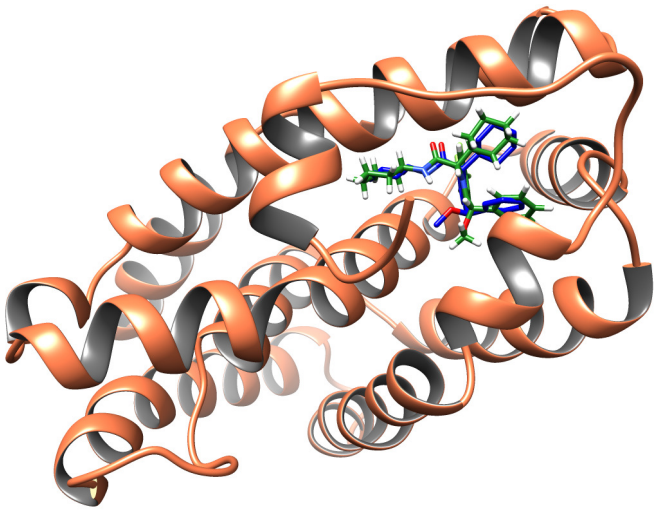
Using Molecular Dynamics and Enhanced Sampling to Predict Binding Poses Beyond The Rigid–Docking Approximation

Emil Åberg

Supervisor: Pär Söderhjelm
Center for Molecular Protein Science, CMPS
Faculty of Engineering (LTH), Lund University

March 1, 2017



The predicted pose for the ligand (green) aligned on the corresponding experimental crystallographic pose (blue) for ligand 22 and the FXR receptor (orange).

Abstract

A computational method is described and tested for prediction of ligand–binding poses between the human farnesoid X receptor and a set of 36 potential agonists, provided by the D3R Grand Challenge 2016. Using tools such as Molecular Docking, Molecular Dynamics, Reconnaissance metadynamics and cluster analysis, the method is an attempt to predict the binding pose without being biased by experimental data. When comparing the predicted poses with the crystal structures, more than half of the ligands were predicted accurately. It is shown that the accuracy of the Molecular Docking is very conformation dependent, as the flexibilty of two α–helices adjacent to the active site makes it difficult for docking to predict the correct pose. Molecular Dynamics are dependent on the predictions from docking, and the force field (GAFF) used for the ligands may be the reason for that only 3 of the accurately predicted poses were refined further. Reconnaissance metadynamics did not result in finding any better poses with the collective variables set used.

More effort is needed to determine a better set of collective variables, which are able to take the flexibility of the α–helices and the positions of the side chains into consideration, as well as possibly enable Reconnaissance metadynamics to overcome the short–comings of docking.

Contents

1 Introduction

Knowledge of the mechanism behind biophysical, biological and chemical phenomena are important for the development of future therapheuticals. If the mechanism is not known, it would have to be determined either in the laboratory or theoretically. One interesting instance where *in silico* methods can be used to study the mechanism, is the binding between a small molecule and a macromolecule.

There are different commonly used physical computational methods available to study binding events, such as molecular docking, MM–PBSA (Molecular Mechanics with Poisson–Boltzmann + Surface Area), RBFE (Relative Binding Free Energy) and ABFE (Absolute Binding Free Energy) [1]. These methods have their benefits and drawbacks, and a common limitation is the trade–off between physical accuracy and speed. MM–PBSA, RBFE and ABFE are methods based on Molecular Dynamics. Unlike for pose prediction determinations, these are used to determine the binding free energy for a given ligand pose. Docking is frequently used for binding pose prediction, and as in the work by *Jiang et al.* 2002, they utilized molecular docking to identify a possible non–peptide HIV fusion inhibitor [2].

Molecular Dynamics, a method commonly used to determine the binding free energy for a bound ligand, has also been used to simulate the binding event [3, 4]. In the work by *Buch et al.* 2011, they simulated how benzamidine found the way to a favorable binding site from a distance of 35 Å outside the protein, as well as its binding pose. Molecular Dynamics is a more physically motivated method in comparison with molecular docking. However, it is vastly more computational and time expensive. Therefore, it is not a common technique to be used for this type of application as it would need a special–purpose machine such as *Anton* used by the Shaw group to be able to perform the long simulations needed [5].

