="0.2104"/>
</NonbondedForce>
```



(b)

Figure 1.5. Excerpt of the XML representing the SMIRNOFF corresponding to methanol. Each of the different forces are boldfaced for readability. The SMIRKS on the left are colour coded to the corresponding atom or bond in the figure to the right. The colour codes correspond to the primary atoms (`HarmonicBondForce` and `NonbondedForce`). Grey corresponds to carbon, red to oxygen, light green to oxygen-carbon, yellow to hydrogen, light purple to hydrogen-oxygen, blue to hydrogen-carbon, and the hydroxyl oxygen is highlighted by magenta. Note that the XML contains the units of the forces within. ***(a)*** shows the SMIRKS harmonic bond force, ***(b)*** shows the SMIRKS nonbonded force. (Mobley et al., 2018).

For the non bonded force type this is not the case. As can be seen on Figure 1.5b, there are more SMIRKS than colours corresponding to the chemical graph to the right. The first line (yellow), showing a single hydrogen is replaced by second line (blue) which is more specific towards a hydrogen bonding with H−C−[N, O, F, S, Cl, Br]. The same is the case for the third line (light purple), which also replaces the first lines for the specialised hydrogen in a hydroxyl group. The last line to replace another, is the purple line which replaces the single oxygen with the specific hydroxyl group which has two bonds, the hydrogen and something else denoted by the `*-[#1]`.

1.4 Reaction energies

The energy of reaction between a host and ligand is written as Equation (1.14) where ΔG_H and ΔG_L are the free energies of the reaction $\Delta_r G$, respectively.

$$\Delta_r G = G_{complex} - G_{solution} \tag{1.14}$$

Reaction (R1) show the corresponding reaction scheme:

$$\mathrm{L_{solution} + H_{solution} \rightleftharpoons LH_{complex}} \tag{R1}$$

with an equilibrium constant K_{eq} of

$$K_{eq} = \frac{[LH_{complex}]}{[L_{solution}][H_{solution}]} \tag{1.15}$$

At equilibrium $\Delta G = 0$ and therefore

$$0 = \Delta_r G + RT \ln K_{eq} \tag{1.16}$$

which can be rewritten to

$$-RT \ln K_{eq} = \Delta_r G \tag{1.17}$$

where $\Delta_r G$ is the free energy of the reaction, R is the gas constant $8.314\,\mathrm{J\,K^{-1}\,mol^{-1}}$, and T is the temperature 298.15 K.

Figure 1.6 illustrates the Potential Energy Surface (PES) of Equation (1.14). The purpose of the PES is to identify where the most likely equilibrium of a system is, corresponding to the minima on a PES (it can be found as a function of time, distance, angle, and dihedral angle, among others). For each minimum, there is some degree of equilibrium. In the case of host-guest interactions, the global minimum of the PES is theoretically where the host and ligand forms an inclusion complex.

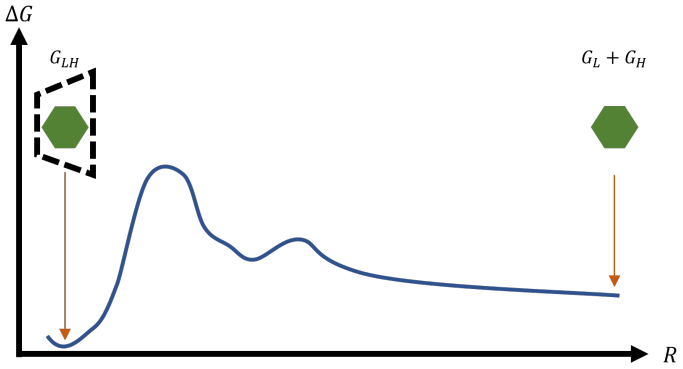


Figure 1.6. Example of a PES between a guest and host solvated in water. The x-axis shows the reaction coordinate R and the y-axis shows the the relative free energy.

Figure 1.6 show two points on the PES; G_{LH} and $G_L + G_H$ where each point represents a minimum corresponding to the free energy between ligand and host at a certain distance. The distance between guest and host increases with R. By increasing the distance in increments the PES of the host-guest relation can be found leading to the minima in free energy. At G_{LH} the two molecules form an inclusion complex resulting in a low ΔG. At distances *not* at G_{LH} the systems energy is higher and increases as the energy barriers after G_{LH} has to be overcome. At $G_L + G_H$ the distance between the two molecules is great enough for them not to "notice" each other.

1.4.1 Relative free energy

As earlier stated, the free energy can be used as a measure of how well a host and ligand bind to each other. Because the free energy is a state function, the path to the results are not important in the sense that the free energy is relative. This means that when a molecule moves from one state to another, be it solid to liquid or from solution to complex is not that important. The reaction schemes shown in Figure 1.7 visualise the change in energy from solution to complex.

In Figure 1.7a there are two paths. **1)** The "real" (horizontal) which shows the free energy of the hosts and ligands A + H and B + H complexing in two different reactions (each of the reactions are comparable to the one shown in Figure 1.6). **2)** The alchemical (vertical) which shows the free energy of the ligand A reacts into ligand B, the host H reacts into the host H (spoiler, the change in free energy is zero), and the complexes AH reacts into BH (Mey et al., 2020).

The reaction circle shown in Figure 1.7b shows change in free energy between different complexes. The reaction circle shows the change in free energy when moving from $AH \rightarrow BH \rightarrow CH \rightarrow$ and back to AH from CH. The "circle" can be enlarged as necessary with as many components as one desires as long as the change between molecules are not too big (Mey et al., 2020).

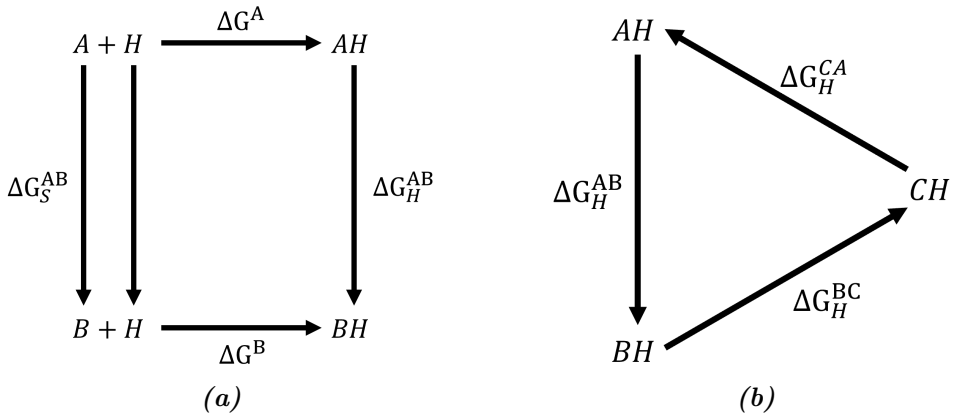


Figure 1.7. Different free energy pathways of host-ligand interactions. H is the host, A and B are the ligands ,and AH and BH are the respective complexes. ΔG^A and ΔG^B are the change in free energy from solution to complex while ΔG_S^{AB} and ΔG_H^{AB} are the change in free energy from one compound to another. ***(a)*** shows the connection between the "real" (horizontal) and alchemical (vertical) path. ***(b)*** shows how the vertical path of ***(a)*** can be used to change the free energy between different inclusion complexes.

A given reaction in which the ligand A forms an inclusion complex with the host H (see the horizontal part of Figure 1.7a) can be written as Reaction (R2):

$$\mathrm{A} + \mathrm{H} \rightleftharpoons^{\Delta G^A} \mathrm{AH} \tag{R2}$$

Similar reactions can be written for the ligands B and C:

$$\mathrm{B} + \mathrm{H} \rightleftharpoons^{\Delta G^B} \mathrm{BH} \tag{R3}$$

$$\mathrm{C} + \mathrm{H} \rightleftharpoons^{\Delta G^C} \mathrm{CH} \tag{R4}$$

If the ligands are very similar, e.g benzene and toluene the energy difference of the inclusion complexes would also be relatively small. By assuming this, one could calculate the difference in Gibbs free energy between the two ligands as Equation (1.18)

$$\Delta\Delta G^{AB} = \Delta G^B - \Delta G^A \tag{1.18}$$

When finding the free energy between inclusion complexes, $\Delta\Delta G$, there are two possible methods; simulation of the whole reaction from start to end (Figure 1.7a horizontal path) or

alchemical free energy calulations (Figure 1.7a vertical path). The first has the advantage of revealing how the energy changes as a function of distance, however it returns only one $\Delta\Delta G$. $\Delta\Delta G$ can be found as Equation (1.19) and expanded to Equation (1.20):

$$\Delta\Delta G = \Delta\Delta G^{AB} + \Delta\Delta G^{BC} + \Delta\Delta G^{CA} \tag{1.19}$$

$$\Delta\Delta G = \Delta G^{A} - \Delta G^{B} + \Delta G^{B} - \Delta G^{C} + \Delta G^{C} - \Delta G^{A} = 0 \tag{1.20}$$

The problem, however is that when cleaning the equation, it will always return zero. To counter this, the standard error of each will be included, so the free energy will be

$$\Delta\Delta G = 0 \pm \frac{\sigma}{\sqrt{n}} \tag{1.21}$$

The argument for doing it this way is that the standard error of $\Delta\Delta G$ is found by

$$\sigma_{\Delta\Delta G} = \sqrt{2\sigma_A^2 + 2\sigma_B^2 + 2\sigma_C^2} \tag{1.22}$$

The rightfulness of the method and results will then depend on Equation (1.22) instead of the $\Delta\Delta G$ which will always be zero.

1.4.2 Simulating the reaction path

As mentioned earlier, Figure 1.6 shows a possible PES when separating a guest-ligand complex, and together with Figure 1.7a it is possible to find the free binding energy of the complex. This approach also has the possible advantage of showing how the ligand behaves in relation to the host.

You et al. (2019) simulated the whole reaction path from solution to inclusion complex through umbrella sampling (US) of a number of frames and finding the Perturbed Mean Free energy (PMF) by Weighted Histogram Analysis Method (WHAM) and Multiple Bennet Acceptance Ratio (MBAR). In umbrella sampling a set of windows (starting configurations) is made and a biasing potential is applied. The biased energy is then found through simulation of each window (further details will be covered in the theory). You et al.'s method yields an estimate of the PES of the reaction from being in solution to inclusion complex. An important aspect is that the method is comparable to how the ligand and host interact in solution *in vitro*. When doing umbrella samplings, each window is set with a specific parameter, in this case the distance R_0, of Equation (1.23):

$$E_{bias} = \frac{1}{2} \cdot K \cdot (R - R_0)^2 \,, \tag{1.23}$$

where E_{bias} is the energy of the biasing potential, K is a force constant and R_0 is the measured distance. Equation (1.23) has the same functional form as Equation (1.9), only here the particles are whole molecules. When doing this kind of simulation, the ligand and host start at the positions of the inclusion complex and are then pulled apart frame by frame. For each frame a specific R is set while the deviation R_0 is measured and sampled. You et al. (2019) investigated the guest-host interactions from 0 to 26 Å with a step size of 0.1 Å between each window, resulting in 260 windows. Each production simulation took 2.5 ns per window, in total 650 ns.

