Extending the scope of alchemical perturbation methods for ligand binding free energy calculations

Eric Fagerberg
Supervisor: Pär Söderhjelm

2016

Contents

1 Abstract

Previously, a method for computing binding free energies between different poses of a ligand bound to a protein using alchemical perturbation was developed. The methodology is to perturb the ligand into a smaller version, common to both poses, from which the difference in free energy between poses can be computed. Here, the method is further improved by finding low-error setups for the method, by investigating different kinds of restraints put on the system during simulation and different kinds of parameters for the soft-core potential. The best low-error setup found was using a 1-1-48 soft-core potential with a water barrier and positional restraints for all the non-hydrogen atoms in the system. Instability was detected for one of the poses, this was investigated. The key to having a stable pose seems to be to understand the effect on stability of changing the Ryckaert-Bellemans parameters for a single rotatable bond.

2 Introduction

Computational modelling can provide chemical properties of a system. A typical use is the binding of a small molecule, called ligand, to a protein. This is relevant to the pharmaceutical industry, where the ligand would be a drug which would modulate the function of the protein upon binding, often by inhibiting the binding of the natural ligand, thereby treating a disease. When a lead for a drug has been found, the goal is to improve this lead. In this process synthesis of new molecules is required but this can be slow and expensive. Using computational methods could help finding possible improvements which could reduce the need for synthesis, accelerating the process [1]. The main property of interest in this case is the free energy of binding, which gives a number on how successful a potential drug could be, not considering body absorption of the drug, side effects or similar aspects. There are different ways of performing this modelling, each approach having different drawbacks and advantages [2]:

- Docking. This method uses a known protein structure and a known ligand structure to rapidly find the optimal conformation for binding, usually treating the protein as rigid. Many conformations and binding sites for ligand are considered, each pose being ranked by a scoring function. It trades physical accuracy for speed, being able to model one compound in seconds but can not very often do more than separate non-binding ligands from binding ligands. In 2008, about 88% [2] of publications related to computational methods were about docking.

- Molecular mechanics with Poisson-Boltzman+surface area, MM-PBSA, uses more extensive conformational sampling than docking, being able to estimate the free energy of binding but is limited in estimating the entropic contribution to the free energy. The increase in physical accuracy increases the computational cost to several hours per compound. 1% [2] of publications related to computational methods in 2008 considered MM-PBSA.

- Relative Binding Free Energies, RBFE, uses molecular dynamics for extensive conformational sampling to compute differences in binding free energy between similar ligands by alchemically transforming one ligand to another. This procedure can cost hundreds of CPU-days, but generally have greater accuracy than MM-PBSA or docking methods. This method requires one bound structure of a similar ligand as a starting point. 11% [2] of publications related to computational methods in 2008 considered RBFE.

- Absolute Bindning Free Energies, ABFE, also molecular dynamics but also involve separate sets of simulations for the solvated ligand and solvated complex. The process is to alchemically transform the full system into “nothing”, in a sense similar to RBFE but instead of comparing two ligands, one ligand and “nothing” is compared. It is called ‘absolute’ since it requires no prior knowledge about binding affinity. In principle, it is the most powerful approach, fairly accurate but only 0.04% [2] of computational methods related publications concerned ABFE in 2008.

Using the RBFE approach, a method to compute the free energy difference between ligands which have different conformational poses has previously been developed by . As a model, the HIV-1 Protease enzyme was used with (3S,4S)-3,4-bis-[(4-carbamoyl-benzene-sulfonyl)-(3-methyl-but-2-enyl)-amino]-pyrrolidine as ligand, which has been shown by Blum et. al. to have two different binding modes(found from crystal structures) denoted orthorhombic(“ortho”) and hexagonal(“hexa”) form, PDB IDs 3CKT and 2ZGA respectively [3]. These are shown in figure 1

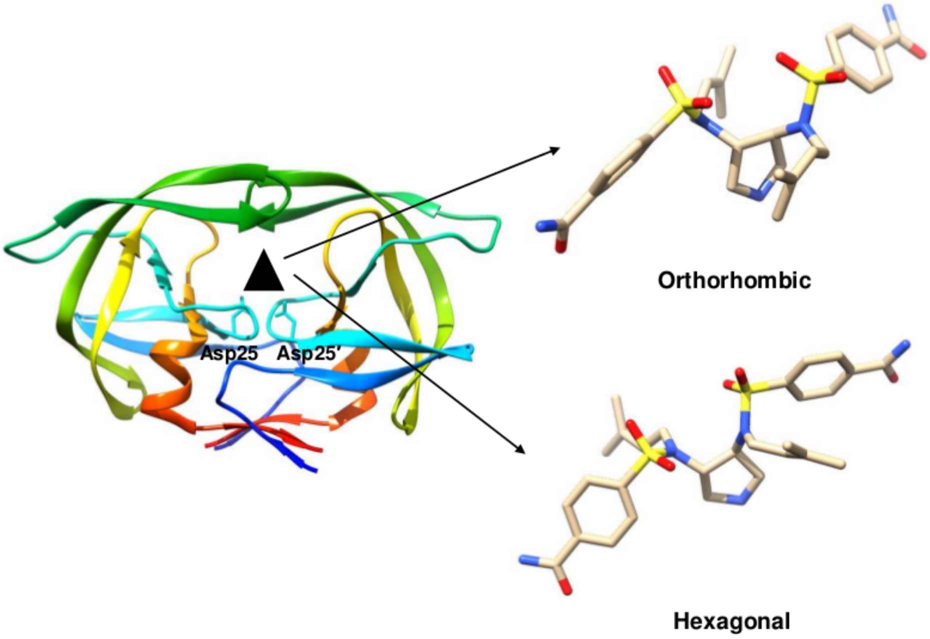

Figure 1: HIV-1 Protease enzyme and the two poses of the ligand.

2.1 Aim

The aim of this project is to improve a method formulated by Teodor Rodin [6]. Prime concern is lowering variance, improving error estimation and understanding why the problems appearing in previous work arise.

3 Theory

3.1 Statistical thermodynamics

A single molecule can be in different states of energy. When considering many molecules together, there is only one way for all molecules to be in the same state but there are several ways on how to distribute different states of energy for a number of molecules. The first molecule could occupy state ϵ_1 and a second molecule occupy state ϵ_2, but there could also be that the first molecule occupy state ϵ_2 and the second molecule occupy state ϵ_1. These two cases would be indistinguishable if the molecules are indistinguishable. Therefore, there is a greater possibility of having such a configuration rather than a configuration of both molecules being in state ϵ_1, because there are more ways to achieve the former. With this reasoning and considering a closed system with constraints that there is a constant total energy and a constant total number of molecules, one can find that the most probable configuration depends on the **Boltzmann distribution**, eq. 1:

$$\frac{N_i}{N} = \frac{e^{-\beta\epsilon_i}}{\sum_i e^{-\beta\epsilon_i}} \tag{1}$$

where N$_i$ number of molecules have an energy of ϵ_i, β is 1/kT, where T is the temperature and k a constant, the Boltzmann constant. The right hand denominator is designated q and called the molecular partition function. The total energy can be written in terms of this function, see equation 2:

$$E(T) = \sum_i N_i\epsilon_i = \frac{N}{q}\sum_i \epsilon_i e^{-\beta\epsilon_i} = -\frac{N}{q}\frac{d}{d\beta}\sum_i e^{-\beta\epsilon_i} = -\frac{N}{q}\frac{dq}{d\beta} \tag{2}$$

Recognising that the derivative is actually a partial derivative and that E(T) is the value of the internal energy relative to the T=0 level, the internal energy is acquired(eq. 3):

$$U(T) = U(0) + E(T) = U(0) - \frac{N}{q}\Big(\frac{\partial q}{\partial\beta}\Big)_V = U(0) - N\Big(\frac{\partial \ln q}{\partial\beta}\Big)_V \tag{3}$$

The partition function can also describe entropy. Using a similar argument as above with the Boltzmann formula(not presented here), it can be shown that entropy is related to the partition function as presented in eq. 4:

$$S(T) = \frac{U(T) - U(0)}{T} + Nk\ln(q) \tag{4}$$

3.1.1 Canonical ensemble

A canonical ensemble is a closed system with some specific volume, composition and temperature that has been replicated $\tilde{N}$ times. All these replicas are in thermal contact so they have the same temperature. The number $\tilde{N}$ can be as large as one would like. As with single systems with molecules, the members of an ensemble may have different energies but some are more probable than others. With the same argument as for single closed systems with molecules, the most probable configuration at constant energy $\tilde{E}$ and fixed total number of members $\tilde{N}$ is as in equation 5, where $\tilde{N}_i$ ensembles have energy E_i:

$$\frac{\tilde{N}_i}{\tilde{N}} = \frac{e^{-\beta E_i}}{Q} \quad \text{where} \quad Q = \sum_i e^{-\beta E_i} \tag{5}$$

Q is the canonical partition function. The difference between q and Q is that Q does not assume independent molecules, which allows for discussion about molecular interactions. The relation betwen Q and q for the case of an ideal gas is as below(eq. 6 and 7):

$$Q = q^N \text{for distinguishable independent molecules} \tag{6}$$

$$Q = \frac{q^N}{N!} \text{for indistinguishable independent molecules} \tag{7}$$

The expressions for internal energy and entropy(equations 3 and 4) becomes a little different using a canonical ensemble, see equations 8 and 9(note that these are not only for the special case of an ideal gas, but are more general):

$$U(T) = U(0) - \Big(\frac{\partial \ln Q}{\partial \beta}\Big)_V \tag{8}$$

$$S(T) = \frac{U(T) - U(0)}{T} + k\ln(Q) \tag{9}$$

3.1.2 Helmholtz and Gibbs energies

The Helmholtz energy is described with internal energy and entropy according to A = U - TS. Using equations 8 and 9, an expression for the Helmholtz energy in terms of the canonical partition function is arrived at in equation 10:

$$A - A(0) = -kT\ln Q \tag{10}$$

Pressure and the Helmholtz energy can be shown to be related by $p = -(\partial A/\partial V)_T$, which after insertion of equation 10 leads to equation 11:

$$p = kT\Big(\frac{\partial \ln Q}{\partial V}\Big)_T \tag{11}$$

Finally, the Gibbs energy is related to the Helmholtz energy by $G = A + pV$. From equations 10 and 11 is the Gibbs energy derived as equation 12:

$$G - G(0) = -kT\ln Q + kTV\Big(\frac{\partial \ln Q}{\partial V}\Big)_T \tag{12}$$

Thus, the Gibbs energy can be calculated from the canonical partition function.

3.2 Molecular Dynamics

The purpose of a simulation is to sample the canonical ensemble. One way to simulate is through Molecular Dynamics(MD). MD simulations uses Newton's laws of motion to describe the motions of atoms from a starting configuration. To define a state in the system, three space coordinates and three coordinates for the momentum(x, y, z) for each atom, giving a total of $6N$ variables for a state with N atoms. A 6N-dimensional *phase space* is thus defined. The Newtonian equations are solved for small timesteps, with the coordinates written out to a file for some specified interval. All these time-dependent coordinates make up the trajectory of the system, which will be the samples of the ensemble. The partition function would be exact if all points in phase space would be visited(such a trajectory is called *ergodic*), but since this is not achievable, only estimates can be achieved. The starting configuration may affect the end result to some extent, which would not be the case if the trajectory would be ergodic.

MD simulations uses the Born-Oppenheimer approximation; electron movements are not considered but molecules are regarded as always being in their electronic groundstate. This avoids quantum mechanical calculations, decreasing the computational load significantly, but phenomena depending on electron distribution cannot be computed, such as chemical reactions.