In the Reconnaissance metadynamics method, the algorithms are constructed in order to overcome the drawbacks of Molecular Dynamics. The simulations is made more extensive in its search for the free–energy minima of the system. In the work by *Söderhjelm et al.* 2011, Reconnaissance metadynamics was used to explore the surface of a protein in order to find favorable binding pockets and poses for the ligand benzamidine [6].

The D3R Grand Challenge 2016 is a competition to make the best binding–pose and binding affinity prediction for a given protein and set of ligands [7]. This is the second *Grand Challenge* conducted by D3R. The first stage involved to determine the binding pose for 36 different ligands to the farnesoid X receptor, FXR, with computational methods.

FXR is a ligand–activated transcription factor, attributed to many bodily functions, e.g. regulation and maintainance of bile acid synthesis, reduction of plasma cholesterol and triglycerides, glucose homeostasis and improvement of insulin sensitivity [8]. Development of FXR agonists are of great therapheutical interest for many different afflictions, such as dyslipidemia and diabetes.

In this thesis a method based on molecular docking, Molecular Dynamics and Reconnaissance metadynamics were tested on a set of agonists, provided by the D3R Grand Challenge 2016, in order to predict their binding poses when bound to the FXR receptor. Further, the pose prediction method described in this thesis is designed in such a fashion as to be able to accuratly predict the binding pose when the knowledge behind the mechanism is limited and/or where no experimental data is available.

1.1 Aim of thesis

The focus of the thesis is to predict the poses of a set of ligands from four chemical series (benzimidazoles, isoxazoles, spiros and sulfonamides) bound to the human farnesoid X receptor target, FXR, without bias from experimental data. The data set is taken from the *D3R Grand Challenge 2016* where the data were provided by Roche Pharmaceuticals [7].

In this work there will be *three* levels of computational pose predictions: Docking, Molecular Dynamics and Reconnaissance metadynamics. These results will help to determine if each step increases the pose prediction accuracy of the method. The RMSD between the predicted poses and the experimental crystallographic data, supplied of the $D3R$ Competition, will be used to determine the accuracy in the predictions. Also, further analysis of the behavior of the pose prediction methods will be studied.

List of abbreviations used throughout in the thesis.

MD	Molecular Mechanics
RMD	Reconnaissance metadynamics
RMSD	Root-mean-square deviation
FXR	Farnesoid X receptor
MM	Molecular Mechanics
QM	Quantum Mechanics
PBC	Periodic Boundary Conditions
GAFF	Generalized Amber Force Field
CVs	Collective Variables
GA	Genetic Algorithms
MC	Monte Carlo methods

2 Background

2.1 Molecular Docking

Molecular Docking is a rapid way to predict the bound conformation of a noncovalent intermolecular complex between two or more molecules [9], commonly between a macromolecule, *receptor*, and a small molecule, *ligand*. However, in comparison with other methods, e.g. Molecular Dynamics, docking trades off the physical accuracy for speed [1]. Therefore, it is a good tool to rapidly filter out molecules that bind from the ones that do not as well as predict the bound ligand pose.

In docking the receptor is often considered rigid and the ligand flexible, although methods that take the flexibility of the side chains in the receptor into account do exist (flexible docking). Although the flexibility of the protein is important for ligand binding, methods that considers the protein flexibility are still in their infancy. For methods where both the ligand and receptor are flexible, the docking simulation may take a couple of days to be performed [10]. This in contrast to the minutes of computational time used when only the ligand is flexible. Therefore, in order to be feasible, the simulation time of flexible receptor and ligand must be decreased. Until then, the flexible ligand and rigid receptor approximation will be more commonly used than flexible docking.

Docking can be seen as two main methods; conformational search and scoring. The conformational search attempts to predict the pose of the ligand when bound to the binding site, whereas the scoring function approximates the binding affinity for that pose. Furthermore, the scoring functions rank the predicted poses to find the best prediction. Random or stochastic algorithms such as Monte Carlo methods, MC, and Genetic Algorithms, GA, are two methods of many which perform conformational search. Monte Carlo methods randomly generate conformational changes in the ligand. The MC alterations on the ligand pose are accepted on criteria based on the Boltzmann probability function. Genetic Algorithms, however, are based on ideas from the theory of biological evolution. In GA the simulation starts with a population of ligand conformations with a defined set of state variables for e.g. the orientation, translation and conformation of the ligand with respect to the protein. To the population of conformations, random generations of mutations are performed until a population that optimizes the scoring function is found [10].