1.4.3 Alchemical free energy calulations

Alchemical free energy calculations is another method for finding the free energy change from molecules in solution to molecules in inclusion complex. The method is only possible in *in vitro* as the laws of physics are broken by introducing chimeric molecules. Chimeric molecules are engineered molecules and does not necessarily exist in reality (Borsari et al., 2020). In alchemical free energy calculation it is used to describe intermediate states, such as the ones showed in Figure 1.8a (Mey et al., 2020). The participation of the blue part of the molecule decreases with an increase in λ. λ is the steps or windows of the simulation, similar to the ones described in the section above. This does not mean that the molecule is not there, rather it's energy contribution is decreased in favour of the orange part.

Figure 1.8b illustrate the probability density function of the potential energy as a function of $\lambda = [0:1]$, respectively. The potential free energies of each λ and it neighbours (λ_{n-1}, λ_n, and λ_{n+1}) has to overlap to make sure the calculations are valid (Mey et al., 2020).

The probability of the potential energy $P(E_{pot})$ as a function of λ is show in Figure 1.8b. At $\lambda = 0$ the probability of phenol contributing to the potential energy is the highest while it decreases as λ. The opposite is true for benzene.

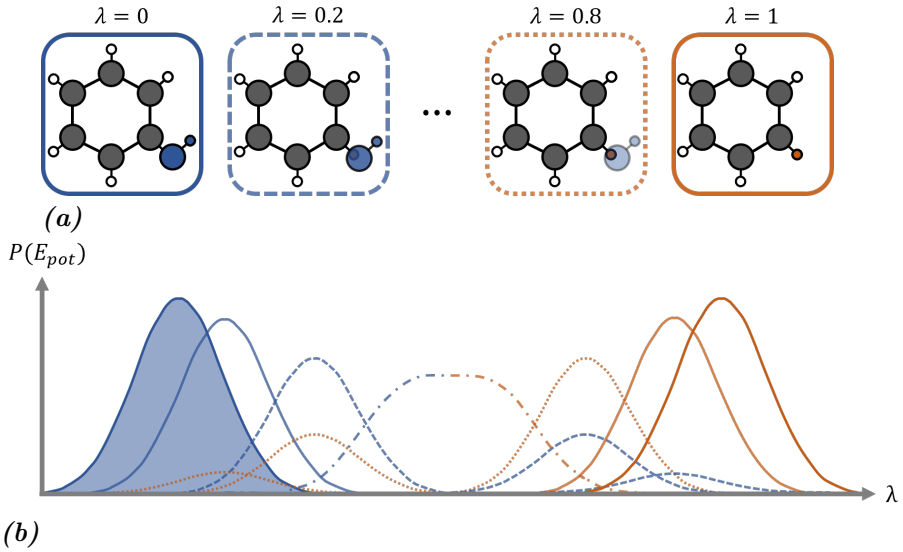


Figure 1.8. Alchemical λ between benzene and phenol. At $\lambda = 0$ phenol is fully interacting, while this is true for benzene at $\lambda = 1$. Inspired by Mey et al. (2020)

As their contribution to the SAMPL6-challenge, Caldararu et al. (2018) made a set of alchemical simulations with 13 windows and with a production time per window of 2 ns totalling to 26 ns.

Choosing the alchemical or reaction path

When comparing the two methods in relation to time, it is obvious that the alchemical procedure is the more efficient (Table 1.1) (the alchemical procedure is repeated three times with different starting conditions to remove simulation artefacts). If one wish to

know more about the reaction path the alchemical cannot be used, as it's only focus is the the two extremes; in solution and in inclusion complex. The concept of the reaction path is relatively to understand as it is very similar to how the experiment would be made in a laboratory in opposition to the alchemical method.

Table 1.1. Comparison of two chemical simulation methods (Caldararu et al. (2018)[a], You et al. (2019)[b]).

	Alchemical path	Reaction path
Number of windows	13[a]	260[b]
Production time per window	2 ns[a]	2.5 ns[b]
Number of repetitions	3[a]	0[b]
Total simulation time	78 ns	650 ns
Reaction path	No	Yes
Simulation time	Short	Medium to long
Ease of implementation	Difficult	Medium

A downside to the reaction path method is that a lot of windows are needed (and therefore simulation time) to obtain a high degree of accuracy. In the specific method applied by You et al. (2019) they increase the distance between host and guest by 0.1 Å per window. On the other hand it has the possibility of producing knowledge of how the host and ligand might react when close together but not quite bound.

1.5 Problem statement

Through computer simulations it is possible to investigate interactions between chemical compounds, among others host-guest interactions. By analysing the resulting energies it is possible to find which complexes are feasible and which are not. Although this can be done by older programs and force fields like AMBER and GLYCAM06, these programs have a steep learning curve.

A relatively new program OpenMM features a much more readable syntax and is completely integrated with Python. This makes it a viable option for non-programmers. It also features the recent force field SMIRNOFF which applies a relative new method called "direct chemical perception" that find the chemical environment from the chemical graph rather then a list like the older programs.

The goal of this project is as follows:

Determine and define a feasible method to apply the biasing potential on an inclusion complex between β-cyclodextrin and different ligands using OpenMM.

Theory 2

This chapter will cover the theory necessary to develop and analyse umbrella sampled simulations. First the biasing energy potential used to separate the host and ligand will be introduced. To figure out which energy ensemble should be used, the two free energies Gibbs and Helmholtz will be discussed. Bennett's Acceptance Ration (BAR) will also be discussed briefly as it is needed to analyse the results of the Umbrella Sampling (US).

2.1 Biasing energy potential

As stated in the introduction the goal of this project is to investigate the inclusion complex of a host and ligand through computer simulation. A relatively easy method is to apply a biasing energy potential (Equation (2.1) same as Equation (1.23)) between the Center of Mass (CoM) of the two molecules (You et al., 2019).

$$E_{bias} = \frac{1}{2} \cdot K \cdot (R - R_0)^2, \tag{2.1}$$

where E_{bias} is the energy potential as a function of the measured distance R, K is a force constant and R_0 is the starting distance. The bigger difference in between R and R_0 will result in a higher potential energy E_{bias}. This difference is influenced (mainly) by the forces of attraction and repulsion i.e., intermolecular forces between the host and ligand and as such the measured distance can be used to quantify the potential energy needed to keep a certain distance. The higher the distance is to either side the more skewed the energy will also be to either side. If the distances has a normal distribution around the minima, the reaction will be in an equilibrium (see Figure 2.1).

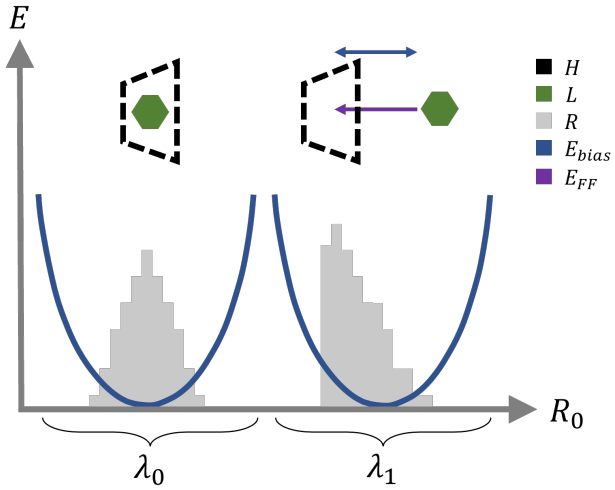


Figure 2.1. Illustration of the energy E_{bias} difference in λ_0 and λ_1 as a function of the measured distance R between CoM of H and L. R_0 is the "set" distance between host and ligand.

Figure 2.1 illustrates the biasing energy potential E_{bias}. At λ_0 the host, H and ligand L are centered at the same point in space and R is distributed normally around the minima of E_{bias}. At λ_1 R_0 is greater, however the two molecules are still close enough to interact. This results in a left skewed distribution of R, increasing E_{bias}.

2.2 MBAR

Before going into too much detail, it is important to note that Bennets Acceptance Ratio (BAR) was defined in 1976, when MD simulations was not yet on par with NVT (the volume is kept constant). This means that the free energy was based on Helmholtz free energy A. With this in mind, the following section shows that Gibbs free energy can just as easily can be used.

2.2.1 Defining the energy of the system

When working with simulation of chemical systems, the aim is to achieve results similar to those found through experimental laboratory work. Often the purpose of both experiments and simulations are to determine whether a reaction is spontaneous or not. A good measure for this are the state functions as described in the introduction. When the change in energy from solvated molecules to inclusion complex is negative, the reaction is spontaneous. The energy of a reaction can be described by Gibbs free energy and Helmholtz free energy, which in many cases are interchangeable.

Gibbs free energy, G, is defined as Equation (2.2)

$$G \equiv H + TS, \tag{2.2}$$

where H is the enthalpy, T is the temperature and S is the entrpy. The change in Gibbs Free Energy ΔG is written as Equation (2.3)

$$\Delta G = \Delta H + T\Delta S \tag{2.3}$$

$$\Delta G = U + PV + T\Delta S, \tag{2.4}$$

where U is the internal energy or the potential energy, P is the pressure and V is the volume. The free energy can also be defined as the Helmholtz Free energy (Equation (2.5)). (Tinoco et al., 1999)

$$A \equiv E - TS \tag{2.5}$$

Here E is the energy of the system. At constant temperature and *pressure* $\Delta G = 0$ while $\Delta A = 0$ at constant temperature and *volume*. (Tinoco et al., 1999) Two generally chosen ensembles in MD simulations are NPT (constant mol, pressure, temperature) and NVT (constant mol, volume, temperature), the first resulting in Gibbs Free energy (Equation (2.4)) and the latter giving Helmholtz Free Energy (Equation (2.5)) (GROMACS development team, 2018). Laboratory experiments are mostly carried out as open systems with ambient pressure. Therefore, when aiming to predict the energies of these experiments through MD simulations, the NPT ensemble is used.