For the calculation of forces, an empirical model referred to as “force fields” are used. Force fields can have different functional form, but an example is provided in equation 13:

$$\begin{aligned}\mathscr{V}(\mathbf{r}^N) = \sum_{bonds} \frac{k_{bonds,i}}{2}(l_i - l_{i,0})^2 + \sum_{angles} \frac{k_{angles,i}}{2}(\theta_i - \theta_{i,0})^2 + \\ \sum_{torsions} \frac{V_n}{2}(1 + cos(n\omega - \gamma)) + \\ \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left(4\epsilon_{ij} \left[(\frac{\sigma_{ij}}{r_{ij}})^{12} - (\frac{\sigma_{ij}}{r_{ij}})^{6} \right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right)\end{aligned} \tag{13}$$

$\mathscr{V}$ is the potential energy, depending on the position($\mathbf{r}$) of all the N atoms in the simulations. In this example, the first term consider bond stretching, with an elementary approach of using Hooke's law, where the energy is related to the difference in bond length l_i to a reference bond lenght $l_{i,0}$. The second term is the angle bending, again exemplified with Hooke's law. The third term are the torsional interactions. The last term are non-bonded interactions like Lennard-Jones potential(where ϵ is the well depth, σ the collision diameter and r the distance between any pair of atoms) for the van der Waals forces and Coulomb potential(q_x is the charge for atom x, ϵ_0 the permittivity in vacuum and r the distance between any pair of atoms) for electrostatics. Other force fields may have more terms, such as explicit terms for hydrogen bonding or have more complicated expressions for the same contribution. Force fields may have different functional form, but they may also have different values of the parameters, i.e. V$_n$, k$_i$ and σ_{ij}, for optimising for different classes of molecules. The force field is a possible source of error, as it omits polarisation, higher-order multipoles, charge transfer and charge penetration [23]

In a MD-simulation, sometimes it is necessary to keep some atoms fixed in a position, to avoid drastic changes for critical parts of the system of study. This is achieved by introducing an energy penalty for when an atom moves away from a reference position. This energy penalty, or restraint, can be manually set to any prefered value.

One way of expressing the potential governing the torsions is by the Ryckaert-Bellemans function, see equation 14.

$$V_{rb}(\phi_{ijkl}) = \sum_{0}^{5} C_n(cos(\phi - 180))^n \tag{14}$$

The C_n are the parameters, which are defined when creating the force field, typical examples have values in the range -50 to 50 $kJmol^{-1}$, but they can of course be any value. The angle $\phi - 180$ is defined by $(\phi - 180)_{trans} = 0$, the "polymer conventions".

MD-simulations create configurations connected in time. This allows for calculation of time-dependent properties, an advantage MD-simulations has over other simulation methods like the Monte Carlo method. However, even for time-independent properties, the *autocorrelation time* is an important factor because an assumption for calculating the free energy is that samples are uncorrelated. To understand this, consider the normalised correlation function C_{xy} between two variables x and y in equation 15:

$$C_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{15}$$

The normalised correlation function give values between -1 and +1, where -1 indicate negative correlation, +1 positive correlation and 0 no correlation. One of the variables can be set to only consider a specific reference time, giving the *time correlation coefficients* which correlates variables at different time points. If the two variables would also be set to be the same, the autocorrelation function is acquired. If there would be no time difference between the variables, the function assumes a value of 1, while if the time difference is large the function assumes a value of 0. The autocorrelation time is the characteristic decay time of the autocorrelation function. Quantities with longer autocorrelation time than the length of the simulation cannot be determined accurately. In the same way, short autocorrelation times in comparison with simulation time give greater accuracy, as the number of data sets with complete relaxation increases. If simulation time is Q steps and autocorrelation time is P steps, then there would be (Q/P) data sets to be used for the determination of a property, with greater precision achieved with more data sets(as in all statistics). Short autocorrelation times can thus in principle reduce the error of a simulation, but cannot be affected unless artificial modifications are made.

3.3 Free Energy Perturbation

From equation 10, the difference in free energy between two states X and Y can be calculated by subtracting their energies. Using the partition function (definition 5) and after some algebraic exercise, the difference can be written as ensemble averages, see equation 16(going from X to Y).

$$\Delta A = -kT \ln \left\langle e^{-(E_y - E_x)} \right\rangle_x \tag{16}$$

An equivalent expression but possibly with different statistical error is by averaging over the y state as in 17.

$$\Delta A = kT \ln \left\langle e^{-(E_x - E_y)} \right\rangle_y \tag{17}$$

This way of calculating the free energy is called "exponential averaging" or the "Zwanzig relationship". However, it is not optimal in terms of statistical performance, more modern methods like Bennett Acceptance Ratio(BAR) [4] or Multistate Bennett Acceptance Ratio(MBAR) are preferred. The exponential averaging is however easier to understand, so it is included here to give an appreciation of how the free energy is calculated. Loosely speaking, the BAR method can use data from one of the ensembles(ex. the x-ensemble through equation 16) more than the other(the y-ensemble through equation 17) if the statistical properties are better for either ensemble. Thus, the BAR method is weighting data points from the two states in an error-minimizing way. This makes it possible to have low error estimates even if one of the states would have a high error, if the other state has lower error. Because of this, it is possible to avoid problematic ensembles to some extent using the BAR method. The free energy difference between two states can thus be computed through the exponential average of a potential difference for the states investigated. But, if the two states have little overlap in phase space, the difference calculated will not be accurate. To resolve this, intermediate steps need to be introduced. These intermediate steps might be non-physical, and the overall transformation from one state X to another state Y is therefore called alchemical. Mathematically, this is described by equation 18:

$$\begin{gathered}\Delta A = A(Y) - A(X) = -kT \ln \frac{Q(Y)}{Q(X)} = \\ \text{introducing intermediates} \\ = -kT \ln \left[\frac{Q(Y)}{Q(1)} \frac{Q(1)}{Q(2)} \frac{Q(2)}{Q(3)} \cdots \frac{Q(N-1)}{Q(N)} \frac{Q(N)}{Q(X)} \right]\end{gathered} \tag{18}$$

How far the simulation has progressed from X to Y is described by the *coupling parameter* λ. λ assumes values between 0 and 1, where 0 and 1 are the end states. For example, an intermediate state which is 50% X and 50% Y is described by $\lambda = 0.5$. The terms in the force field are written as a linear combination of X and Y with the help of λ. Any intermediate state, using the force field in 13, could be described with equation 19:

$$V(\lambda) = \lambda V_Y + (1 - \lambda) V_X \tag{19}$$

The perturbation may be performed step-wise, by first changing electrostatics and then changing the Lennard-Jones interactions.

The error in a perturbation is connected to the standard deviation of the energy calculated, however, one more often considers the standard deviation of the derivative of the energy depending on λ instead. This has the advantage of giving information about whether a lower standard deviation can be achieved if there is a greater number of intermediates(=more dense λ) included in the simulation. This derivative, henceforth refered to as $\frac{\partial H}{\partial \lambda}$(or dhdl)[1], can be computed for each value of lambda, making it possible to appreciate which intermediate steps in the perturbation are the most error introducing, which would need to be divided into smaller intermediates.

[1] The dependence of λ is the same in $\frac{\partial H}{\partial \lambda}$ as in $\frac{\partial V}{\partial \lambda}$, so they may be used interchangeably

3.3.1 Soft-core interactions

Using FEP, atoms may become deleted or introduced in the system. When lambda is at its extremes, the interactions between such atoms could be very weak, so that other particles may come very close which would lead to large fluctuations in the energy derivative. To overcome this, the interacting term(Lennard-Jones potential) is modified so that the interaction attains a finite value as the distance between particles goes to zero. This means that the Lennard-Jones potential changes into equation 20:

$$\begin{aligned} V_{ij}^{LJ}(\lambda) = 4(1-\lambda)\epsilon_X \left(\frac{\sigma_X^{12}}{\left[\alpha\lambda^p\sigma_X^6 + r_{ij}^6\right]^2} - \frac{\sigma_X^6}{\left[\alpha\lambda^p\sigma_X^6 + r_{ij}^6\right]} \right) + \\ 4\lambda\epsilon_Y \left(\frac{\sigma_Y^{12}}{\left[\alpha(1-\lambda)^p\sigma_Y^6 + r_{ij}^6\right]^2} - \frac{\sigma_Y^6}{\left[\alpha(1-\lambda)^p\sigma_Y^6 + r_{ij}^6\right]} \right) \end{aligned} \tag{20}$$

where p is the soft-core λ power and α is the soft-core parameter. The soft-core λ power p is often chosen to be 1 or 2. An example of how this soft-core potential behaves can be found in figure 2, where the soft-core parameter is set to 1.0.

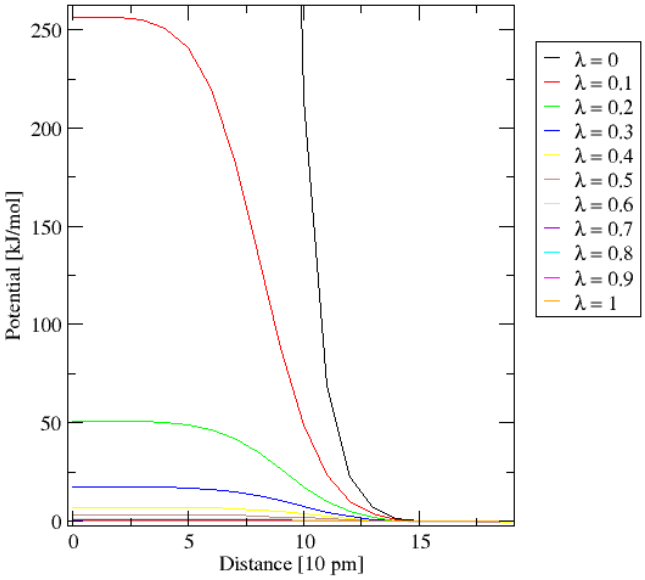

Figure 2: Soft-core potential example. The behaviour for different values of λ is shown for a soft-core parameter α of 1.0

One alternative given in [5] that should give lower and more evenly distributed statistical noise uses an exponent of 48 in the σ_Y, σ_X terms in the denominator and has a denominator bracket power of $\frac{1}{48}$. This was found by using calculus of variations, trying to minimize the variance to bring forth a minimal variance pathway for the alchemical perturbation, which gave lower variance for most values of λ when simulating an uncharged 3-methylindole in water compared to previous "best practices" of having an exponent of 6.

4 Project

The model system, HIV-1 protease binding with a pyrrolidine-based ligand has two poses, denoted “Ortho”(PDB id=3CKT) and “Hexa”(PDB id=2ZGA) with different protein-ligand interactions. These were found by cocrystallizing the protein and the ligand, but it was not expected to find two different crystal structures from this procedure, it was a serendipitous discovery [3].

The ligand can roughly be described as a pyrrolidine ring with four “arms”, see figure 3, where the numbering of the non-hydrogen atoms is also included. The different poses were shown in [3] to have the ligand “arms” in different pockets of the protein, thereby having different protein-ligand interactions. In order to find the difference in free energy, in this study both the Hexa pose and the Ortho pose is perturbed so that their “arms”(atoms in black in figure 3) disappear to the common element of the pyrrolidine ring, making the poses identical.

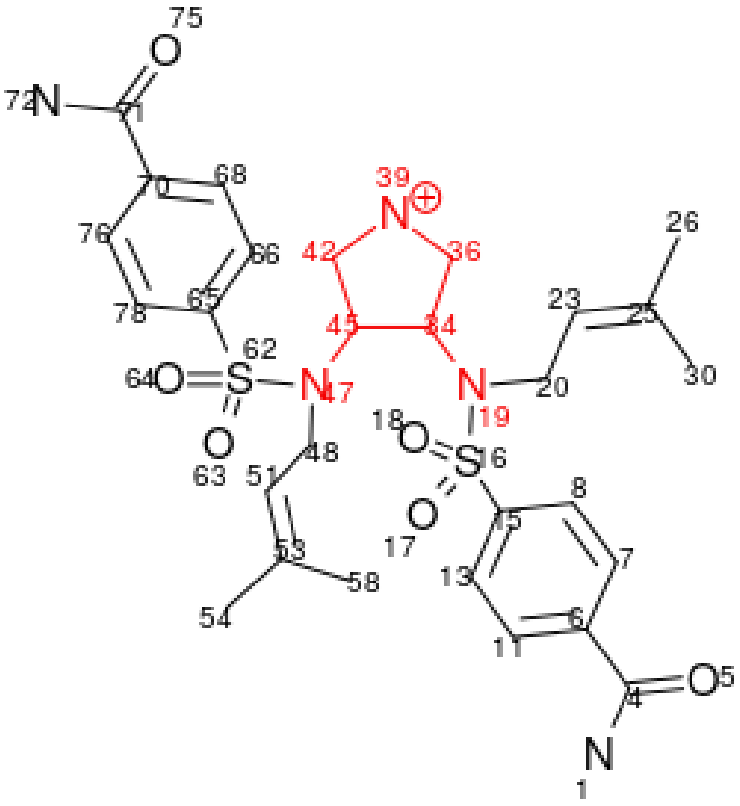

Figure 3: Ligand for the studied system. Atoms in black are perturbed.