The scoring function is commonly based on Molecular Mechanics calculations in order to attempt to approximate the binding free energy for the conformation. These force field based scoring functions are commonly used to calculate the binding free energy, more on force fields in section 2.2.2. Other functions such as strictly empirical or knowledge–based scoring functions also exist, however these are highly dependent on the availability of information concerning the given ligand or similar molecules. The force field based scoring function also has its limitations, such as the absence of solvation and entropic terms.

One docking software based on the flexible ligand and rigid receptor approximation is AutoDock Vina. AutoDock Vina utilizes both MC, GA and other methods to predict the pose. The stochastic Iterated Local Search global optimizer is also used to predict and speed up the sampling of ligand conformations in AutoDock Vina. With this algorithm, iterations consisting

of a random mutation and local optimization of the ligand conformation are performed and scored. For the scoring in AutoDock Vina, the iteration is accepted or rejected according to the Metropolis criterion, i.e. it is accepted if the conformation has a lower free energy than the previous iteration (based on force–field–based scoring function calculations) [11]. The conformations are ranked based on their score and the one with the highest binding affinity is ranked as the top pose.

The major drawbacks with docking is that both the conformational search and the scoring functions have their limitations and inaccuracies. As the conformational search does not commonly take the full flexibility of both ligand and protein into account, the pose prediction could be inaccurate. Moreover, there are no explicit water molecules interacting with the ligand or the protein. This could lead to misrepresentation of the bound ligand conformation, as well as to a decrease in accuracy of the scoring functions.

2.2 Molecular Dynamics

Molecular Mechanics, MM, and Quantum Mechanics, QM, are two ways of describing the potential energy, V, for the system. However MM, unlike QM, ignores the electrons movements and electronic state. It calculates the energy in the system as a function of the nuclear positions only. The Born–Oppenheimer approximation is one of many assumptions which allows MM to function this way. Still MM can in some cases provide predictions as good as QM at a fraction of computer time. However, since it approximates and disregards the electron distribution, it cannot provide information regarding scenarios depending on the electron distribution such as chemical reactions, proton transfer etc.

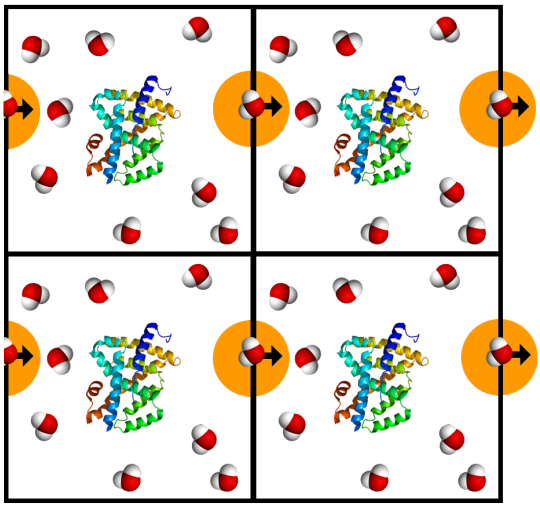


Figure 1: An example of a PBCs, depicting how a water molecule leave and return on the opposite side of the boundary of the cell. Furthermore, the replication of the cell, a FXR receptor and a few water molecules are shown in the figure.

Molecular Dynamics, MD, is an N-body computer simulation method where numerical integrations of Newton's laws of motion are applied to the potential energy, V, and generates successive configurations of a system. Solving these Newtonian equations, for each time step, results in a trajectory which describes the position and velocities for every atom in the system at a given time. Thus the trajectory contains the information for the three space coordinates (x, y, z) and *three momentum vectors* which defines the 6N-dimensional phase space, with N amount of particles, for the system. Molecular Dynamics simulations are usually based on empirical MM calculations and what are commonly referred to as "Force Fields", see section 2.2.2.
In contrast to docking, MD is more physical accurate as it takes explicit water molecules and more accurate force fields into account. However, here MD trades off speed for physical accuracy.