2.2.2 Introducing BAR

BAR can be used to find the energy difference between two states, λ_l and λ_{l+1}, which have to be close to each other. For BAR to be a viable method, the energy difference cannot be too big, the energy potentials have to overlap and each λ has to be sampled independent of each other. The same restrictions apply to Multistate BAR (MBAR), however this method, as the name implies, is extended to handle more λ than just two. The application of MBAR lets the energies of each sampled λ to affect the others until some level of precision is reached. This allows for interpolation of the energy between each λ, predicting the energy with fewer samples. (Shirts and Chodera, 2008; König, 2010)

$$\Delta G(\lambda_l \to \lambda_{l+1}) = k_b T \left(\frac{\sum_{\lambda_{l+1}} f(U(\lambda_l) - U(\lambda_{l+1}) + C)}{\sum_{\lambda_l} f(U(\lambda_{l+1}) - U(\lambda_l) - C)} \right) - \ln \frac{n_{\lambda_{l+1}}}{n_{\lambda_l}} + C \tag{2.6}$$

n_{λ_l} and $n_{\lambda_{l+1}}$ is the number of coordinate states at λ_l and λ_{l+1}, respectively and C is ratio of the classical partition functions Z_{λ_l} and $Z_{\lambda_{l+1}}$ given by:

$$C = k_b T \ln \frac{Z_{\lambda_l} n_{\lambda_{l+1}}}{Z_{\lambda_{l+1}} n_{\lambda}} \tag{2.7}$$

and f is the Fermi Function

$$f(x) = \frac{1}{1 + \exp \frac{x}{k_b T}} \tag{2.8}$$

As the C Equation (2.6) is an unknown factor we cannot benefit from the expression in its current form. However, by iterating Equation (2.9) until satisfied, the value of C can be found thereby making Equation (2.6) applicable. (Shirts and Chodera, 2008; König, 2010)

$$\sum_{\lambda_{l+1}} f(U(\lambda_l) - U(\lambda_{l+1}) + C) = \sum_{\lambda_l} f(U(\lambda_{l+1}) - U(\lambda_l) - C) \tag{2.9}$$

2.2.3 Sampling for BAR

As mentioned BAR works on exactly two different configurations as will be shown in the following. Say you have two configurations, λ_l and λ_{l+1}, each with a different starting condition and you want to move λ_l to λ_{l+1}. This requires some amount of energy, E_{bias}. As described in the introduction, E_{FF}, contains the energy contribution from the bonded and non-bonded forces meaning that we can write the energy of the system as $E_{sys} = E_{FF} + E_{bias}$.

$$E_{sys} = E_{FF} + E_{bias} = U \tag{2.10}$$

Because of Equation (2.10), Equation (2.6) can be extended to Equation (2.11):

$$\Delta G^{bias}(\lambda_l \to \lambda_{l+1}) = k_b T \frac{a}{b} - \ln \frac{n_{\lambda_{l+1}}}{n_{\lambda_l}} + C, \tag{2.11}$$

where ΔG^{bias} is the free energy of E_{bias}. For the sake of transparency the numerator and denominator in Equation (2.11) is assigned to a and b, respectively.

$$a = \sum_{\lambda_{l+1}} f(E_{bias}(\lambda_l) - E_{bias}(\lambda_{l+1}) + C) \tag{2.12}$$

and

$$b = \sum_{\lambda_l} f(E_{bias}(\lambda_{l+1}) - E_{bias}(\lambda_l) - C) \tag{2.13}$$

From the above we can assign a part of the free energy to the biasing potential:

$$\Delta G^{bias}(\lambda_l \to \lambda_{l+1}) = k_b T \left(\frac{\sum_{\lambda_{l+1}} f(E_{bias}(\lambda_l) - E_{bias}(\lambda_{l+1}) + C)}{\sum_{\lambda_l} f(E_{bias}(\lambda_{l+1}) - E_{bias}(\lambda_l) - C)} \right) - \ln \frac{n_{\lambda_{l+1}}}{n_{\lambda_l}} + C \quad (2.14)$$

The application of BAR is done in two main steps: 1) the free energy of λ_l and $\lambda_{(l+1)}$ is sampled and the average is found, 2) the energy of the two configurations are interpolated and a new average is found. When applying MBAR, (step 2) BAR is extended to depend on more configurations.

The reduced energy matrix **A** is made from the sampled distances. The size of **A** is MxN, where M is the number of frames and N is the number of samples in each frame. A is dimensionless. FastMBAR is used to find the potential mean force (PMF). To extend the PMF into pertubated PMF the matrix LxN is created. When FastMBAR has finished, the number of energies are the same as L. For an energy to be accepted into L, the energy difference cannot be too big, otherwise it will result in $+\infty$ and be discarded. The diagonal of LxN is the PMF found from A, while over and under are found both through iterations of C and comparison of the energies (if they are too big they will be counted as $+\infty$ as they would be too big to fit the system). (Ding et al., 2017, 2019a)

2.3 Umbrella sampling

Umbrella sampling starts with something resembling equilibrium (closed umbrella) and then pulls or moves the system until a new equilibrium is found (fully open umbrella). In the case of inclusion complexes, the guest is pulled from the host frame by frame. For each frame some change in the system is measured and in this case it is the distance between the two. To pull the host and guest apart, a starting system is made for each distance, i.e. R_0 (Equation (2.1)) is set to a specific distance. This allows the diverging distance to be measured for each snapshot in the frame. As all of the frames are unique and have their own starting configurations the frames can be said to be independent of each other, while the snapshots of each frame are not.

Method 3

This chapter aims to show how Python can be used to implement the OpenMM framework for simple simulations and umbrella sampling. The simulation setup and the ligands in question will be show and lastly the resulting configurations/snapshots are then analysed by implementing FastMBAR in python.

3.1 Setting up an OpenMM simulation

The following section presents an introduction to the OpenMM framework and how to programmatically perform simulations through the Python API.

When starting any simulation the simulation object must be built. It concist of a molecule, a force field, a system, and an integrator. It is possible to customize all of these but here we will focus on the simulation object and only the applied settings will be described in detail.

3.1.1 Setting up the `Molecule` and `ForceField`

The `ForceField` object contains one or more forcefields which must be submitted as xml-files or the SMIRNOFF version offxml-files.

The force field must contain the correct molecules as the topology is otherwise unknown to the `ForceField` object. The molecules can be added into the force field by multiple ways including SMILES strings and sdf-files. It is possible to add more molecules or atoms to the force field than needed like the two ions Na^+ and Cl^- but one should of course not add unnessary clutter. A complete list of ways to create the molecules can be found in the documentation (Open Force Field Consortium, 2022).

In Listing 3.1 below, the molecule is imported from either a smiles string (ions) or a sdf-file (host/guest molecules). The `ForceField` only contains the molecules which are added through the `add_molecules` method.

Listing 3.1. OpenFF toolkit and Molecule

```
1  from openff.toolkit.topology import Molecule
2  from openmmforcefields.generators import SMIRNOFFTemplateGenerator
3
4  smirnoff = SMIRNOFFTemplateGenerator(forcefield='openff-2.0.0.offxml')
5
6  # Adding the two ions Na+ and Cl- as SMILES strings to smirnoff
7  # The two ions are added to ensure charge neutralization if the added molecules has a charge
8  # The ion will only be used if nesseary
```

```
9
10      ion_smiles = ['Na+', 'Cl-']
11      for ion_smile in ion_smiles:
12          ion = Molecule.from_smiles(ion_smile)
13          smirnoff.add_molecules(ion)
14
15      # Adding the host/guest molecules to smirnoff
16      sdf_files = ['host.sdf', 'guest.sdf']
17      for i in range(len(sdf_files)):
18          offmol = Molecule.from_file(sdf_files[i])
19          smirnoff.add_molecules(offmol)
```

`ForceField` is imported from the OpenMM and is acting in the foreground. The three xml-files which the force field is generated from loads the specific parameters of the molecules. `'amber14-all.xml'` is a dependency of both `'tip3p.xml'` and `'GLYCAM_06j-1.xml'`. The TIP3P is added to be able to parameterise water as a solvent and GLYCAM06 is added to provide the parameters of β-CD, specifically the residue 4GA (1,4$-\alpha$-glucose).

Listing 3.2. Setting up the OpenFF forcefield

```
20      # Setting up the OpenMM forcefield and adding the OpenFF forcefield containing the molecules
21      # This lets the OpenFF force field work its magic in the background
22      from openmm.app import ForceField
23
24      forcefield = ForceField('amber14-all.xml', 'tip3p.xml', 'amber/GLYCAM_06j-1.xml')
25      forcefield.registerTemplateGenerator(smirnoff.generator)
```

The `forcefield` is now ready to handle the water model tip3p and the GLYCAM06-j parameters which will be added next and the system is ready to be set up.

3.1.2 Setting up the system

Modeller

The modeller provides a representation of the molecule(s). To function properly, the force field **must** contain the molecules in question. The modeller takes a the topology and positions of the molecules, which can be loaded through the OpenMM's PDBFile. The topology and positions are then transfered to the `Modeller` object

```
26      from openmm.app import PDBFile
27      from openmm.app import Modeller
28
29      pdb_file = PDBFile('pdb_file.pdb')
30      modeller = Modeller(pdb_file.topology, pdb_file.positions)
```

To add water to the system, some volume is needed. The easiest is to use the `padding` command, which takes some distance in nanometer as argument. However, to make a NPT ensemble periodicity must be enforced and this requires well defined borders. The `boxVectors` is used to define these and takes a tuple of three `n3` vectors by `Vec3()`.

```
31      from openmm.vec3 import Vec3
32
33      a_vec = Vec3(5, 0, 0)
34      b_vec = Vec3(0, 5, 0)
35      c_vec = Vec3(0, 0, 5)
36      modeller.addSolvent(forcefield, model='tip3p', boxVectors=(a_vec, b_vec, c_vec))
```

If nothing else is given `modeller.addSolvent()` will automatically add counter ions to neutralize the system if nesseary. The counter ions are by default Na^+ and Cl^-.

3.1.3 Making a system

The simplest system possible in OpenMM is made by the `forcefield.createSystem()` and only need a topology as input which is taken from the `modeller`. However, this is not enough when creating an NPT ensemble. To do so the non-bonded method need to be set to Particle Mesh Ewald (PME), which ensures periodicity and makes it possible to keep the pressure of the system constant.

```
37  from openmm.app import PME
38
39  system = forcefield.createSystem(modeller.topology, nonbondedMethod=PME)
```

Adding forces to a system

When working with the interactions of two or more molecules a force is needed. This can be added to the system by the `system.addForce()`. A custom force was be built by the `CustomCentroidBondForce()`, specifying the amount of groups (at least 2) and a energy function. A way to define the groups will be covered in detail later.

In this project two different force functions were utilized in the system; one to center the host and guest molecules on top of each other (Equation (3.1)) and one to keep them at a specific distance (Equation (3.2)):

$$F_{center} = K_{center} \cdot |R - R_0| \tag{3.1}$$

$$F_{pull} = \frac{1}{2} \cdot K_{pull} \cdot |R - R0|, \tag{3.2}$$

where R is the measured distance between the two groups `g1` and `g2`. The keyword `distance(g1, g2)`, is specific for `CustomCentroidBondForce` and finds the centroid of the two groups e.g. the weighted geometrical center and then the distance between the two.

Please note that the force function given in the `CustomCentroidBondForce()` is **not** written with pythons normal math syntax and is written as a string. In `pull_force` the function is further split up in three parts, each devided by a semicolon (;). The method does not understand the phrase `0.5*K_pull*( (distance(g1, g2)- R0)^2 )`, hence the difference in syntax.

Listing 3.3. Setting up `pull_force`

```
40  from openmm.openmm import CustomCentroidBondForce
41  from openmm.openmm import unit
42
43  pull_force = CustomCentroidBondForce(2, '0.5 * K_pull * dR^2; dR = (R-R0);
44      R = distance(g1, g2)')
45  pull_force.addGlobalParameter('K_pull', 0.0 * unit.kilocalories_per_mole /
46      1.0 * unit.angtroms ** 2)
47  pull_force.addGlobalParameter('R0', 0.0)
```

The two global parameters `K_pull` and `R0` were added as global parameters to be able to update them later on. When the system is finished they can be changed by the method `system.method.setParameter('R0', R0)`, where the first argument works similar to a libary and the second argument is a variable.