The perturbation is performed step-wise, first turning off the electrostatics for the “arms”, then turning off the Lennard-Jones interactions, effectively making the atoms on the arms non-interacting dummy atoms. This will yield a thermodynamic cycle as shown in figure 4, where the non-direct route from Ortho to Hexa is as valid as the direct route because free energy is a state function. Looking at figure 3, the atoms coloured red are the ones that will be left after perturbation.

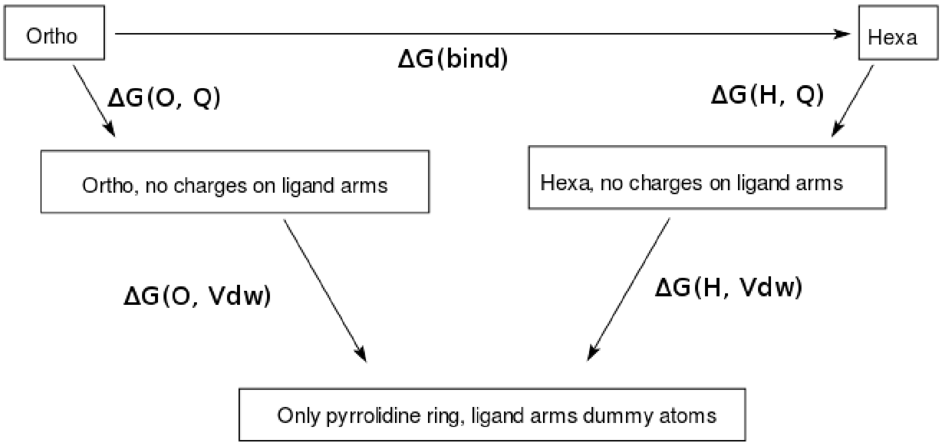
Ortho
ΔG(bind)
Hexa
ΔG(O, Q)
ΔG(H, Q)
Ortho, no charges on ligand arms
Hexa, no charges on ligand arms
ΔG(O, Vdw)
ΔG(H, Vdw)
Only pyrrolidine ring, ligand arms dummy atoms

Figure 4: Thermodynamic cycle. Any route can be used to calculate the free energy difference between Ortho and Hexa as the free energy is a state function.

From the thermodynamic cycle in figure 4, the difference in binding free energy between the Ortho pose and the Hexa pose is computed through equation 21:

$$\Delta G(bind) = \Delta G(O, Q) + \Delta G(O, \mathrm{Vdw}) - (\Delta G(H, \mathrm{Vdw}) + \Delta G(H, Q)) \tag{21}$$

4.1 Previous work

In previous unpublished work by Teodor Rodin in the Söderhjelm research group [6], initial simulations according to the suggested method were made, but there were some problems. The variance of the calculated free energy was high, particularly for the transformation where the Lennard-Jones interactions are turned off. This decreased the accuracy of the computation. The obtained results were also unexpectedly low, for example the free energy difference of the ligand in the presence of protein or in vacuum was only 10 kJ mol^{-1}. In these simulations, a restraint was introduced to keep the system in its pose, but the effect of this inclusion was never evaluated.

In this work, focus has primarily been to understand the problem, then lowering the variance and improving error estimation.

5 Methods

5.1 Simulation details

Molecular dynamics was performed using GROMACS [7], version 5.0.4, with the Plumed plugin [8], version 2.1. Perturbation simulations has been made using eleven lambdas, equally spaced in the interval 0 - 1. Simulations were performed at a temperature of 300 K, using a velocity rescaling thermostat with a stochastic term [11] and a Parrinello-Rahman barostat [12], thus generating NPT-ensembles. The step length was set to 0.002 ps, a Verlet cut-off scheme with cut-off = 1.0 nm was used. In order to achieve an equilibrated system, the first three nanoseconds are discarded before computing the standard deviation of $\frac{\partial H}{\partial \lambda}$ for each λ, as recommended by [13]. This is no guarantee of having an equilibrated system, but should bring the system close to equilibrium. For the calculation of autocorrelation times, only the first nanosecond was discarded, as autocorrelation times are not as sensitive to un-equilibrated systems. If nothing else is stated, a soft-core parameter α of 1.0 is used.

5.2 System preparation

The preparation of the system, made previously in the Söderhjelm group, was conducted according to the following procedure: The Amber ff99sb force field [14] was used for the protein, the Generalized Amber force field (GAFF) [15] for the ligand and TIP3P [16] for water. The antechamber program from Amber 10 [17] was used to select GAFF parameters for the ligand. For each ligand pose (ortho and hexa), the ligand was geometry-optimized at the B3LYP/6-31G* level. Atomic charges were determined by the RESP method [18] based on the HF/6-31G* electrostatic potential, at points selected according to the Merz-Kollman scheme [19]. Finally, the charges were averaged over the two poses to enable the use of an identical force field in the two cases without being biased to one pose or the other. Each system was first energy-minimized using the steepest-descent method and then equilibrated in the NPT ensemble for 1 ns with the temperature gradually increased from 0 to 300 K during the first 0.2 ns and with all backbone atoms and all non-hydrogen ligand atoms restrained to the crystal structure with a force constant of 1000 kJ/mol nm^{-2}. This was followed by a 10 ns equilibration with all C_α atoms restrained with the same force constant, but no restraints on the ligand. Finally, a 20 ns MD simulation with weaker restraints on the C_α atoms, 200 kJ/mol nm^{-2} was performed, and this run was also used for analysis. The last snapshot of each such MD simulation was used as a starting point for the alchemical perturbation calculations.

5.3 Restraints

For positional restraints, a harmonic restraint force constant of 200 kJ/mol nm^2 was used, which was found in previous work to be a good balance between having an effective restraint and as minimal restraint as possible. For dihedral restraints, the angle fluctations in an unrestrained MD simulation were fitted with a Gaussian function and the force constant set to the approximate value of σ, which was 100 kJ mol^{-1} rad^{-2}. When doubling the force constant, it was validated that the fluctuations decreased by a factor of $\sim \sqrt{2}$.

The Plumed "UPPER_WALLS" function was used to create the water barrier when irregular water interactions were found. The "UPPER_WALLS" function defines a wall for a collective variable, which activates a restraint on the system when the value of the collective variable is greater than a certain limit("AT"). The collective variable in this case was set as the coordination between two groups of atoms. The five atom-membered ring was defined as one coordination group and all the water oxygen atoms as another coordination group. The coordination number is calculated according to a switching function, shown in equation 22.

$$\sum_{i\in A}\sum_{i\in B} s_{ij}, s_{ij} = \frac{1-(\frac{r_{ij}-d_0}{r_0})^n}{1-(\frac{r_{ij}-d_0}{r_0})^m} \tag{22}$$

In equation 22, the n parameter has been set to 16, the m parameter set to 32, the r$_0$ parameter which is the distance between the atoms of the two groups needed to be below to be counted is set to 0.8 nm and the d$_0$ parameter has been left at the default of 0 nm. The restraining potential(or barrier) acting on the system when the coordination exceeds the limit “”’AT” is governed by equation 23.

$$\sum_i k_i((x_i - a_i + o_i)/s_i)_i^e \tag{23}$$

The barrier was set with a force constant κ (k$_i$) set to 2.5 kJ mol^{-1} and the limit “AT” (a$_i$ in equation 23) set to 33. Other constants were left at default.

5.4 System analysis

Visual inspection was performed using Visual Molecular Dynamics program(“VMD”) [9], version 1.9.1. Clustering is performed using the Gromacs gmx cluster program, using the gromos method [10] with a 0.05 nm RMSD cutoff considering the the non-hydrogen atoms of the ligand. Restraints were evaluated using a pseudo-pertubation method where a number of simulations are made where the restraints are decreased down to zero, multiplied by the change in restraint in percent and divided by the restraint. An exponential averaging is made on these values, which can then be summed to give the contribution in energy(and its error) of having a restraint.

Free energies were computed with the Gromacs “gmx bar” software. Evaluation of different setups is done by considering both the standard deviation in $\frac{\partial H}{\partial \lambda}$ and the autocorrelation time. Both of these measures should influence the error computed by the BAR method, but neither of these measures should be seen as perfect.

In the investigation of residue-wise energy contributions, the finite difference method was used, see equation 24.

$$\text{Derivative of residue interaction energy} = \frac{V_i(\lambda + 0.001) - V_i(\lambda - 0.001)}{(\lambda + 0.001) - (\lambda - 0.001)} \tag{24}$$

In equation 24, V_i is the energy contribution for residue i, including water-ligand and internal ligand interactions. Adding up all the residue-wise contributions would give $\frac{\partial H}{\partial \lambda}$. The difference in λ was chosen to be 0.002. The λ values to be investigated were chosen based on which λ had the highest standard deviation in $\frac{\partial H}{\partial \lambda}$.

6 Results and Discussion

6.1 Finding the optimal setup for low-variance simulations

The high variance found in previous work was believed to be caused by large molecular fluctuations in the transformation to dummy atoms, where the Lennard-Jones interactions are suppressed. A postulated solution gave that another kind of restraint could avoid these. If the restraint as well could be minimal, the contribution of the restraint to the free energy could be low, making evaluation easy. For comparison, the BAR result for not having any restraint at all is -68.0 ± 2.9 kJ mol^{-1}. The important number is the error, not the actual free energy, as any restraint would give a contribution to the free energy, but the error is comparable between simulations.

6.1.1 Restraining ten dihedrals in the ligand

The first restraint suggested was to find key dihedrals in the ligand and apply a restraint onto those. Each dihedral chosen would be a representative of a rotatable bond. Ten such dihedrals were identified by visual inspection. In table 1 the atoms of each dihedral is given, refering to the numbering given in figure 3.

Table 1: Table of dihedrals, each representative of a rotatable bond. Atom numbers refer to the numbering used in figure 3.

Designation	Atoms
t1	8, 15, 16, 18
t2	34, 19, 16, 18
t3	34, 19, 20, 23
t4	20, 19, 34, 36
t5	25, 23, 20, 19
t6	63, 62, 65, 66
t7	63, 62, 47, 45
t8	51, 48, 47, 45
t9	42, 45, 47, 48
t10	47, 48, 51, 53

The restraint values for these dihedrals were determined by analysing a MD-simulation with only a restraint on the C_α atoms on the protein for stability. Based on the histogram of each dihedral angle for the Ortho simulation(see figure 5), we concluded that the distribution was approximately normal, so an average for these could be used. The Hexa pose seemed to have two sub-poses in terms of some dihedrals; one major and one minor, see figure 6. Further, it would seem that once the pose changed from the starting pose, it would not return. For this simulation, the average of the major pose was chosen. The existence of multiple poses is discussed in section “The states of Hexa”. To avoid problems with multiple poses, only Ortho is considered when investigating the effect of various restraints.