2.2.1 Initial Velocity and Periodic Boundary Conditions

In order to set up a MD simulation an initial configuration for the system is needed. Therefore, all the coordinates (x, y, z) and initial velocities for all atoms in the system must be specified. The starting velocities are usually taken from the Maxwell–Boltzmann distribution. The initial starting positions for the atoms in the system is however much more important than the starting velocities.

Periodic boundary conditions enable the particles in the simulation to experience bulk fluid–like forces even with a smaller amount of particles. This is achieved by using a cell of a specific shape e.g. cubic or octahedral, which is replicated in all directions. The cells are all images of each other and if one particle leaves the cell through one boundary, it is replaced by an identical particle entering from the opposite direction to keep the amount of particles constant in the system. In Figure 1 the highlighted water molecule is shown to leave and return to the opposite side of the cell. The benefit of using PBCs is that the solute of interest never experiences solvent free surface effects and thus, as stated above, the simulation is more like a bulk fluid. However, it is important that the dimension of the cell is large enough so that the solutes do not affect itself through the boundaries.

2.2.2 Force Fields

The Force Field, FF, makes up an empirical mathematical representation of the intra– and intermolecular forces within the system, thus describing the potential energy as a function of the coordinates of all atoms. There are many different force field models but a way to describe a simple force field is to separate the potential energy contributions into four components: *bond stretching*, *angle bending*, *torsional terms* and *non-bonded interactions*. Each component contributes to the potential energy as can be seen in the following equation for the AMBER FF [9, 12]:

$$V(\mathbf{r}^N) = \sum_{Bonds} \frac{k_i}{2}(l_i - l_{i,0})^2 + \sum_{Angles} \frac{k_i}{2}(\theta_i - \theta_{i,0})^2 + \sum_{Torsions} \frac{V_n}{2}\left(1 + \cos\left(n\omega - \gamma\right)\right) + \underbrace{\sum_{i=1}^{N}\sum_{j=i+1}^{N}\left(4\epsilon_{ij}\left[(\frac{\sigma_{ij}}{r_{ij}})^{12} - (\frac{\sigma_{ij}}{r_{ij}})^{6}\right] + \left[\frac{q_i q_j}{4\pi\epsilon_0 r_{ij}}\right]\right)}_{\text{Non-bonded interactions}} \quad (1)$$

where $V(\mathbf{r}^N)$ is the potential energy as a function of the postions, $\mathbf{r}$, of N particles. The right-hand side of the equation, eq. 1, will be described more in–depth below together with the parameters, variables and constants.

The bond stretching describes the energy of the harmonic potential between two bonded atoms as the bond length, l, deviates from the reference length, l_0. By using Hooke's laws formula where the difference in bond lengths and the Hookean spring force constant, k, describe the potential energy contribution, $v(l)$, as follows:

$$v(l) = \frac{k}{2}(l - l_0)^2 \quad (2)$$

A common approximation is to keep the bond lengths constant throughout MD simulations.

Angle bending is similar to bond stretching as it is also described by Hooke's Law and the deviation in bond angle, θ, from the reference angle, θ_0, changes the potential energy for the bond, according to

$$v(\theta) = \frac{k}{2}(\theta - \theta_0)^2 \tag{3}$$

Less energy is required to distort the angle than to stretch or compress the length. Still, both are regarded as "hard" degrees of freedom, i.e. they both require a large amount of energy in order to change to a value that differs significantly from the reference value. Most variations in molecular structure are therefore connected to the interaction between the torsional and non-bonded contributions which requires less energy for conformational changes.

The torsional terms depict the torsion potential, v, as the bond rotates in a dihedral (A-B-C-D, where the capital letters are the atoms in the dihedral) between the two central atoms (B and C).

ω is the torsion angle, V_n is the "barrier" height and

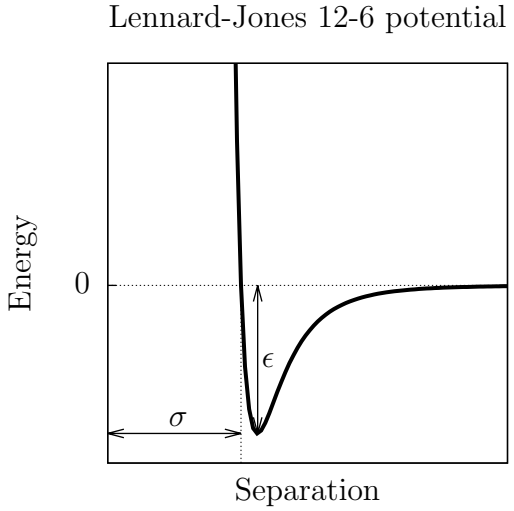


Figure 2: The Lennard–Jones 12-6 potential. Where σ is the collision diameter and ϵ is the well depth between the x-axis and the Lennard–Jones potential function.