The `host` and `guest` were added to `pull_force` as two sets of indecies. If these are taken from a pdb-file the indecies must be arrays of the lines containing the relavant groups.

Listing 3.4. Adding a force to a system object

```
48    host = pull_force.addGroup(host_indecies)
49    guest = pull_force.addGroup(guest_indecies)
50    pull_force.addBond([host, guest])
51    system.addForce(pull_force)
```

Almost the same procedure is used for the centering force with the exception of the variable `K_center`, which was the only global parameter.

Listing 3.5. Setting up `center_force`

```
center_force = CustomCentroidBondForce(2, 'K_center * distance(g1, g2)')
center_force.addGlobalParameter('K_center', 0.0 * unit.kilocalories_per_mole / 1.0 *
    unit.angstroms ** 2)
```

OpenMM has no trouble reading PDB-files from Maestro, however the output differs. OpenMM will include keywords like MODEL, TER and ENDMDL and also CRYST1 if water is added. From a simulation point of view this does not mean much but if one wishes to extract data from the PDB-files some programming is needed.

Integrators

Integrators take care of moving the simulation along in time steps. The simplest version is the `VerletIntegrator(stepSize)` which takes the step size as its only argument. As it is quite light it is a good choice for debugging the `Simulation()`. We will, however, use the `LangevinMiddleIntegrator(temperature, stepSize, frictionCoef)` instead. The LangevinMiddleIntegrator obviously provides more customisation off the integrator (temperature, step size and friction coefficient) and also simulates the use of a heat bath to keep a constant temperature in the simulation.

When creating the integrator an object is made. It can either move the simulation along on its own or be attached directly to a simulation object. The integrator itself is independent of the molecules, forcefield and system.

Simulation

The `Simulation` object is the simplest way to produce simulations. A number of methods are directly attached to the system, where the most prominent are the`Simulation.minimizeEnergy()` and `Simulation.step()` but also writing `pdb`-files and printing how the simulation is moving along with number of steps, energy parameters, volume, and more to either a file or the console

The unit system used in OpenMM

The unit system in OpenMM is based on ParmED's UnitSystem (Shirts et al., 2016). The only difference is that the loading command is `from openmm.openmm import unit`. `unit()` has a wide array of possible units already set up, with the possiblity to create more if needed. To convert a unit, say joule to kilo calories, the method `in_units_of(unit())` should be used (Listing 3.6)

Listing 3.6. Changing units in OpenMM

```
G_joule_per_mole = 4184 * unit.joule / unit.mole
G_kcal_per_mole = G_joule_per_mole.in_units_of(unit.kilocalorie_per_mole)
```

$$G_{joule\_per\_mole} = 4184\,\mathrm{J\,mol^{-1}} \tag{3.3}$$

$$G_{kcal\_per\_mole} = 1\,\mathrm{kcal/mol} \tag{3.4}$$

3.2 Simulation setup

The simulation protocol follow You et al. (2019) in regards to the minimisation, equalibration and production. The step length in the displacement was, however increased from 0.1 Å to 0.25 Å.

For all simulations the temperature was set to 298.15 K, the friction coefficient was set to $10.0\,\mathrm{ps^{-1}}$, and the step size was set to 1 fs.

3.2.1 Starting point

The starting point of all simulations containing a complex was a pdb file with the host and chosen ligand in close proximity. The two molecules were first minimised for 4000 steps and then pulled towards each other's center by a force constant K_{pull} of $10\,000\,\mathrm{kcal\,mol^{-1}\,Å^2}$. After centering the molecules on top of each other, they were displaced to orego of the coordinate system.

3.2.2 Displacement

The ligand was displaced from the β-CD in both the direction of the primary and secondary end. For each displacement of the ligand, the ligand was displaced and the complex was minimised in place. The displacement was made by moving the ligand to a set distance R_0 in Python. The complex was then minimised for 1500 steps with a force constant K_{pull} of $1000\,\mathrm{kcal\,mol^{-1}\,Å^2}$. The ligand was displaced either 13 Å (long simulation time)) or 26 Å (short simulation time) with a increment of 0.25 Å

3.2.3 Production

For the production run a barostat was added to ensure periodic boundaries and the molecules were solvated in water. The pressure was set to 1 bar and the box size was set to 50 Å on each side. The system was minimised for 1000 steps and then equilibrated for 1 ns with a force constant K_{pull} of $1000\,\mathrm{kcal\,mol^{-1}\,Å^2}$.

Short simulation time

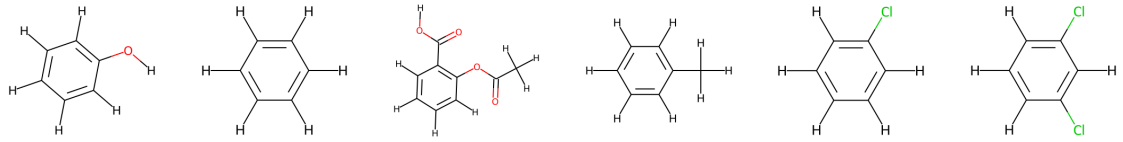
For the short production the system was simulated for 2.5 ns with a K_{pull} of 100 kcal mol^{-1} Å^2. A snapshot of the distance was taken every 1250 steps totalling to 2000 snapshots per simulation.

Long simulation time

The system was simulated for 5.0 ns with a K_{pull} of 100 kcal mol^{-1} Å^2. A snapshot of the distance was taken every 1000 steps totalling to 5000 snapshots per simulation.

3.3 Chosen ligands

For this project, six compounds were chosen as the ligands in connection with *beta*-cyclodextrin (Figure 3.1). All of the ligands have benzene as a base.



(a) ***(b)*** ***(c)*** ***(d)*** ***(e)*** ***(f)***

Figure 3.1. Six molecules, all with benzene as the central structure, used as ligands in the simulations of the host-guest interactions with β-CD. The ligands are as follows: ***(a)*** phenol, ***(b)*** benzene, ***(c)*** aspirin, ***(d)*** toluene, ***(e)*** chlorobenzene, and ***(f)*** 1,3-dichlorobenzene.

3.4 FastMBAR

Implemention of FastMBAR is done in four main steps (Ding et al., 2019b):

1. The files containing the snapshot information are read and assigned to a list. The number of snapshots N per frame is recorded.
2. The reduced energy matrix, $A = [MxN]$ where M is the number of frames, is made by the absolute difference between the set distance, R_0, and the actual distance R.
3. FastMBAR is setup with the reduced energy matrix and number of conformations resulting in free energies for each of the frames.
4. $L = mM$ is introduced to the matrix $B = LxN$ which is used to calculate the free energies of the perturbed states. This means that the energies of all states are spread out on a wider 'area'.

Results and discussion 4

This chapter contains the results and a following discussion of these. First the data will be compared internally by a large number of samples and bootstrapping. Then the different ligands will be compared with each other and the literature.

4.1 Data convergence

When doing molecular dynamic simulations time is of the essence. It is important to use a sufficient amount of time to get reliable results while at the same time, the simulation should also be within a realistic time span. To this end, the long simulation protocol (Section 3.2.3) was applied to the complex containing phenol and β-CD and the snapshots were sampled 5000 times. The potential mean force was found by applying FastMBAR (Section 2.2) to the measured distances. The FastMBAR results were bootstrapped 1000 times with replacement. From the 5000 snapshots a number of subsets were made $n_{snapshots} = [10, 20, 100, 250, 500, 750, 1000, 2000, 3000, 4000, 5000]$.

The simulations were based on an umbrella sampling with $M = 52$ windows. The energies from the umbrella sampling were then divided into n_{bins} by the following scalars $[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 30, 35, 40]$ resulting in $n_{bins} = [52, 104, 156, 208, 260, 312, 364, 416, 468, 520, 780, 1040, 1560, 1820, 2080]$. In total 165 combinations of n_{bins} and $n_{snapshots}$ were made. As the main interest is the change in energy from solution to complex (Equation (1.14)), the minima of the complex G_{LH} were found in the range 0 Å to 4 Å and a minimum in solution $G_L + G_H$ was found in the range 8 Å to 13 Å (see Figure 1.6 for an example). For evaluation purposes the smallest value was subtracted from the data set containing the free energy ΔG to give the relative change in free energy $\Delta\Delta G$.

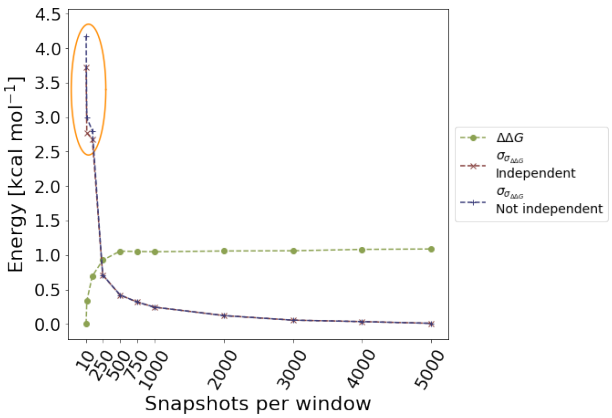


Figure 4.1. The change in free energy and standard deviation of phenol and β-CD in inclusion complex.

Figure 4.1 show the energy and two standard deviations of phenol complexing with β-CD from 0 Å to 13 Å as a function of number of snapshots per window and a bin size of 1 per window. The smallest energy was found with a sample size of 10 snapshots per window. The energy converges at $1\,\mathrm{kcal\,mol^{-1}}$ with a sample size of 500. Each of the standard deviations are the standard deviations of the standard deviations $\sigma_{\sigma_{\Delta\Delta G}}$ found by FastMBAR.

When applying FastMBAR to a set of data, each data point has to be independent of the other. After implementing FastMBAR, the resulting free energy is no longer independent due to how the states λ_l and λ_{l+1} are handled (Equation (2.14)) (Shirts and Chodera, 2008; Ding et al., 2019a). While the free energies can simply be added to each other, this is not the case for standard deviations. Instead each $\sigma_{\Delta\Delta G}$ has to be changed into the variance, added together and then changed back into $\sigma_{\sigma_{\Delta\Delta G}}$ (Lane, 2022).

To test if $\sigma_{\Delta\Delta G}$ were independent of each other Pearson's correlation value were found for Equations (4.1) and (4.2). If $\sigma_{\Delta\Delta G}$ are independent of each other $\sigma_{\sigma_{\Delta\Delta G}}$ can be found by Equation (4.1).

$$\sigma_{\sigma_{\Delta\Delta G}} = \sqrt{s_{\sigma_1}^2 + s_{\sigma_2}^2}, \tag{4.1}$$

where s_{σ_1} and s_{σ_2} are the variance of the standard deviation $\sigma_{\Delta\Delta G}$. If, however the $\sigma_{\Delta\Delta G}$ are *not* independent Equation (4.2) are used instead.