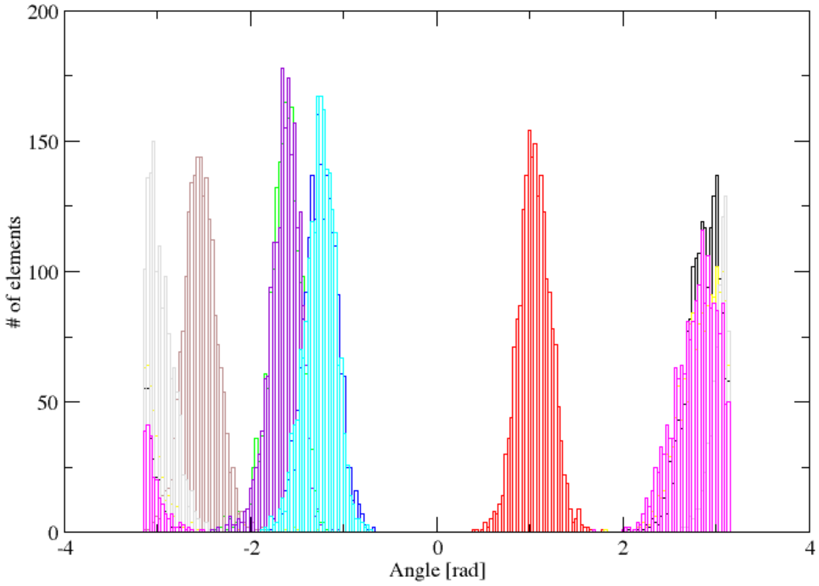

Figure 5: Angle distributions for the ten selected dihedrals in the Ortho pose. On the x-axis, the angle in radians is shown, on the y-axis is the frequency of each angle for each dihedral.

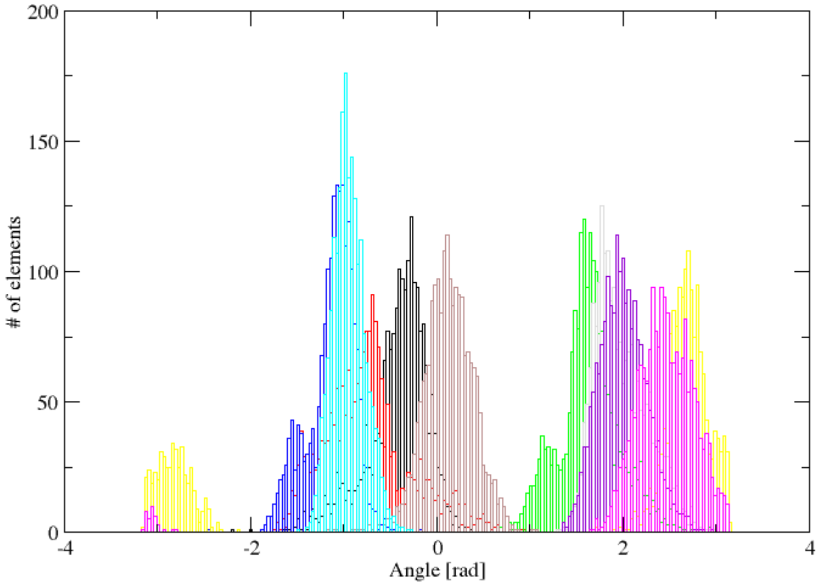

Figure 6: Angle distributions for the ten selected dihedrals in the Hexa pose.On the x-axis, the angle in radians is shown. The scale is periodic so that pi and -pi correspond to the same angle. On the y-axis is the frequency of each angle for each dihedral.

With this restraint, a perturbation to dummy atoms was made for the Ortho pose. Compared to a perturbation without restraints, this gave no significant improvement, see figures 7 and 8. The autocorrelation time, which is a measure of how long time is needed for uncorrelated data sets, even increases, which decreases the amount of statistically useable data. This is reflected in the BAR result, -51.4 ± 5.0 kJ mol^{-1}.

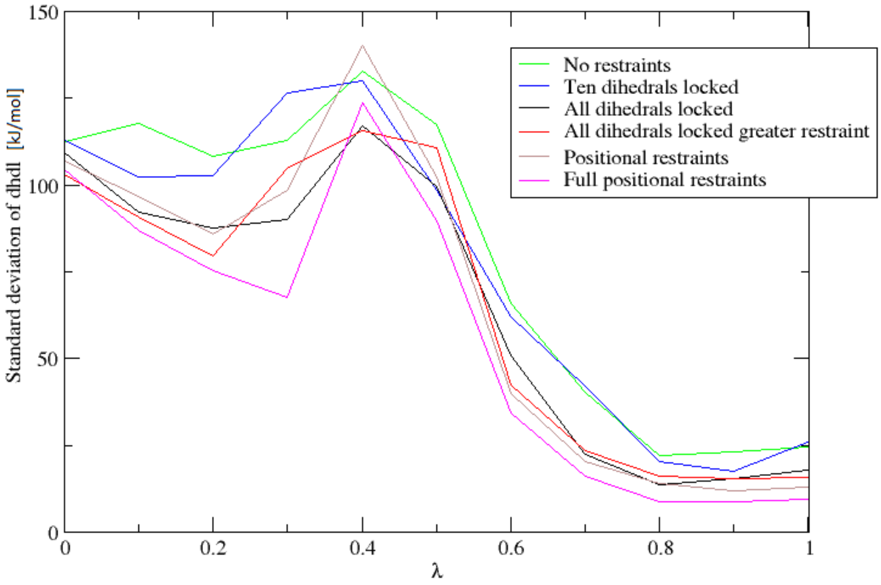

Figure 7: Standard deviation of dhdl for different restraints and different values of λ represented on the x-axis. Each restraint corresponds to a 5 ns simulation where the three first nanoseconds have been discarded, the standard deviation applies to the last two nanoseconds of the simulation, see “Methods”.

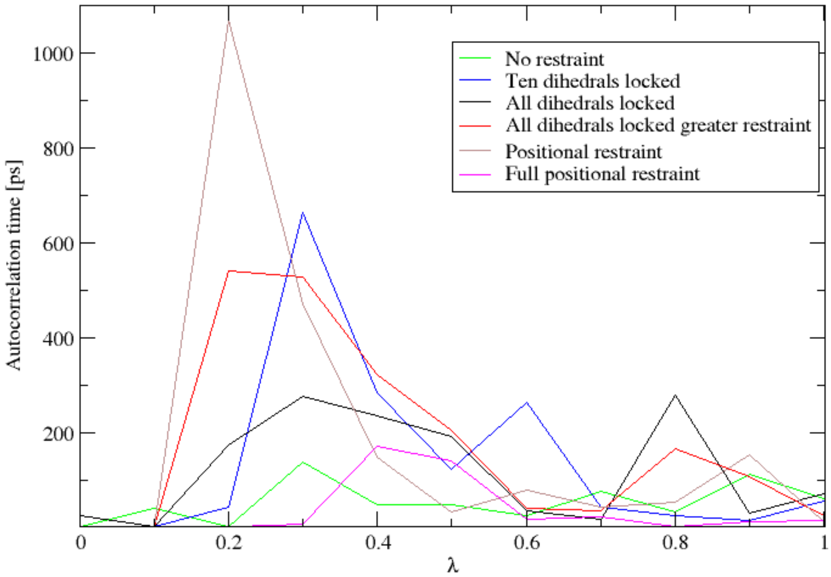

Figure 8: Autocorrelation time for different restraints and different values of λ represented on the x-axis. Each restraint corresponds to a 5 ns simulation where the first nanosecond have been discarded, the last four nanoseconds have been used for the calculation of the autocorrelation, see “Methods”.

The α value of the soft-core potential was varied. As seen in figures 9 and 10, this mostly moved the problematic values of λ to other values of lambda. Decreasing the soft-core α value to 0.5(from 1.0) seemed to have the best effect, both considering autocorrelation time and standard deviation in $\frac{\partial H}{\partial \lambda}$ which also shows in the BAR result of -65.8 ± 3.4 kJ mol^{-1} although it is still worse than having no restraint at all, when considering autocorrelation and the BAR result. Increasing it to 1.5 would seem to increase the standard deviation in $\frac{\partial H}{\partial \lambda}$, but improve the autocorrelation time, with a total effect of becoming worse when looking at the BAR result, -50.9 ± 6.8 kJ mol^{-1}.

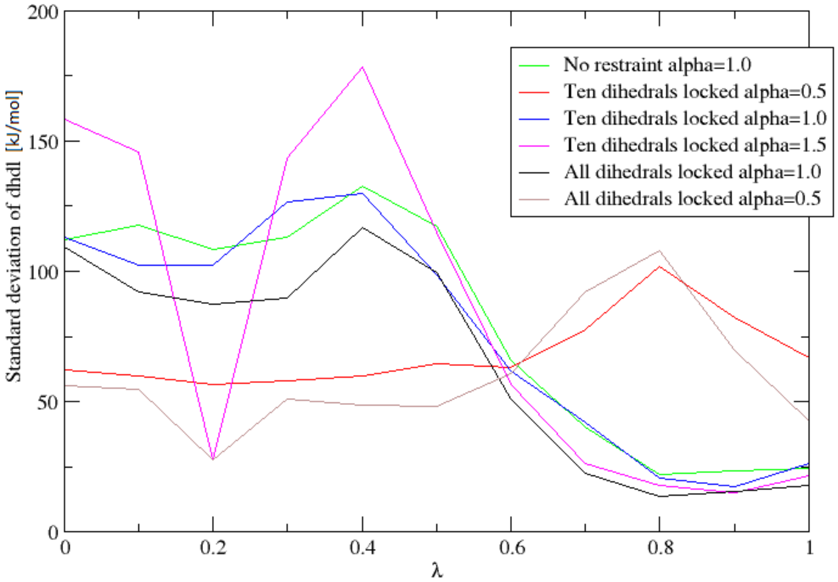

Figure 9: Standard deviation of dhdl for different restraints with different soft-core α values. Different values of λ are represented on the x-axis. Each restraint correspond to a 5 ns simulation where the three first nanoseconds has been discarded, the standard deviation applies to the last two nanoseconds of the simulation, see "Methods".

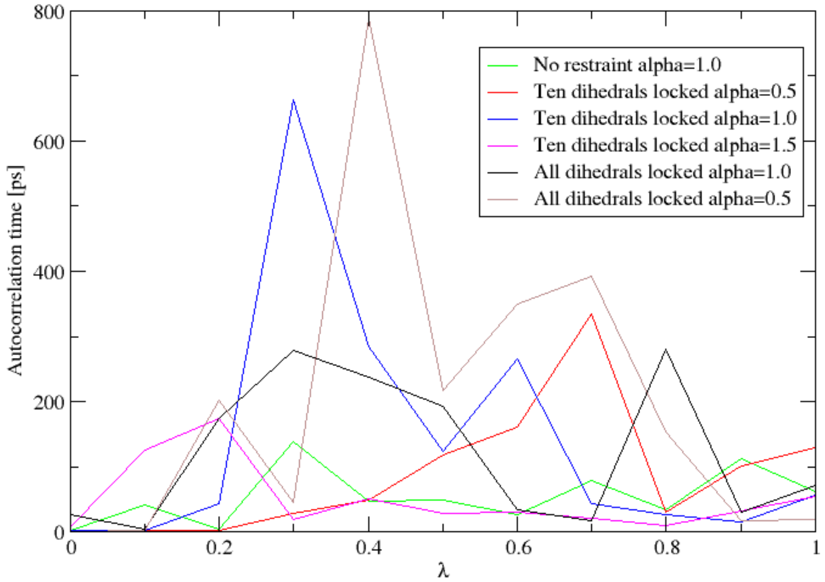

Figure 10: Autocorrelation time for different restraints with varying soft-core α values. Different values of λ are represented on the x-axis. Each restraint correspond to a 5 ns simulation where the first nanosecond has been discarded, the last four nanoseconds have been used for the calculation of the autocorrelation, see “Methods”.

6.1.2 Positional restraint: restraint to C_α atoms in protein, non-hydrogen atoms in ligand

A less minimalistic restraint than restraining a few dihedrals is obtained if all the non-hydrogen atoms of the ligand and all the C_α atoms of the protein are restrained at their positions. A perturbation into dummy atoms was performed with this setting for the Ortho pose, see figures 7 and 8 for the evaluation of this. This was seen as a slight improvement compared to the case when ten dihedrals were restrained in terms of $\frac{\partial H}{\partial \lambda}$, but a considerable worsening in terms of autocorrelation. However, the BAR result is surprisingly good, with a value of -36.1 ± 3.1 kJ mol^{-1}, which is a bit suspicious. However, the worsening is concentrated to one value of λ in the autocorrelation graph, so the BAR method would have neighbouring simulations to draw low-error data from, giving this unexpectedly low error compared with the case of restraining ten dihedrals.