$$\sigma_{\sigma_{\Delta\Delta G}} = \sqrt{s_{\sigma_1}^2 + s_{\sigma_2} + 2p \cdot s_{\sigma_1} \cdot s_{\sigma_2}}, \tag{4.2}$$

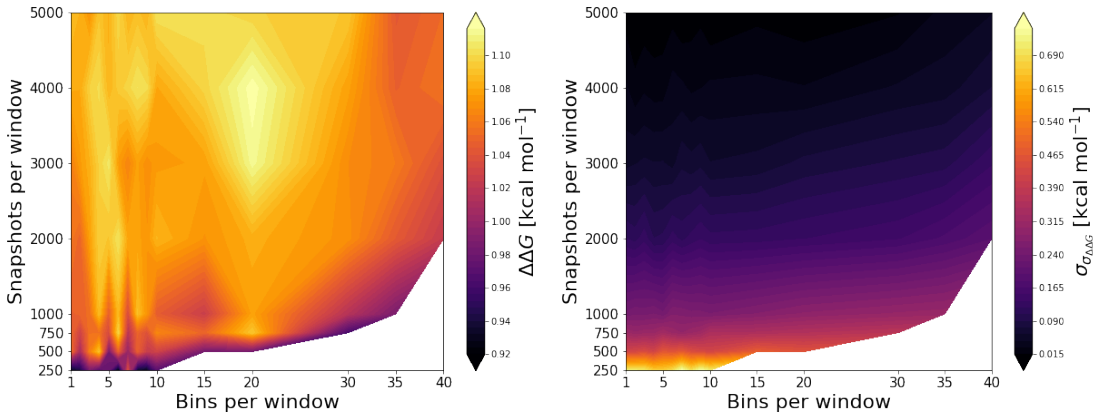
p is Pearsons correlation value (Lane, 2022). While there were a small discrepancy in $\sigma_{\sigma_{\Delta\Delta G}}$ at low sample sizes (see the orange circle at Figure 4.1), the two methods cannot be told apart at sample sizes larger than 250 snapshots. The correlation values were found to be between 0.09 and 0.21. Equation (4.2) is therefore discarded and Equation (4.1) will be applied for any further investigations.

At sample sizes of 10, 20 and 100 (orange ellipse), the standard deviations are about the same as the corresponding free energies, hence the sample sizes are too small and will not be a part of the further investigation. At sample sizes of 2000 snapshots to 5000 snapshots the standard deviations ranges from $0.29\,\mathrm{kcal\,mol^{-1}}$ to $0.19\,\mathrm{kcal\,mol^{-1}}$.

Although Figure 4.1 shows the relationship between the free binding energy and the standard deviation as a function of snapshots, it does not consider the bin size. To evaluate how the bin size and number of snapshots affect the energy and standard deviation, they were plotted against each other in Figure 4.2.

Figure 4.2 shows two contour plots with the change in free energy and standard deviation of phenol complexing with β-CD from 0 Å to 13 Å as function of bins per window (x-axis) and snapshots per window (y-axis). The white areas show the combination of bin size and number of snapshots per window where the energy cannot be calculated (Not a Number (NaN)). For instance at 250 snapshots per window the bin size has to be lower than 15 bins per window, to produce meaningful data. The NaN's were sorted based on the data of the PMF's; if even one NaN was present in the 1000 bootstraps both the data set of the PMF and standard deviation were discarded.

The energy scale of Figure 4.2a goes from dark (low energy) to bright (high energy). The lowest energies are found with a sample size of 250 to 750 snapshots and increasing sizes (1 to 30 bins per window), while the highest are found with a sample size of 2000, 3000 and 4000 snapshots with a bin size ranging from 6 to 20 bins per window. The energy seems to stabilise with a sample size bigger than 2000, at least when the number of bins per window are below 30. When looking at the change in free energy, it will always be tempting to pick the lowest energy. It is, however, important to take the standard deviation into consideration (Figure 4.2b). This shows that with a sample size of 250 snapshots yields a standard deviation which is about a quarter of the corresponding change free energy. A sample size of 5000 snapshots on the other hand yield the lowest standard deviation at all bin sizes.



(a) *(b)*

Figure 4.2. The change in free energy of the inclusion complex between phenol and β-CD and the corresponding standard deviations of a bootstrap with a total of 165 data points. ***(a)*** shows the change in energy $\Delta\Delta G = (G_1 - G_0) - G_{min}$ from solution to complex. ***(b)*** shows the standard deviation of the $\Delta\Delta G$.

From 2000 snapshots to 5000 snapshots the change in relative free energy is very small and likewise for the standard deviation. If more samples equals better, then a sample size of 5000 snapshots would inherently be better than one with 2000 snapshots. While this *is* the case, at least when considering the standard deviation, the increase in precision is negligible (the difference is approximately 0.16 kcal Å^{-1}). The measured energy is also about equal with a sample size above 2000.

4.1.1 Partial conclusion

In the above it is shown that the energies and standard deviations converge with a sample size of more than 750 depending on the bin size. With an increasing bin size it is necessary to sample more snapshots to get meaningful answers. This makes sense as there should be a certain amount of data per bin in order to trust the results.

All of the different sample sizes were split into the aforementioned bin sizes. This could either produce some number or a NaN, depending on the sample size versus bin size. With a sample size of 10 snapshots, only the bin size of 1 yielded a number. In fact the sample

size had to consist of 2000 snapshots or more before all the bin sizes (the maximum was 40 bins) yielded a number. With a sample size of 2000 and a bin size of 20 or 40 per window, there would be an average of 100 or 50 snapshots per bin in each window, respectively. The general thought is that more data equals higher precision, however the increase in precision vs. time spent collecting the data also has to be considered. A sample size of 2000 snapshots and a bin size of 20 per window were found be a good compromise between precision and time spent collecting the data.

4.2 Analysing phenol in inclusion complex with b-CD

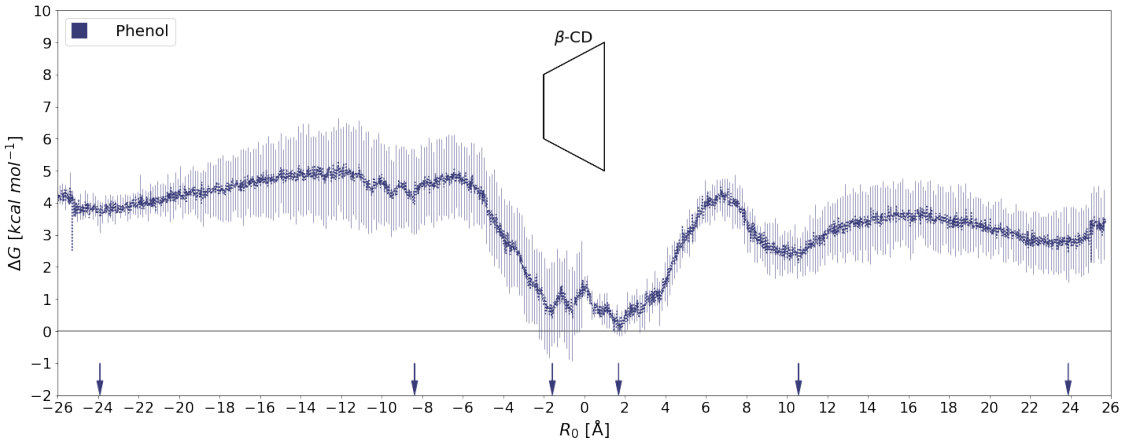
The focus of the analysis above was to find sufficient analysis parameters through data convergence. The following will focus on the direct implications of letting phenol (Figure 3.1a) forming an inclusion complex with β-CD. For readability, the minimum of ΔG was normalised to zero.

The two PES shown in Figure 4.3 are based on the same data, although the method differs a bit. The energy of the secondary PES of Figure 4.3a is found by moving $R_0 = -26\,\text{Å}$ to $R_0 = 0\,\text{Å}$ whereas the secondary PES of Figure 4.3b is found starting with $R_0 = 0\,\text{Å}$ moving to $R_0 = -26\,\text{Å}$. It is clear that the initial starting point affects where on the PES the standard deviation appears even though (on visual inspection) it appears to increase with the same rate. The minima in the free energy of both figures appear at the same R_0 as will be covered in the following. The primary part of the PES is produced from the same data and moves from 0 Å to 26 Å.

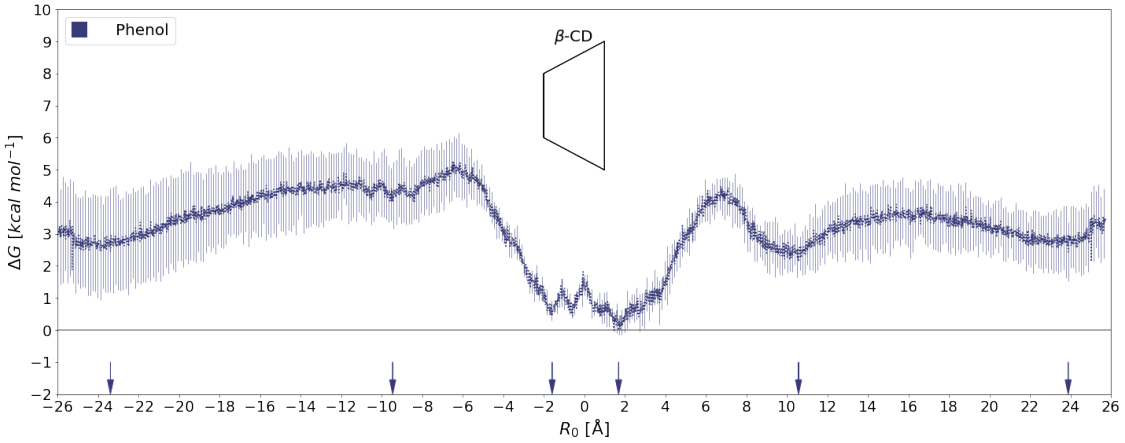
Figure 4.3a shows the free energy of binding between phenol and β-CD in the interval $-26\,\text{Å}$ to $26\,\text{Å}$. Due to computation limits, the analysis was split in two from $-26\,\text{Å}$ to $0\,\text{Å}$ and $0\,\text{Å}$ to $26\,\text{Å}$. The arrows along R_0 show where the minima were found in the following ranges: $\pm[0:4], \pm[8:12], \pm[14:24]$. The β-CD shape shows how the ligand was pulled in relation to β-CD.