6.1.3 Restraining all dihedrals in the ligand

Instead of restraining down a few key dihedrals, one could instead restrain all the dihedrals present in the ligand. Due to the number of dihedrals(213), the average angle was computed programmatically, after taking care of periodicity. Histograms of angle distributions were visually inspected and it was found that some dihedrals had multiple states; the angles were not normally distributed. Closer inspection showed that these dihedrals contained hydrogen atoms, like methyl groups, which naturally have multiple minima for the dihedral angle. The solution was to leave out all such dihedrals so that only dihedrals not containing hydrogens remained.

The evaluation of this restraint(found in figures 7 and 8) showed it to be a slight improvement, but not satisfactory. The BAR result also shows this, with a value of -20.7 ± 3.7 kJ mol^{-1}.

The idea that the force constant for the restraint had not been enough was tested by doubling the force constant for the restraint (to 200 kJ mol^{-1} rad^{-2}). The results shown in figures 7 and 8 indicate that the standard deviation is similar while the autocorrelation time is different but neither was satisfactory considering standard deviation in $\frac{\partial H}{\partial \lambda}$ and autocorrelation. However, the BAR result actually indicated that it performed better than having no restraint at all, -30.6 ± 2.6 kJ mol^{-1}, which is somewhat suspicious.

We tested whether changing the α of the soft-core potential could improve the result. As seen in figures 9 and 10, this was not the case, the autocorrelation became much greater(in particular for $\lambda = 0.4$) while the standard deviation in $\frac{\partial H}{\partial \lambda}$ was improved, providing a similar result as using the "ten dihedral" restraint. However this had a very low error according to the BAR result(-16.8 ± 2.2 kJ mol^{-1}). One should keep in mind that the worsening in autocorrelation is in particular related to one λ, which has low-autocorrelation neighbours for the BAR method to draw low-error data from.

6.1.4 Full positional restraint: restraint to all non-hydrogen atoms in protein and ligand

We tested having a positional restraint on all non-hydrogen atoms of both the ligand and the protein. In the evaluation of this restraint(result in figures 7 and 8), it was considered the best restraint so far, but still not satisfactory. The BAR outcome was comparable with the positional restraint, -11.6 ± 3.2 kJ mol^{-1}.

6.1.5 The major cause of fluctuations

Given figure 7, it would seem that $\lambda = 0.3$, 0.4 has the greater fluctuations and would be most interesting to investigate. The approach was to calculate the residue-wise energy contributions for the value of $\frac{\partial H}{\partial \lambda}$ at the specific λ. The finite difference method(equation 24 in "Methods" section) was used to calculate the derivative for the chosen λ by evaluating the energy for λ -0.001 and λ + 0.001. The results are shown in figures 11 and 12.

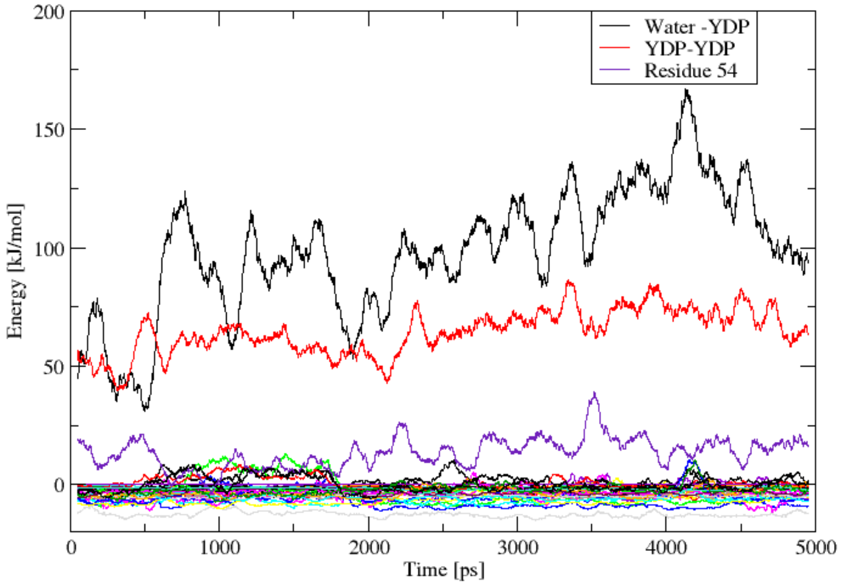

Figure 11: Contributions to dhdl for each protein residue, internal interactions and water interactions for $\lambda = 0.3$. Only residues contributing more than 4 kJ/mol at any time are considered.

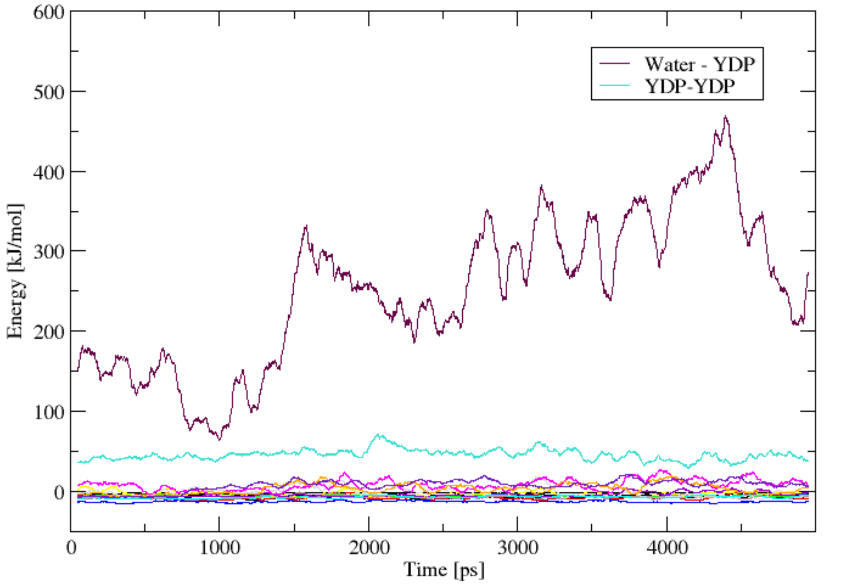

Figure 12: Contributions to dhdl for each protein residue, internal interactions and water interactions for $\lambda = 0.4$. Only residues contributing more than 4 kJ/mol at any time are considered.

Both cases indicated that water - ligand("YDP") was the greatest contributor, followed by internal ligand interactions. For $\lambda = 0.4$, the internal interactions seemed well-behaved as there seemed to be little fluctuations(see figure 12), and no single residue seemed to have a much higher energy contribution than any other residue. For $\lambda = 0.3$, residue 54 seemed to contribute more to the total energy than any other residue. Water interactions with the ligand would be present for a ligand solvation simulation, but as figure 13 shows, there is little error considering autocorrelation for the case of a solvated ligand.

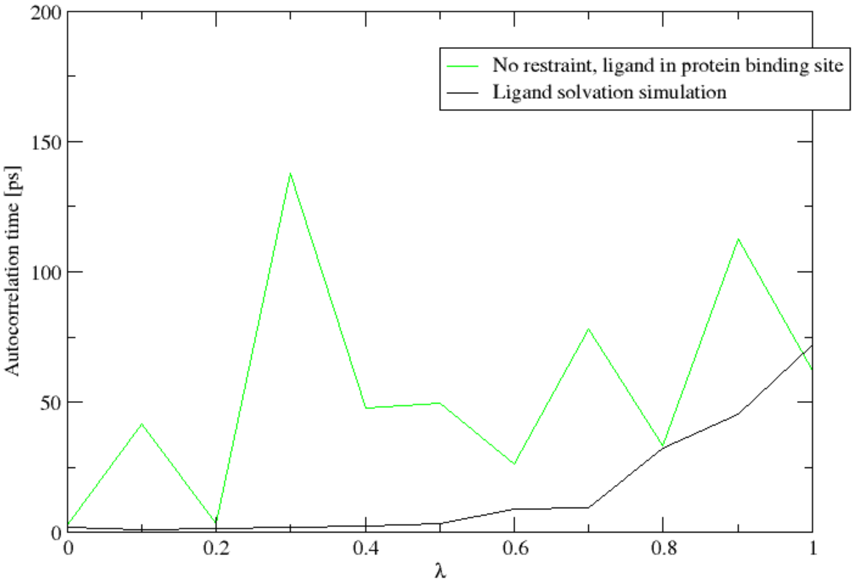

Figure 13: Autocorrelation time for a ligand solvation simulation without protein and compared to a protein-ligand simulation with no restraints.

Comparing figures 13 and 8, it would seem autocorrelation is low when having no restraint at all. A hypothesis was formulated: as λ increases, there is an empty pocket of space created between the protein and the ligand. In case of restraints on the protein, this space would not be filled up by the protein. Instead, there would be water molecules that would be able to move into this space, but the protein would partly hinder this movement, thereby creating large timescale fluctuations.

A solution where a barrier is be set up to stop entrance of water (apart from the water molecules already there) was conceived and applied together with full positional restraints, which yielded a considerable improvement, see figures 14 and 15. In these figures it is also shown that the barrier does not affect the early λ values, which is in accordance with the hypothesis. The improvement shown in figures 14 and 15 is in agreement with the BAR result, 114.5 ± 1.6 kJ mol^{-1}. A barrier to prevent water molecules from entering could thus be a tool for bringing down the error.

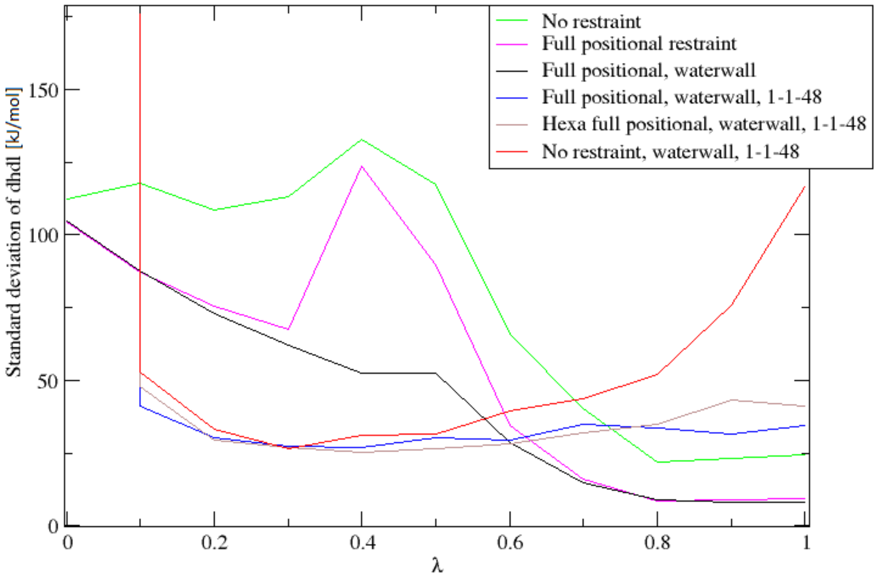

Figure 14: Standard deviation of dhdl for different restraints, different soft-core potentials and with or without water barrier. Different values of λ are represented on the x-axis. Each restraint correspond to a 5 ns simulation where the three first nanoseconds has been discarded, the standard deviation applies to the last two nanoseconds of the simulation, see “Methods”.