Moving from $R_0 = 0\,\text{Å}$ in the secondary direction first shows a minimum of $\Delta G = 0\,\text{kcal mol}^{-1}$ at about $R_0 = 2\,\text{Å}$ after which the energy increases to $\Delta G = 4.5\,\text{kcal mol}^{-1}$ with a sharp peak ($R_0 = 7\,\text{Å}$). At $R_0 = 10.5$ the next minimum is found $\Delta G = 2.25\,\text{kcal mol}^{-1}$ and the energy rises slowly again until $R_0 \approx 16\,\text{Å}$ where it falls again towards $R_0 = 24\,\text{Å}$. In the primary direction there are two sharp minima from $R_0 = 0\,\text{Å}$ to $-2\,\text{Å}$ showing the supposedly best binding distances in this direction, although the binding energy is better in the secondary direction. The slope between the minimum and peak at $R0 = -2\,\text{Å}$ to $-7\,\text{Å}$ seems very linear when compared to the primary direction, which might be caused by the narrow pathway and lesser room for rearrangement of the ligand inside the cavity of the host. In the range of R_0 from $-8\,\text{Å}$ to $-12\,\text{Å}$ there are several energy minima where one might expect a single larger one. The energy decreases from there towards $R_0 = -24\,\text{Å}$.

Even though there are obvious differences in the behaviour depending on the primary and secondary pathway of phenol leaving β-CD, the minima are found to be in the same regions. This can probably be ascribed to the difference in size of two ends β-CD. The



(a)



(b)

Figure 4.3. Phenol in inclusion complex with β-CD. The arrows show the different minima found along R_0. ***(a)*** ΔG was found from -26 Å to 0 Å (primary path) and 0 Å to 26 Å (secondary path) while ***(b)*** was found from 0 Å to ± 26 Å (both paths). This means that the initial values were different, mainly affecting the standard deviation but also the free energy path.

behaviour of the standard deviation on the other hand is more difficult to reason for (Figure 4.3b). As stated earlier, the data was split in two due to computational limits. As the two data sets are calculated independently of each other, this probably accounts for the discontinuity of the standard deviation. The energy calculations of the two pathways were started at 0 Å towards -26 Å and 0 Å towards 26 Å for the primary and secondary path, respectively. The standard deviations in the primary pathway seem to become increasingly larger, the further from $R_0 = 0$ they move. A possible explanation is that the calculations of FastMBAR becomes more uncertain as the energy diverges from its origin. On a side note, it is interesting that the standard deviations of the primary path in both Figures 4.3a and 4.3b starts to increase after approximately 5 Å.

In conclusion, the secondary pathway shows a better binding affinity between phenol and β-CD (Figure 4.3b). The height of a β-CD is about 8 Å, which might explain the following minima in free energy, at least in the secondary pathway (Davis and Brewster, 2004). If the benzene ring interacts with the relative open cavity of β-CD and the alcohol interacts with the water, it can explain the minima. It might also because the alcohol of the phenol

interacts with one of the secondary alcohols while the benzene ring of the phenol interacts with the cavity. The standard deviation also seem to decrease in the region of the different minima, suggesting that the complex becomes more stable. The corresponding minima of the primary path are not as distinct, which might be ascribed to the primary end of β-CD being more narrow than the secondary. This might decrease the possibilities for interactions between the benzene ring and the cavity of the β-CD. The standard deviations slowly increase with the distance from $R_0 = 0$ in the primary pathway.

After 25 Å in either direction there is a drop and peak in the energy which seem out of order with the rest of the data (Figure 4.3b). This is thought to be an artefact caused by the boundary conditions being overstepped; the simulation run for 26 Å in either direction. As the box size is 50 Å on all sides and has periodic boundaries the simulation will able to "feel" itself when half of the box size is overstepped. For this reason the boundary will be set to be ±25 Å in the following analyses.

4.2.1 Benzene and β-CD

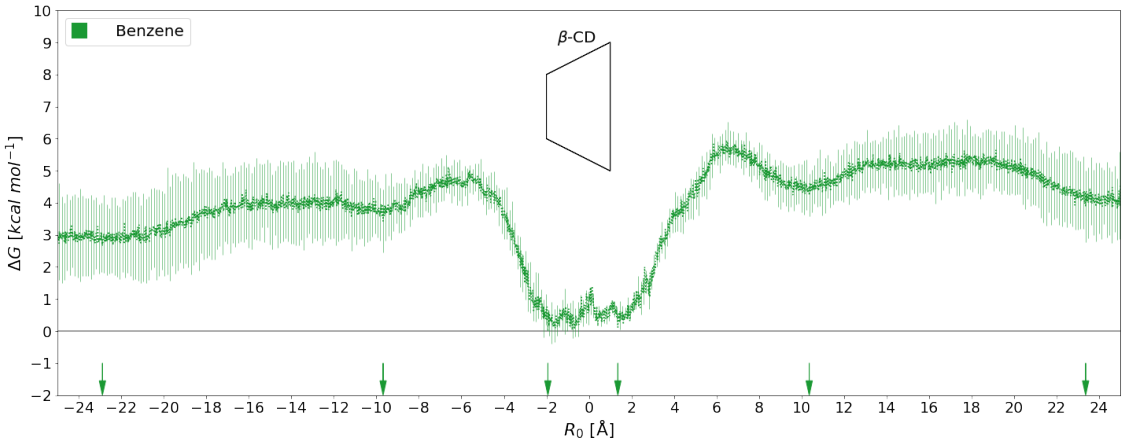


Figure 4.4. Benzene

The only difference between the two molecules, benzene (Figure 3.1b) and phenol, is that in phenol one hydrogen is replaced by an alcohol group. This makes benzene less soluble in water compared to phenol and it would make benzene more energetically favoured inside the cavity rather than outside. Figure 4.4 shows the free energy between benzene and β-cyclodextrin. While the PES is similar to the one of phenol and β-CD, there is a slight disconnect in the free energy at $R_0 = 0$ between the primary and secondary path. This is probably an artefact owing to the fact that the primary and secondary path are two different data sets.

The first half of the slope in the secondary direction from $R_0 \approx 2$ Å has quite a steep increase and lessens at $R_0 \approx 4$ Å. This might owe to the height of β-CD (4 Å on each side of $R_0 = 0$), as the benzene get more room to wiggle and find a less strained placement. The following peak at $R_0 \approx 7$ Å is relative sharp at the left side, while smooth at the right side. There is a clear minimum at $R_0 \approx 10$ Å after which the free energy is relatively stable until $R_0 \approx 20$ Å where it drops towards $R_0 = 24$ Å. The free energy

of the primary path is lower than the secondary path although the minima are located in the same regions (the first minimum at both paths are almost indistinguishable). The free energy has approximately the same drop after the first peak ($R_0 \approx 5.5\,\text{Å}$). The primary path towards the first peak seem more linear (much like phenol), than its corresponding slope at the secondary path, probably owing to the lack of room. It is interesting to note that the standard deviation is higher for the primary path than the secondary. So while the more energetically favourable path is the primary, the higher standard deviation of the free energy indicate that it is more unstable. The lower free energy in the primary pathway might be ascribed to the less exposed cavity of the β-CD, reducing the interactions between the guest and host, thereby lowering the pull from the non-polar inside of β-CD.

Table 4.1. The free binding energy between host and ligand found in the literature and at the primary and secondary paths. All ΔG have the unit kcal mol^{-1}. The free energies of $\Delta G_{\text{primary}}$ and $\Delta G_{\text{secondary}}$ are found by Equation (1.14) All ΔG have the unit kcal mol^{-1}. [a]Lewis and Hansen (1973), [b]Gómez-Orellana and Hallén (1993).

Ligand	**pH**	$\mathbf{\Delta G_{literature}}$	$\mathbf{\Delta G_{primary}}$	$\mathbf{\Delta G_{secondary}}$
Phenol	7	-4.6 ± 1.4^{a}	-2.08 ± 1.56	-2.55 ± 1.01
Benzene	7	-2.77 ± 0.19^{b}	-2.42 ± 1.25	-3.68 ± 1.11

Table 4.1 shows ΔG found in the literature and the primary and secondary path of phenol and benzene. The free energy between β-CD and benzene shows good compliance with the literature in the primary direction, while the secondary free energy is within the scope if the standard deviations are taken in to consideration. The free binding energies of phenol are off by about a third. The found binding energies of benzene are lower than those of phenol, in contrast to what is found by the literature (Lewis and Hansen, 1973; Gómez-Orellana and Hallén, 1993).

4.3 Simulating aspirin and β-CD in inclusion complex

Aspirin has previously been shown to form a good inclusion complex with β-CD both experimentally and through MD simulations (Fukahori et al., 2006; You et al., 2019). There are some differences in the approach developed by You et al. (2019) and this project which will be covered first. Second the results will be evaluated and compared with the previous findings.

4.3.1 Comparing parameters

You et al. (2019) chose three specific conformations of β-CD, previously shown to have a high population i simulations. In opposition, this project let both the β-CD and ligand change their confirmations freely. The only external force applied (in both works) was Equation (1.23). It is important to note that the only purpose of Equation (1.23) was to keep a certain distance between the ligand and the host, and not steer the molecules along a path. Accordingly, both the host and ligand were free to move about in relation to each other.

In the current method, the distance per window was set to 0.25 Å in opposition to 0.1 Å. With a total distance of 26 Å in each approach, this gave 104 windows and 260 window in total, respectively. While You et al. (2019) did not report how many snapshots they sampled, the initial resolution was higher than the one used here. A reason for the difference between the free binding energies in this project and You et al. (2019) might be the number of bins per window. For each window You et al. (2019) divided their data into two bins of 0.05 Å, whereas the current project divided the data into 20 bins per window of 0.0125 Å with an average of 100 snapshots per bin. Although not reported here, it was observed that a simulation with a larger distance per window resulted in a PES with a similar shape but ultimately further off the experimentally found free binding energies.

If the method, described in this project, is to be applied for further research, it might be prudent to investigate how a decreasing distance between each window with a set number of snapshots and an increasing number of bins per windows correlates. The aim of this investigation should be to optimise the wall-time used on simulations as the equilibration and production runs are by far the more expensive/time-consuming part of the analysis.

4.3.2 Inclusion complex between aspirin and β-CD

Figure 4.5 shows the free energy of the inclusion complex between aspirin and β-CD together with the three PMF's of You et al. (2019): Conf 1, Conf 2, and Conf 3. The graphs show much of the same features as phenol and benzene (Figures 4.3b and 4.4) so here the main focus will be on similarities and differences between the two data sets. Please note that the carboxylic acid group of aspirin in this project was not deprotonated and the pH of the water was 7. You et al. (2019) did not report a specific pH, although they did compare their results with data from Fukahori et al. (2006), whom reported a pH of 1.75 (Table 4.2).