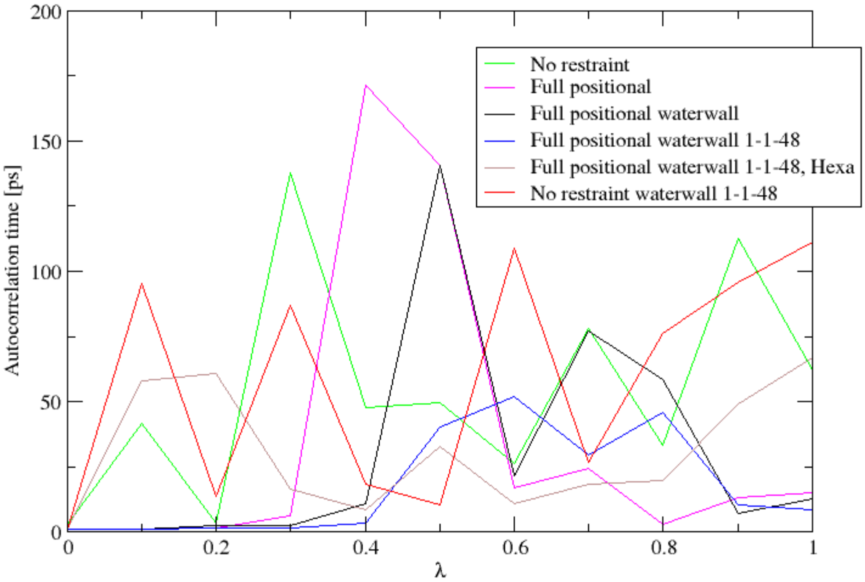

Figure 15: Autocorrelation time for different restraints, different soft-core potential and with or without water barrier. Different values of λ are represented on the x-axis. Each restraint correspond to a 5 ns simulation where the first nanosecond has been discarded, the last four nanoseconds have been used for the calculation of the autocorrelation, see "Methods".

6.1.6 The 1-1-48 soft-core potential

Visual studies in VMD showed that solvent molecules could enter into the hole of the benzene ring in the ligand arms for $\lambda = 0.5$. This behaviour is attributed to the soft-core potential, as the interaction energy between any two non-bonded atoms does not increase very much when they come very close to each other for higher values of λ, see figure 2. At some λ, water molecules would be able to enter; this would be a big difference compared to the neighboring λ and thus lead to large errors.

Other kinds of soft-core potentials were therefore considered, where the 1-1-48 soft-core potential in [5] was an alternative. A simulation was set up to try this soft-core potential, including the best settings found so far (full positional restraints, water barrier). The result(shown in figure 14) seems to verify the findings of [5] and have a favourable improvement of the method. This statement may seem strange as there is a "spike" in the $\frac{\partial H}{\partial \lambda}$ standard deviation graph which assumes a values of $\sim$ 10 000. The cause of the spike is the much greater non-linearity of the $V(\lambda)$ curve for the 1-1-48 soft-core potential. This would have been a severe problem for methods like thermodynamic integration which would then have needed very dense λ values near the spike, but does not affect the BAR method because of its error-minimizing weigthing.The BAR method yielded 112.7 ± 1.1 kJ mol^{-1}, which is in agreement with the results of autocorrelation time and standard deviation in $\frac{\partial H}{\partial \lambda}$(figures 14, 15).

The success of the 1-1-48 soft-core potential gave rise to hopes that a positional restraint would not be necessary. This was found to not be the case, as seen in figure

14, but did have an acceptable correlation time according to figure 15 and a BAR result of -1.8 ± 1.8 kJ mol^{-1}, somewhat unexpected considering the standard deviation in $\frac{\partial H}{\partial \lambda}$, but is due to the error-minimizing weighting of the BAR method. These settings (full positional restraints, water barrier, 1-1-48 soft-core potential), which so far only had been tried for the Ortho pose, was now tried for the Hexa pose, with similar results, see figures 14 and 15. Surprisingly, the BAR method did not give a similar result, 93.9 ± 1.8 kJ mol^{-1}(for Ortho, the BAR result was 112.7 ± 1.1 kJ mol^{-1}).

6.1.7 Evaluation of restraint

A way to evaluate the effect of the positional restraints (found in Methods) was tried out for Ortho, which show the feasibility of having an accurate estimation if the number of intermediate steps are sufficient, see appendix. In the appendix, where all intermediate data is shown, the key data is the total error: using many intermediate steps returns a lower error than having just a few intermediates. For 20 intermediates, the total error becomes 13.895 kJ/mol, while 94 intermediates returns a total error of 7.843 kJ/mol. Therefore, with even more intermediates the error could be lowered further until a satisfactory error is achieved. If this would not have been the case, then the introduction of a restraint would introduce errors, which would ultimately not make any improvement in having an accurate free energy of binding estimation. The thermodynamic cycle changes to that shown in figure 16 when using restraints. A limitation here is that this cannot be performed for the Hexa pose; as the restraint approaches zero one would ultimately have the unstable simulations that the restraint was meant to prevent. Also, it would seem that a low error requires many intermediates, which may be not computationally feasible. If only few intermediates can be used, one may consider not having a restraint thus avoiding this evaluation and only use the 1-1-48 potential and the water barrier setup. With few intermediates, a larger error may be had if using a restraint than if not.

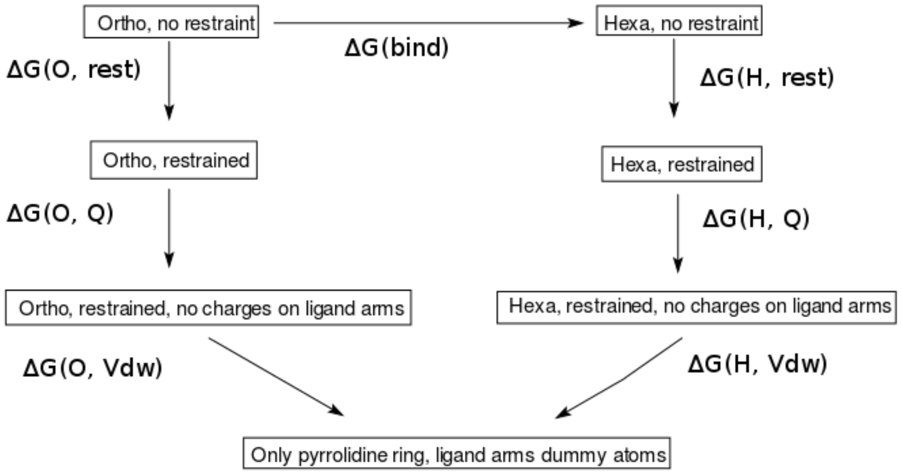

Figure 16: Thermodynamic cycle, with an evaluation of the restraints. Free energy is a state function, any route can be used to calculate the free energy difference between Ortho and Hexa.

From the thermodynamic cycle in 16, the equation describing the free energy difference between Ortho and Hexa becomes equation 25, compare with equation 21.

$$\Delta G(bind) = \Delta G(O, \text{rest}) + \Delta G(O, Q) + \Delta G(O, \text{Vdw}) - (\Delta G(H, \text{Vdw}) + \Delta G(H, Q) + \Delta G(H, \text{rest})) \tag{25}$$

6.2 The states of Hexa

As shown in section “Restraining ten dihedrals in the ligand”, the Hexa pose comprises multiple states identified by different dihedral angles. Once the ligand diverted from the starting pose, it did not return, so the structure could be said to “slip away” from the starting structure, which thus seems unstable. This gives rise to some questions:

- Do the different angles actually correspond to different subposes or do they have no effect structurally?
- Can the relative energy of these poses be estimated?
- Can the force field be altered so that stability is achieved?

A number of trajectories were concatenated (from four 20 ns Hexa simulations and one 5 ns Ortho simulation) and clustered according to RMSD as described in the “Methods” section. The angle of the dihedrals were compared with the clusters formed, see figures 17 and 18.

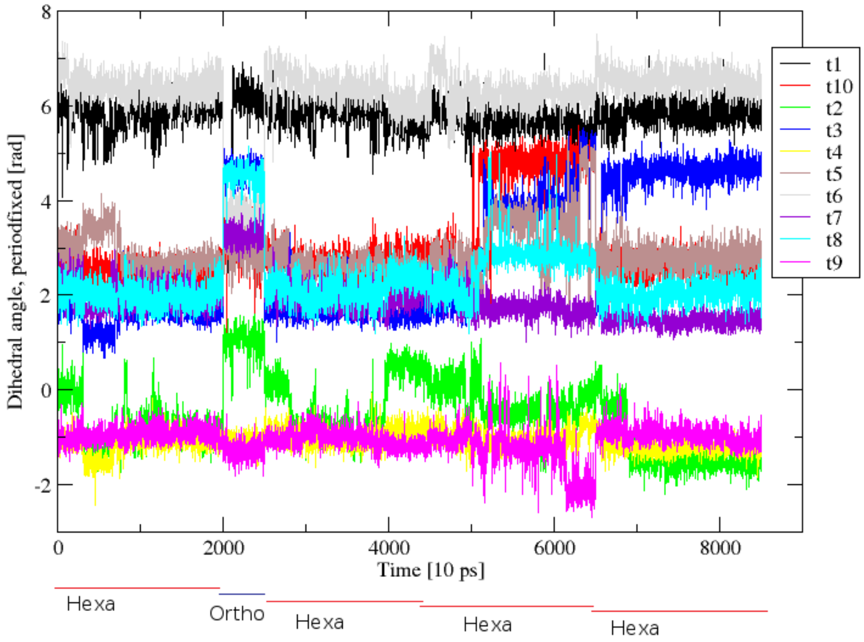

Figure 17: Dihedral angles for concatenated trajectory, beginning with a 20 ns Hexa simulation, then a 5 ns Ortho simulation(starting at 20 000 ps and ending at 25 000 ps) then three consecutive 20 ns Hexa simulations(starting at 25 000 ps, 45 000 ps and 65 000 ps respectively).

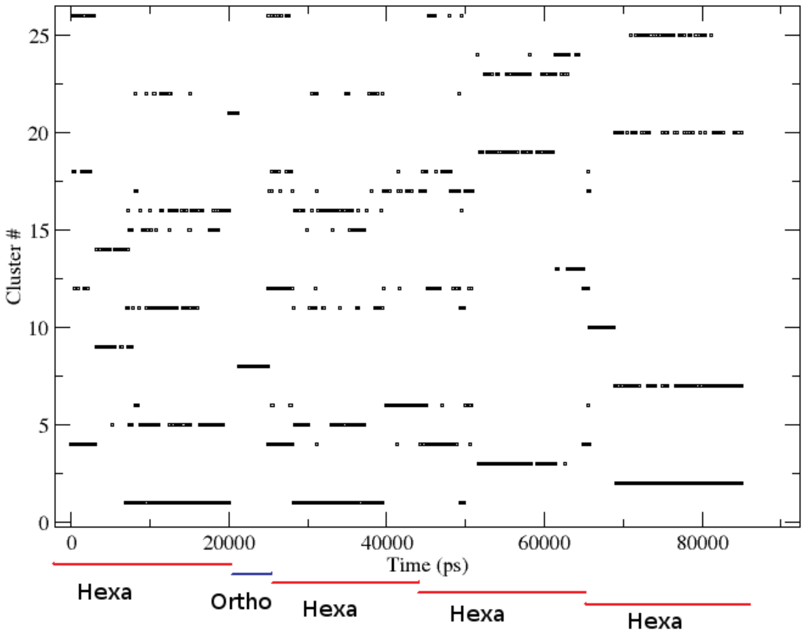

Figure 18: Clusters formed for the same concatenated trajectory as in figure 17, beginning with a 20 ns Hexa simulation, then a 5 ns Ortho simulation(starting at 20 000 ps and edning at 25 000 ps) then three consecutive 20 ns Hexa simulations(starting at 25 000 ps, 45 000 ps and 65 000 ps respectively). Clustering is according to RMSD. The y-axis is the number of the cluster; lower number indicates a greater amount of snapshots have been sorted to that cluster. Only the 25 clusters with the most snapshots shown here, there are in total 400 clusters.

The average potential energy was calculated for the 8 clusters with the greatest amount of snapshots in in them, as shown in table 2.

Table 2: Table of potentials for the most frequent clusters found in figure 18.

Cluster #	Mean of Potentials [kJ mol^{-1}]	Std of Potentials	# snapshots
1	-90.2	22.7	1064
2	-114.7	22.1	1054
3	-94.3	24.0	552
4(Crystal structure, Hexa)	-62.5	22.1	480
5	-112.6	23.7	450
6	-51.3	23.5	445
7	-110.5	22.5	271
8(Ortho cluster)	-33.7	26.4	265

The clustering graph should be interpreted as cluster # 1 being most common, cluster # 2 being second most common and so on. Even if it may seem so, different conformers(which the clusters represent) are not visited simultaneously; for each time step only one conformer is visited. This illusion is because it is very small time steps, one would have to zoom in to observe singular data points for individual time steps. How the clusters differ from each other is seen in figure 19, where the representative structures for the three largest clusters have been superimposed.

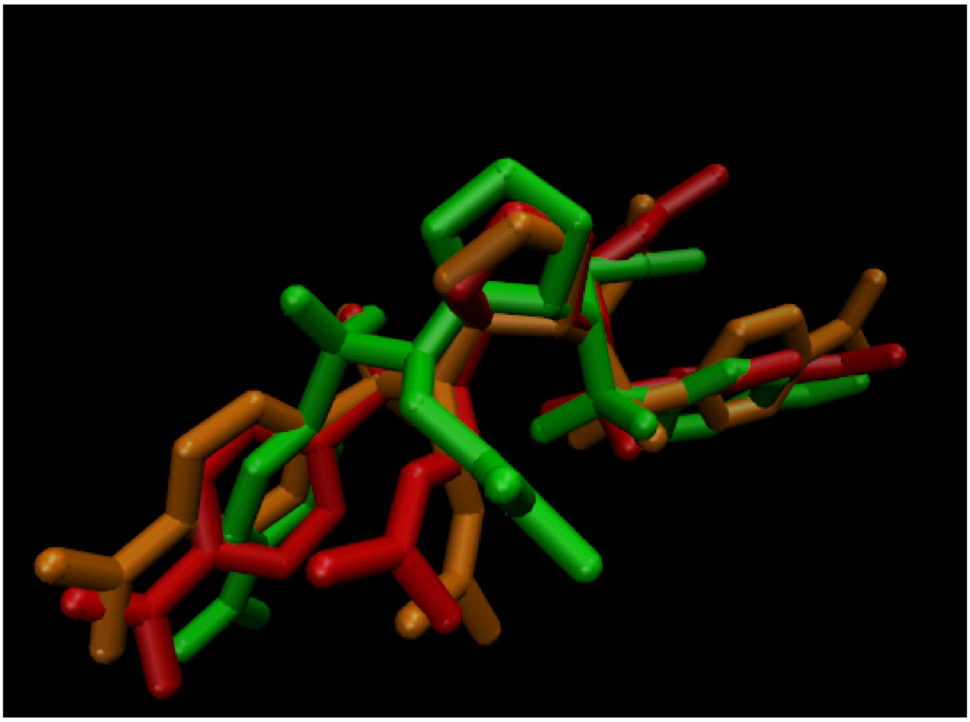

Figure 19: Representative structures for the three largest Hexa clusters from figure 18 superimposed. Red structure is cluster # 1, green structure is cluster # 2 and orange structure is cluster # 3.

It is observed that the Hexa poses are separate from the Ortho pose. The difference between the Hexa poses seem connected to changes in the "t2" dihedral (see table 1 for which atoms this dihedral represents). Surprisingly, the most common cluster does not have the lowest/most negative potential energy, which one would have expected. This shows that there are more effects to consider than just the potential energy for having an accurate energy landscape.

6.2.1 Changing the force field

As the "t2" dihedral seemed to be crucial for the pose assumed, we hypothesised that changing the Ryckaert-Bellemans function governing the potential for this particular dihedral and any redundant dihedrals associated with "t2" could produce a force field in which the Hexa crystal structure would be stable. All these dihedrals had only one parameter in the Ryckaert-Bellemans function being non-zero, namely the C_2 term with a value of 26.21694 kJ mol^{-1}. Investigations primarily focused on changing this value, but also introducing non-zero terms for the other parameters one at the time. One example is the change of the C_2 term into 11.14354 kJ mol^{-1} for four of the dihedrals of the rotatable bond that "t2" represents. The change in Ryckaert-Bellemans function compared to the original function for all the four dihedrals combined is shown in figure 20. Figure 20 was conceived through writing the total potential for the rotatable bond as a function of only the "t2" dihedral angle.

To be able to express the total Ryckaert Bellemans function of several dihedrals describing the same rotatable bond as a function of a single dihedral angle, a constant shift between each pair of coupled dihedrals was assumed. Their values were determined from the distributions obtained from an MD simulation of the Hexa state. The standard deviation of the dihedrals was in the range 0.139 rad to 0.171 rad, while the standard deviation of the difference between the "t2" dihedral and the other dihedrals were in the range 0.075 rad to 0.192 rad, i.e. in the same order of magnitude as the fluctuations of the dihedrals themselves. Overall, the fluctuations were small enough to validate the assumption of a constant shift.

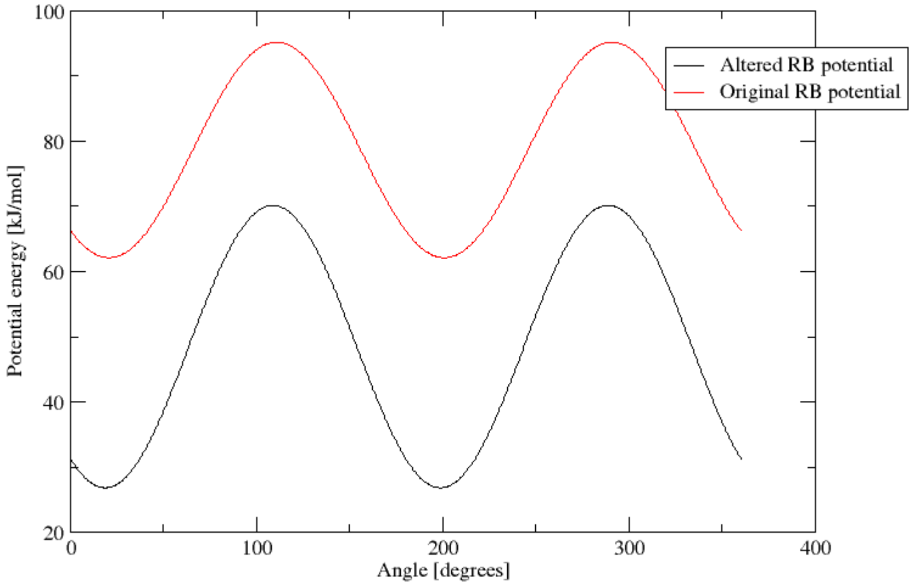

Figure 20: Comparison of the Ryckaert-Bellemans function of the original force field(red line) and the altered force field(black line). Each line is the total potential of four dihedrals modelling the same rotatable bond.

The appearance of the Ryckaert-Bellemans function should be connected to the dynamics of the ligand pose. The changed behaviour of the "t2" dihedral can be seen in figure 21, which should be compared with the t2 dihedral behaviour found in figure 17.

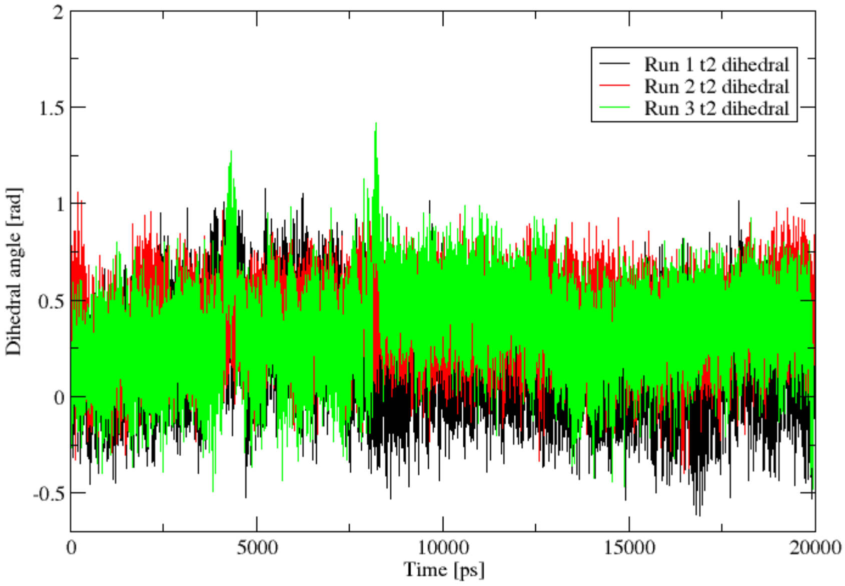

Figure 21: Angle for the t2 dihedral for three independent Hexa simulations using an altered force field, where the Ryckaert-Bellemans C2 term for t2 and associated dihedrals has been changed from 26.21694 kJ mol^{-1} to 11.14354 kJ mol^{-1}.

The “t2” dihedral would seem to be more stable due to the change in the force field, the dihedral does not “slip away” from the starting angle. The total conformational change is better represented by the RMSD, which is shown in figure 22. The results should be compared with the corresponding graph for the original force field, which is shown in figure 23.

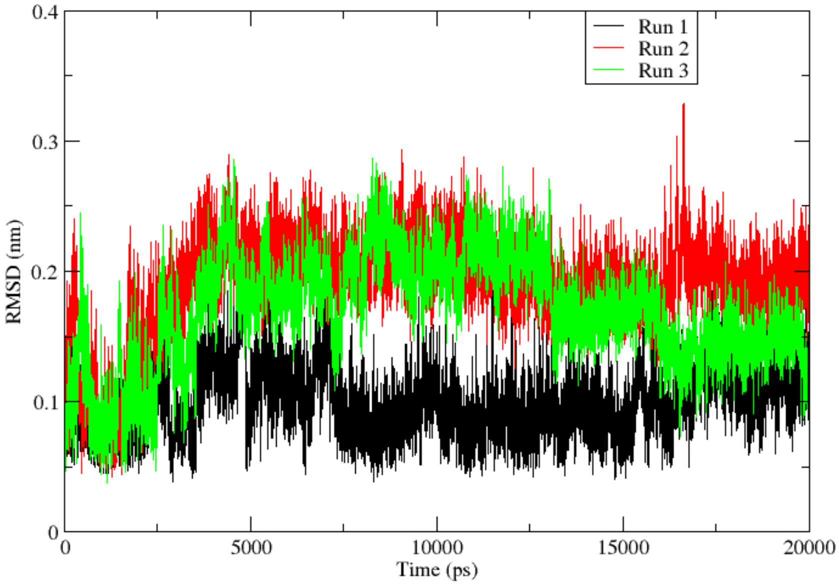

Figure 22: RMSD(nm) of the ligand after alignment to the C_α of the protein, using an altered force field. Three independent simulations are considered. The simulations starts from the Hexa crystal structure.

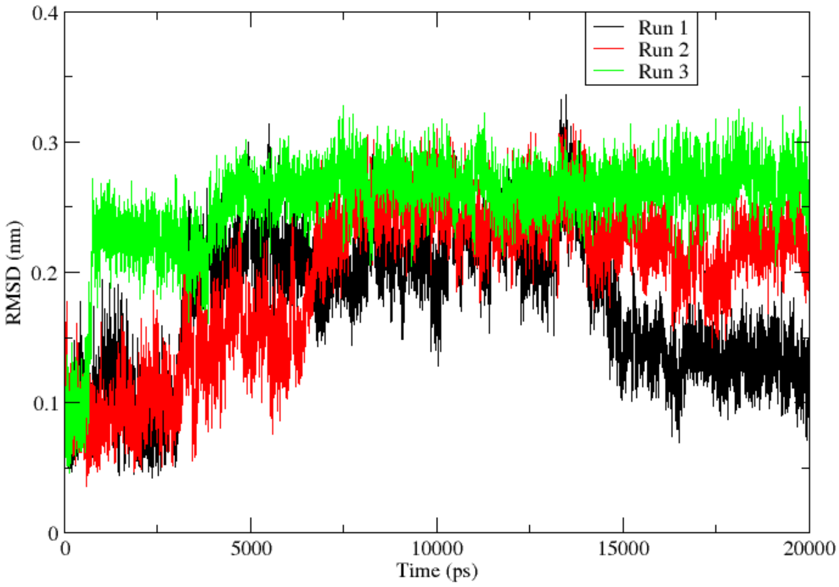

Figure 23: RMSD(nm) of the ligand after alignment to the C_α of the protein, using the original force field. Three independent simulations are considered. Simulation starts from the Hexa crystal structure, no restraints applied.

The goal would be to have the RMSD return to a level which correspond to the fluctuations that the experimental structure would have, which is found where the RMSD just have stabilized but before any conformational changes(after a few 100ps). The changes made does not achieve this according to figure 22.