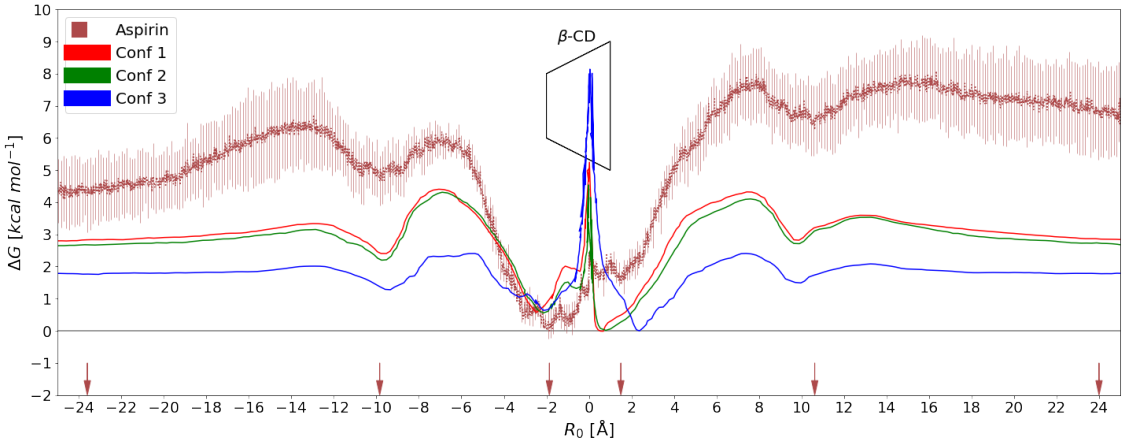


Figure 4.5. Plot showing the free energy between β-CD and aspirin. The red, green and blue graphs are taken from You et al. (2019) and show their results of US-sampling with aspirin and β-CD. The β-CD shape shows the direction of the ligand's placement at the start.

While there are some differences between the PES of aspirin and Conf 1, Conf 2, and Conf 3, the minima and peaks are positioned in the same ranges of R_0 (Figure 4.5). The lowest absolute free energy is found at the primary direction of β-CD, while the highest energy

barrier is found in the secondary direction (Table 4.2). The opposite is observed for the data of You et al. (2019).

Table 4.2. The free binding energy between host and ligand found in the literature and at the primary and secondary paths. All ΔG have the unit $\mathrm{kcal\,mol^{-1}}$. *=calculated from K_a and the relationship $K_{eq} = \frac{K_a}{K_d}$ and Equation (1.17). **=Assumed, as nothing else was reported. [a]Fukahori et al. (2006), [b]You et al. (2019) *in silico*.

Ligand	**pH**	$\mathbf{\Delta G_{literature}}$	$\mathbf{\Delta G_{primary}}$	$\mathbf{\Delta G_{secondary}}$
Aspirin	1.75	-3.74 ± 0.41[a]*	/	/
Aspirin	6	-2.33[a]*	/	/
Aspirin	7	/	-4.18 ± 1.09	-5.12 ± 1.36
	pH	$\mathbf{\Delta G_{Conf\ 1}}$	$\mathbf{\Delta G_{Conf\ 2}}$	$\mathbf{\Delta G_{Conf\ 3}}$
Aspirin	7**	-2.8[b]	-2.7[b]	-1.8[b]

Figure 4.6 shows the inclusion complex between β-CD and aspirin at different R_0 corresponding to the different minima for the primary pathway (Figure 4.5). At $R_0 = -24$ there seem to be no interaction between the two molecules. This, however, increases as the distance becomes smaller.

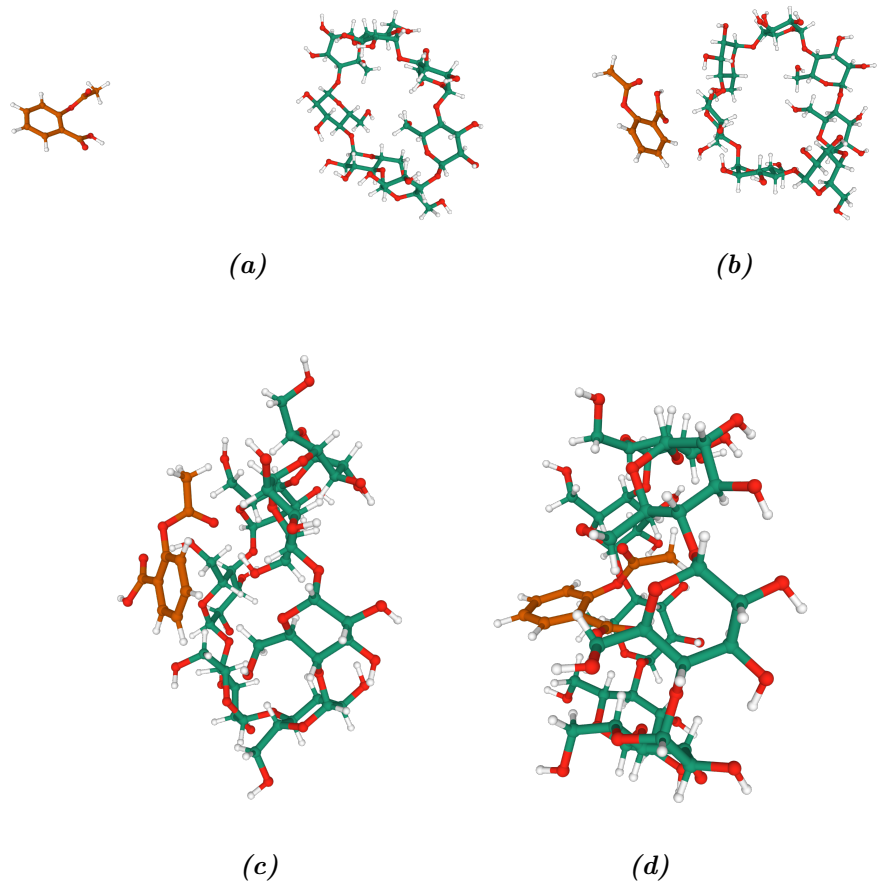


Figure 4.6. Snapshots of the inclusion complex between aspirin and β-CD at ***(a)*** $R_0 = -24$, ***(b)*** $R_0 = -10$, ***(c)*** $R_0 = -5$, and ***(d)*** $R_0 = -2$. The red and white bonds represent oxygen and hydrogen. The green and brown bonds represent the carbon bonds in the host and ligand, respectively.

Please note how the shape of the β-CD seem less distorted as R_0 is decreased. At $R_0 = -24\,\text{Å}$ and $R_0 = -10\,\text{Å}$ (Figures 4.6a and 4.6b) there is at least one glucose ring which is turned around. Although difficult to see, the glucose rings become more evenly distributed at $R_0 = -5\,\text{Å}$ and at $R_0 = -2\,\text{Å}$ the cone-like shape which is expected of a β-CD in complex, is visible. The decrease in distortion as a function of a smaller distance, is substantiated the fact that this is where the global minimum was found and with the expectation of a more stable β-CD when is forms an inclusion complex with a ligand in water (Larsen et al., 2005; Sabadini et al., 2006).

4.4 Comparison of similar compounds

Figure 4.7 shows the free binding energy between toluene, chlorobenzene, and 1,3−dichlorobenzene and β-cyclodextrin. The three compounds show similar behaviour along the secondary pathway whereas the primary pathway is more differentiated. Toluene has the lowest free binding of the three compounds shown here. As toluene is highly alphatic this in unexpected, especially because the free energy barrier is relatively high, even higher than that of benzene (Figure 4.4). This suggests that β-CD has a relative short reach in relation to toluene. Chlorobenzene and 1,3−dichlorobenzene each have a PES which are somewhere in between those of the inclusion complexes between β-CD and phenol, β-CD and benzene (Figures 4.3b and 4.4), as well as β-CD and aspirin (Figure 4.5).

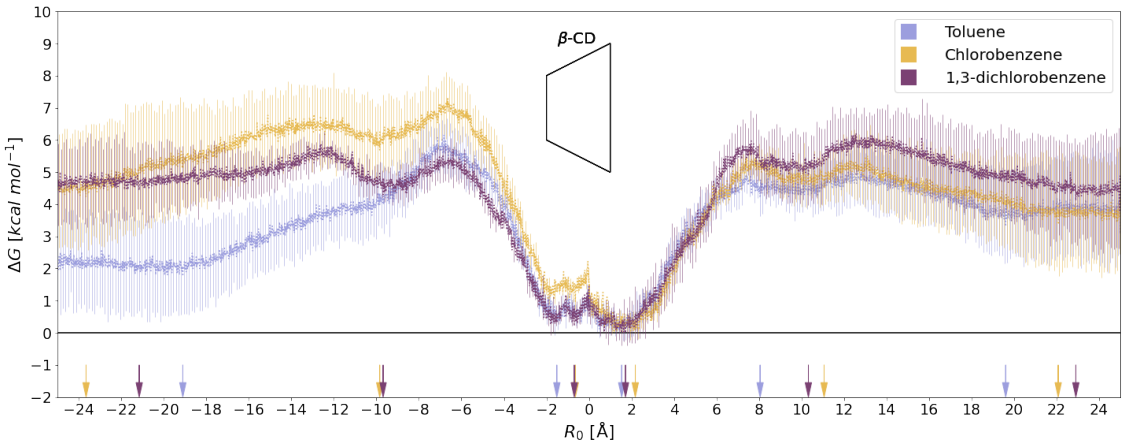


Figure 4.7. Toluene, chlorobenzene and 1,3-dichlorobenzene in complex with β-CD. The arrows correspond to the different energy minima found. Note: No second minima could be found for toluene in the primary pathway (no arrow).

According to Table 4.3 toluene, chlorobenzene, and 1,3−dichlorobenzene have the best binding affinity in the secondary direction of β-CD. Toluene has by far the highest free binding energy in the primary direction. Even so, in order to reach the solution, it needs to overcome an energy barrier equal to those of chlorobenzene and 1,3−dichlorobenzene at $R_0 \approx -7\,\text{Å}$. From this perspective it seems unlikely that toluene will make an inclusion complex with β-CD at the primary direction, whereas it seems more likely to happen with the two other compounds.

Table 4.3. The free binding energy between host and ligand found in the literature and at the primary and secondary paths. All ΔG have the unit $\mathrm{kcal\,mol^{-1}}$. *=calculated from K_a and the relationship $K_{eq} = \frac{K_a}{K_d}$ and Equation (1.17). [a]Sanemasa and Akamine (1987), [b]Takuma et al. (1990).

Ligand	**pH**	$\mathbf{\Delta G_{literature}}$	$\mathbf{\Delta G_{primary}}$	$\mathbf{\Delta G_{secondary}}$
Toluene	7	-2.93 ± 1.36^{a}*	-1.52 ± 1.5	-3.47 ± 1.46
Chlorobenzene	7	-3.01 ± 1.36^{b}*	-3.13 ± 1.83	-3.44 ± 1.74
1,3-dichlorobenzene	7	/	-4.11 ± 1.2	-4.18 ± 1.67

4.5 A critical perspective

The method in this project did not control how the host and ligand faced each other. This means that both molecules had the possibility of rotating around their own axes as they were pushed by the water molecules and their inter acting forces. As all ligands were relatively small, their free rotation would probably not influence the free energy much on its own. However, by analysing the free energies it was shown that the direction of the β-CD in relation to the ligand did matter in this regard. The distance was only measured from CoM to CoM of the two molecules without regard to the intended direction of the β-CD.