7 Conclusions

A setup has been found that decreases the error associated with estimating the free energy of binding using alchemical perturbation when the ligand has different poses in the start and end state. This setup uses a water barrier to decrease fluctuations in water interactions, a 1-1-48 soft-core potential for the perturbation of Lennard-Jones interactions and a positional restraint on all non-hydrogen atoms present in both the ligand and the protein.

Throughout these investigations, it can be noted that different measures of the error did not always give a coherent picture of the error. This would advice for caution when trying to make error estimations. There are also inherent limitations in the force field model, which could produce an erroneous free energy of binding, even though the estimated error is small. For exact values, quantum mechanical calculations are necessary.

During simulation the Hexa pose diverges from the crystal structure and does not return to the crystal structure. Investigations show that the ligand pose is governed primarily by one rotatable bond. Force field modifications to make the Hexa pose stable should therefore focus on this rotatable bond. Changes to the Ryckaert-Bellemans function parameters for the rotatable bond have been shown to change the behaviour of the rotatable bond. The changed behaviour seems to improve the stability in the RMSD, approaching the level of RMSD typical for local fluctuations. Understanding how the Ryckaert-Bellemans function affects the ligand dynamics seems key in finding a stable force field for the Hexa pose. However, it is not certain that a stable Hexa pose can be achieved by only changing the ligand force field, other changes may be needed as well.

7.1 Future work

Essential future work would be to do force field development in order to have the Hexa pose stable, including rationalizing the force field modifications with quantum mechanical calculations. The current settings could also be further optimized by trying different values of α in the 1-1-48 soft-core potential, as mentioned in [5]. With a working setup for precise perturbation simulations, it might be possible to obtain results at the quantum mechanics/molecular mechanics level, as in [23].

7.2 Acknowledgements

The author declares no conflict of interests. The computations were performed on computer resources provided by the Swedish National Infrastructure of Computing(SNIC) at HPC2N at Umeå University and LUNARC at Lund University. Many thanks to my supervisor Pär Söderhjelm for his guidance and Soumendranath Bhakat and Sven Wernersson for additional help.

References

[1] David L. Mobley and Pavel V. Klimovich
Perspective: Alchemical free energy calculations for drug discovery
The Journal of Chemical Physics **137**, 230901, 2012.
DOI: 10.1063/1.4769292

[2] David L. Mobley and Ken A. Dill
Binding of Small-Molecule Ligands to Proteins: "What You See" is Not Always "What You Get"
Structure *17* 2009
DOI: 10.1016/j.str.2009.02.010

[3] Andreas Blum, Jark Böttcher, Stefanie Dörr, Andreas Heine, Gerard Klebe and Wibke E. Diedrich
Two Solutions for the Same Problem: Multiple Binding Modes of Pyrrolidine-Based HIV-1 Protease Inhibitors
J. Mol. Biol. (2011) **410**, 745-755
DOI: 10.1016/j.jmb.2011.04.052

[4] C.H. Bennett
Efficient estimation of free energy differences from Monte Carlo data
Journal of Computational Physics, **22**, 2, 245-268 (1976)
DOI: 10.1016/0021-9991(76)90078-4

[5] Tri T. Pham and Michael R. Shirts
Identifying low variance pathways for free energy calculations of molecular transformations in solution phase
The Journal of Chemical Physics **135**, 034114 (2011)
DOI: 10.1063/1.3607597

[6] Teodor Rodin
Theoretical characterisation of two binding modes for the same ligand to HIV-1 protease
Lund University, Söderhjelm Research Group (2014)

[7] M.J. Abraham, T. Murtola, R. Schulz, S. Páll, J.C. Smith, B. Hess, E. Lindahl
GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers
SoftwareX **1-2**, 19-25 (2015) DOI: 10.1016/j.softx.2015.06.001

[8] G.A. Tribello, M. Bonomi, D. Branduardi, C. Camilloni, G. Bussi
PLUMED 2: New feathers for an old bird
Computer Physics Communications **185**, 2, 604–613 (2014) DOI: 10.1016/j.cpc.2013.09.018

[9] W. Humphrey, H. Dalke, K. Schulten
VMD - Visual Molecular Dynamics
Journal of Molecular Graphics, **14**, 33-38 (1996)
<http://www.ks.uiuc.edu/Research/vmd/>

[10] X. Daura, K. Gademann, B. Jaun, D. Seebach, W.F. van Gunsteren, A.E. Mark
Peptide folding: when simulation meets experiment
Angewandte Chemie International Edition **38**, 236-240 (1999)
DOI: 10.1002/(SICI)1521-3773(19990115)38:1/2<236::AID-ANIE236>3.0.CO;2-M

[11] G. Bussi, D. Donadio, M. Parrinello
Canonical sampling through velocity rescaling
Journal of Chemical Physics, **126**, 014101 (2007) DOI: 10.1063/1.2408420

[12] M. Parrinello, A. Rahman
Polymorphic transitions in single crystals: A new moleculardynamics method.
Journal of Applied Physics, **52**, 182–7190 (1981) DOI: 10.1063/1.328693

[13] Alchemistry.org, 2013. *Simulating states of interest* [online]. Available at <http://www.alchemistry.org/wiki/Simulating_States_of_Interest>. [Accessed 24 August 2016]

[14] Viktor Hornak, Robert Abel, Asim Okur, Bentley Strockbine, Adrian Roitberg, and Carlos Simmerling
Comparison of multiple Amber force fields and development of improved protein backbone parameters
Proteins: Structure, Function, and Bioinformatics **65**, 3, 712-725 (2006)
DOI: 10.1002/prot.21123

[15] J. Wang, R.M. Wolf, J.W. Caldwell, P.A. Kollman, and D.A. Case
Development and testing of a general amber force field
Journal of Computational Chemistry, **25**, 9, 1157-1174 (2004)
DOI: 10.1002/jcc.20035

[16] W.L. Jorgensen, J. Chandrasekhar, J.D. Madura, R.W. Impey and M.L. Klein
Comparison of simple potential functions for simulating liquid water
The Journal of Chemical Physics, **79**, 926 (1983)
DOI: 10.1063/1.445869

[17] D.A. Case, T.A. Darden, T.E. Cheatham III, C.L. Simmerling, J. Wang, R.E. Duke, R. Luo, M. Crowley, R.C. Walker, W. Zhang, K.M. Merz, B. Wang, S. Hayik, A. Roitberg, G. Seabra, I. Kolossvary, K.F. Wong, F. Paesani, J. Vanicek, X. Wu, S.R. Brozell, T. Steinbrecher, H. Gohlke, L. Yang, C. Tan, J. Mongan, V. Hornak, G. Cui, D.H. Mathews, M.G. Seetin, C. Sagui, V. Babin, and P.A. Kollman
Amber 10, University of California, San Francisco, United States (2008)

[18] C.I. Bayly, P. Cieplak, W.D. Cornell and P.A. Kollman
A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model
Journal of Physical Chemistry, **97**, 10269 (1993)
DOI: 10.1021/j100142a004

[19] Brent H. Besler, Kenneth M. Merz and Peter A. Kollman
Atomic charges derived from semiempirical methods
Journal of Computational Chemistry, **11**, 4, 431-439 (1990)
DOI: 10.1002/jcc.540110404

8 References for Theory section

[20] *Atkins' Physical Chemistry* Peter Atkins and Julio de Paula 9th ed, 2010 Oxford University Press. ISBN: 978-0-19-954337-3

[21] *Molecular Modelling: Principles and Applications* Andrew R. Leach 2nd ed 2001 Prentice Hall. ISBN: 0-582-38210-6

[22] M.J. Abraham, D. van der Spoel, E. Lindahl, B. Hess. and the GROMACS development team *GROMACS USER MANUAL Version 5.0.4* <www.gromacs.org> (2014)

[23] Martin A. Olsson, Pär Söderhjelm and Ulf Ryde
Converging Ligand-Binding Free Energies Obtained with Free-Energy Perturbations at the Quantum Mechanical Level
Journal of Computational Chemistry, **37**, 1589-1600 (2016)
DOI: 10.1002/jcc.24375

[24] Alchemistry.org, 2013. *Exponential Averaging* [online]. Available at <http://www.alchemistry.org/wiki/Exponential_Averaging>. [Accessed 17 August 2016]

9 Appendix

9.1 Evaluation of restraints

Table 3: Table of error in evaluating restraint

Force constant	Many intermedirates		Few intermediates	
	Step	Absolute error[kJ/mol]	Step	Absolute error[kJ/mol]
f=0.02	0.01	0.0667	0.1	0.738
f=0.04	0.01	0.1238		
f=0.06	0.01	0.0443		
f=0.08	0.01	0.0357		
f=0.10	0.01	0.0582		
f=0.12	0.01	0.0843		
f=0.14	0.01	0.0471		
f=0.16	0.01	0.0154		
f=0.18	0.01	0.0675		
f=0.2	0.1	0.2160	0.1	0.236
f=0.4	0.1	0.1198	0.1	0.121
f=0.6	0.1	0.0802	0.1	0.078
f=0.8	0.1	0.0775	0.1	0.087
f=1	0.1	0.0569	0.1	0.061
f=1.2	0.1	0.0832	0.1	0.085
f=1.4	0.1	0.0524	0.1	0.046
f=1.6	0.1	0.0420	0.1	0.045
f=1.8	0.1	0.0627	0.1	0.066
f=2	0.1	0.0482	0.1	0.044
f=2.2	0.1	0.0459		
f=2.4	0.1	0.0569		
f=2.6	0.1	0.0459		
f=2.8	0.1	0.0493		
f=3.0	0.1	0.0405		
f=3.2	0.1	0.0326		
f=3.4	0.1	0.0394		
f=3.6	0.1	0.0390		
f=3.8	0.1	0.0274		
f=4.0	0.1	0.0365		
f=4.2	0.1	0.0291		
f=4.4	0.1	0.0289		
f=4.6	0.1	0.0224		
f=4.8	0.1	0.0397		
f=5.0	0.1	0.0332		
f=5.2	0.1	0.0255		
f=5.4	0.1	0.0290		
f=5.6	0.1	0.0214		
f=5.8	0.1	0.0240		
f=6.0	0.1	0.0282		

f=6.2	0.1	0.0256		
f=6.4	0.1	0.0288		
f=6.6	0.1	0.0201		
f=6.8	0.1	0.0213		
f=7.0	0.1	0.0276		
f=7.2	0.1	0.0201		
f=7.4	0.1	0.0213		
f=7.6	0.1	0.0292		
f=7.8	0.1	0.0196		
f=8.0	0.1	0.0280		
f=8.2	0.1	0.0245		
f=8.4	0.1	0.0181		
f=8.6	0.1	0.0193		
f=8.8	0.1	0.0184		
f=9.0	0.1	0.0208		
f=9.2	0.1	0.0206		
f=9.4	0.1	0.0198		
f=9.6	0.1	0.0190		
f=9.8	0.1	0.0174		
f=10	1	0.2823		
f=12	1	0.3188		
f=14	1	0.3457		
f=16	1	0.1632		
f=18	1	0.1113		
f=20	1	0.1115	10	5.098
f=22	1	0.1698		
f=24	1	0.1265		
f=26	1	0.1685		
f=28	1	0.1054		
f=30	1	0.0994		
f=32	1	0.0964		
f=34	1	0.0889		
f=36	1	0.0647		
f=38	1	0.0747		
f=40	1	0.0686	10	5.098
f=42	1	0.0723		
f=44	1	0.0576		
f=46	1	0.0535		
f=48	1	0.0577		
f=50	1	0.0515		
f=52	1	0.0544		
f=54	1	0.0492		
f=56	1	0.0440		
f=58	1	0.0522		
f=60	2	0.0888	10	0.779
f=64	2	0.0904		

f=68	2	0.0826		
f=72	2	0.0831		
f=76	2	0.1044		
f=80	10	0.4414	10	0.476
f=100	10	0.4643	10	0.510
f=120	10	0.3590	10	0.362
f=140	10	0.2608	10	0.305
f=160	10	0.2172	10	0.225
f=180	10	0.2671	10	0.293
Total		7.84331		13.895