Conclusion 5

It has been shown that it is possible to use OpenMM and OpenFF to simulate inclusion complexes between a ligand and host, in this case β-CD. The relative new force field Sage (`openff-2.0.0.offxml`) was used to parameterise the ligands through direct chemical perception, while the parameters for β-CD were taken from GLYCAM06. Through umbrella sampling, a set of simulations were carried out ranging from $-26\,\text{Å}$ to $26\,\text{Å}$ with an increment of $0.25\,\text{Å}$ where free energy of the inclusion complexes were found. For each window in the umbrella sampling, the distance between the ligand and β-Cd were controlled by an external biasing energy potential. In each snapshot, the distance was measured and analysed by FastMBAR. The resulting potential energy surface showed good compliance with other simulation works from the literature, even though the free energies were a bit higher. This might be explained by the larger distance increment used in this project and at the same time a smaller number of initial windows in the umbrella sampling. The described simulation method applying the biasing potential on inclusion complexes between β-CD and each of the following ligands: benzene, toluene and chlorobenzene showed good correpondance with results from laboratory experiments with similar parameters.

Bibliography

Chiara Borsari, Darci J. Trader, Annalisa Tait, and Maria P. Costi. Designing chimeric molecules for drug discovery by leveraging chemical biology. Journal of Medicinal Chemistry, 63(5):1908–1928, February 2020. doi: 10.1021/acs.jmedchem.9b01456. URL https://doi.org/10.1021/acs.jmedchem.9b01456.

Octav Caldararu, Martin A. Olsson, Majda Misini Ignjatović, Meiting Wang, and Ulf Ryde. Binding free energies in the SAMPL6 octa-acid host–guest challenge calculated with MM and QM methods. Journal of Computer-Aided Molecular Design, 32(10): 1027–1046, September 2018. doi: 10.1007/s10822-018-0158-2. URL https://doi.org/10.1007/s10822-018-0158-2.

D.A. Case, V. Babin, J.T. Berryman, R.M. Betz, Q. Cai, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, H. Gohlke, A.W. Goetz, S. Gusarov, N. Homeyer, P. Janowski, J. Kaus, I. Kolossváry, A. Kovalenko, T.S. Lee, S. LeGrand, T. Luchko, R. Luo, B. Madej, K.M. Merz, F. Paesani, D.R. Roe, A. Roitberg, C. Sagui, R. Salomon-Ferrer, G. Seabra, C.L. Simmerling, W. Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu, and P.A. Kollman. Amber 14, 2014. University of California, San Francisco.

Mark E. Davis and Marcus E. Brewster. Cyclodextrin-based pharmaceutics: past, present and future. Nature Reviews Drug Discovery, 3(12):1023–1035, December 2004. doi: 10.1038/nrd1576. URL https://doi.org/10.1038/nrd1576.

Xinqiang Ding, Jonah Z. Vilseck, Ryan L. Hayes, and Charles L. Brooks. Gibbs sampler-based λ-dynamics and rao-blackwell estimator for alchemical free energy calculation. Journal of Chemical Theory and Computation, 13(6):2501–2510, May 2017. doi: 10.1021/acs.jctc.7b00204. URL https://doi.org/10.1021/acs.jctc.7b00204.

Xinqiang Ding, Jonah Z. Vilseck, and Charles L. Brooks. Fast solver for large scale multistate bennett acceptance ratio equations. Journal of Chemical Theory and Computation, 15(2):799–802, January 2019a. doi: 10.1021/acs.jctc.8b01010. URL https://doi.org/10.1021/acs.jctc.8b01010.

Xinqiang Ding, Jonah Z. Vilseck, and Charles L. Brooks III. Fastmbar documentation, 2019b. URL https://fastmbar.readthedocs.io/en/latest/usage.html.

Takanori Fukahori, Minako Kondo, and Sadakatsu Nishikawa. Dynamic study of interaction between β-cyclodextrin and aspirin by the ultrasonic relaxation method. The Journal of Physical Chemistry B, 110(9):4487–4491, February 2006. doi: 10.1021/jp058205n. URL https://doi.org/10.1021/jp058205n.

Isabel Gómez-Orellana and Dan Hallén. The thermodynamics of the binding of benzene to β-cyclodextrin in aqueous solution. Thermochimica Acta, 221(2):183–193, July 1993. doi: 10.1016/0040-6031(93)85062-e. URL https://doi.org/10.1016/0040-6031(93)85062-e.

GROMACS development team. Gromacs documentation release 2019-rc1. https://manual.gromacs.org/documentation/2019-rc1/manual-2019-rc1.pdf, December 2018. Accessed: 2022-04-05.

Frank Jensen. Introduction to computational chemistry. John Wiley & Sons, Nashville, TN, 3 edition, February 2017.

Jan H Jensen. Molecular Modeling Basics. CRC Press, Boca Raton, FL, February 2009.

Karl N. Kirschner, Austin B. Yongye, Sarah M. Tschampel, Jorge González-Outeiriño, Charlisa R. Daniels, B. Lachele Foley, and Robert J. Woods. GLYCAM06: A generalizable biomolecular force field. carbohydrates. Journal of Computational Chemistry, 29(4):622–655, September 2007. doi: 10.1002/jcc.20820. URL https://doi.org/10.1002/jcc.20820.

Gerhard König. Methodological Studies concerning Free Energy Simulations. PhD dissertation, University of Vienna, 2010.

David Lane. 4.7: Variance sum law ii - correlated variables, Apr 2022. URL https://stats.libretexts.org/Bookshelves/Introductory_Statistics/Book%3A_Introductory_Statistics_(Lane)/04%3A_Describing_Bivariate_Data/4.07%3A_Variance_Sum_Law_II_-_Correlated_Variables.

Kim L. Larsen, Finn L. Aachmann, Reinhard Wimmer, Valentino J. Stella, and Ulrich Madsen Kjølner. Phase solubility and structure of the inclusion complexes of prednisolone and 6α-methyl prednisolone with various cyclodextrins. Journal of Pharmaceutical Sciences, 94(3):507–515, March 2005. doi: 10.1002/jps.20192. URL https://doi.org/10.1002/jps.20192.

Kim Lambertsen Larsen. Large cyclodextrins. Journal of Inclusion Phenomena and Macrocyclic Chemistry, 43(1/2):1–13, 2002. doi: 10.1023/a:1020494503684. URL https://doi.org/10.1023/a:1020494503684.

Edwin A. Lewis and Lee D. Hansen. Thermodynamics of binding of guest molecules to α- and β-cyclodextrins. J. Chem. Soc., Perkin Trans. 2, pages 2081–2085, 1973. doi: 10.1039/p29730002081. URL https://doi.org/10.1039/p29730002081.

Antonia S.J.S. Mey, Bryce K. Allen, Hannah E. Bruce Macdonald, John D. Chodera, David F. Hahn, Maximilian Kuhn, Julien Michel, David L. Mobley, Levi N. Naden, Samarjeet Prasad, Andrea Rizzi, Jenke Scheen, Michael R. Shirts, Gary Tresadern, and Haufeng Xu. Best practices for alchemical free energy calculations [article v1.0]. Living Journal of Computational Molecular Science, 2(1), 2020. doi: 10.33011/livecoms.2.1.18378. URL https://doi.org/10.33011/livecoms.2.1.18378.

David L. Mobley, Caitlin C. Bannan, Andrea Rizzi, Christopher I. Bayly, John D. Chodera, Victoria T. Lim, Nathan M. Lim, Kyle A. Beauchamp, David R. Slochower, Michael R. Shirts, Michael K. Gilson, and Peter K. Eastman. Escaping atom types in force fields using direct chemical perception. Journal of Chemical Theory and Computation, 14(11):6076–6092, October 2018. doi: 10.1021/acs.jctc.8b00640. URL https://doi.org/10.1021/acs.jctc.8b00640.

Open Force Field Consortium. Openff toolkit documentation release 0.10.2+9.gc29c9359.dirty. https://open-forcefield-toolkit.readthedocs.io/en/0.10.2/, January 2022. Accessed: 2022-02-02.

Edvaldo Sabadini, Terence Cosgrove, and Fernanda do Carmo Egídio. Solubility of cyclomaltooligosaccharides (cyclodextrins) in h2o and d2o: a comparative study. Carbohydrate Research, 341(2):270–274, February 2006. doi: 10.1016/j.carres.2005.11.004. URL https://doi.org/10.1016/j.carres.2005.11.004.

Isao Sanemasa and Youko Akamine. Association of benzene and alkylbenzenes with cyclodextrins in aqueous medium. Bulletin of the Chemical Society of Japan, 60(6): 2059–2066, June 1987. doi: 10.1246/bcsj.60.2059. URL https://doi.org/10.1246/bcsj.60.2059.

Michael R. Shirts and John D. Chodera. Statistically optimal analysis of samples from multiple equilibrium states. The Journal of Chemical Physics, 129(12):124105, September 2008. doi: 10.1063/1.2978177. URL https://doi.org/10.1063/1.2978177.

Michael R. Shirts, Christoph Klein, Jason M. Swails, Jian Yin, Michael K. Gilson, David L. Mobley, David A. Case, and Ellen D. Zhong. Lessons learned from comparing molecular dynamics engines on the SAMPL5 dataset. September 2016. doi: 10.1101/077248. URL https://doi.org/10.1101/077248.

Casper Steinmann and Jan H. Jensen. Using a genetic algorithm to find molecules with good docking scores. January 2021. doi: 10.26434/chemrxiv.13525589.v2. URL https://doi.org/10.26434/chemrxiv.13525589.v2.

József Szejtli. Introduction and general overview of cyclodextrin chemistry. Chemical Reviews, 98(5):1743–1754, June 1998. doi: 10.1021/cr970022c. URL https://doi.org/10.1021/cr970022c.

Tatsuyoshi Takuma, Toshio Deguchi, and Isao Sanemasa. Association of halobenzenes with cyclodextrins in aqueous medium. Bulletin of the Chemical Society of Japan, 63 (4):1246–1248, 1990. ISSN 0009-2673.

Ignacio Tinoco, Kenneth Sauer, James C Wang, and Joseph D Puglisi. Physical chemistry. Pearson, Upper Saddle River, NJ, 4 edition, December 1999.

Kaneto Uekama, Fumitoshi Hirayama, and Tetsumi Irie. Cyclodextrin drug carrier systems. Chemical Reviews, 98(5):2045–2076, June 1998. doi: 10.1021/cr970025p. URL https://doi.org/10.1021/cr970025p.

Junmei Wang, Wei Wang, Peter A. Kollman, and David A. Case. Automatic atom type and bond type perception in molecular mechanical calculations. Journal of Molecular Graphics and Modelling, 25(2):247–260, October 2006. doi: 10.1016/j.jmgm.2005.12.005. URL https://doi.org/10.1016/j.jmgm.2005.12.005.

Wanli You, Zhiye Tang, and Chia en A. Chang. Potential mean force from umbrella sampling simulations: What can we learn and what is missed? Journal of Chemical Theory and Computation, 15(4):2433–2443, February 2019. doi: 10.1021/acs.jctc.8b01142. URL https://doi.org/10.1021/acs.jctc.8b01142.

Zhijun Zhang, Xinzijian Liu, Kangyu Yan, Mark E. Tuckerman, and Jian Liu. Unified efficient thermostat scheme for the canonical ensemble with holonomic or isokinetic constraints via molecular dynamics. The Journal of Physical Chemistry A, 123(28): 6056–6079, May 2019. doi: 10.1021/acs.jpca.9b02771. URL https://doi.org/10.1021/acs.jpca.9b02771.