$$v(\omega) = \sum_{n=0}^{N} \frac{V_n}{2}\left(1 + \cos\left(n\omega - \gamma\right)\right) \tag{4}$$

The non-bonded interactions for the AMBER force field are the contributions from electrostatic and van der Vaals interactions. The electrostatic interaction between different parts of the molecule is calculated with Coulomb's Law, eq. 5 below, where the sum of interactions between point charges are calculated.

$$V = \sum_{i=1}^{N_A}\sum_{j=1}^{N_B} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \tag{5}$$

N_A and N_B are the numbers of point charges in the two molecules while q_i and q_j are the values of the point charges i and j. r_{ij} is the distance between the charges and ϵ_0 the electric permittivity in vaccuum.

The van der Vaals interactions are commonly described using the Lennard–Jones 12-6 potential function, eq. 6.

$$\sum_{i=1}^{N}\sum_{j=i+1}^{N} 4\epsilon_{ij}\left[(\frac{\sigma_{ij}}{r_{ij}})^{12} - (\frac{\sigma_{ij}}{r_{ij}})^{6}\right] \tag{6}$$

σ is the collision diameter, ϵ is the well depth and r_{ij} is the distance between the atoms. The

Lennard–Jones potential describes the attraction, r_{ij}^{-6}, and the repulsion part, r_{ij}^{-12}, between two atoms. As can be seen in Figure 2, the Lennard–Jones potential has an energy minimum where the attraction balances the repulsion.

2.3 Adding Bias to Molecular Dynamics Simulations

One of the drawbacks with Molecular Dynamics is that the simulation is very time consuming. In principle MD has the possibility to visit every energy minimum with enough time, but high energy barriers prevent the exploration during the time frame for the simulation. The energy landscape is therefore not fully explored in practical situations and only one or a few metastable states are visited during the simulation. A way to circumvent the time–scale problem with MD is to add bias to the simulation in such a way that the system is driven out from the current local minimum in order to discover more of the energy landscape. One method to do that is Metadynamics.

2.3.1 Metadynamics

By adding a bias potential to the MD simulation for some selected degrees of freedom, the Metadynamics method allows for enhanced sampling of the free–energy surface, FES, and for simulating rare events e.g. conformational changes. The degrees of freedom that the bias potentials act on are commonly called *collective variables*, CVs. A bias potential is constructed as a sum of Gaussians added to the potential energy for the system [13]. This addition, in the CV-space of the trajectory, the potential discourages the system to sample states previously visited.

For a set of d CVs, $\mathrm{S}_\alpha(\mathbf{R})$, $\alpha \in [1,\, d]$ and where $\mathbf{R}$ is the set of microscopic coordinates of the system, the bias potential can then be written as the following equation, eq. 7, at time t.

$$V_G(\mathbf{S}, t) = \omega \sum_{t'=\tau_G,\ 2\tau_G,\ldots}^{t'<t} \exp\left(-\sum_{\alpha=1}^{d} \frac{(S_\alpha - S_\alpha(\mathbf{R}(t')))^2}{2\delta S_\alpha^2}\right) \tag{7}$$

V_G is the metadynamics potential, ω is the Gaussian height, $\delta\mathbf{S}_\alpha$ is the width of the Gaussian for CV α and τ_G is the deposition frequency of Gaussians.

The CVs can be any function of R, e.g. a distance or an angle, as follows:

$$\mathbf{S}(\mathbf{R}) = (S_1(\mathbf{R}), ..., S_d(\mathbf{R})) \tag{8}$$

If the simulation time is sufficiently long, eq. 7 can be estimated to be approximately equal to the free–energy surface with opposite sign, see eq. 9 [14].

$$\lim_{t\to\infty} V_G(\mathbf{S}, t) = -F(\mathbf{S}) \tag{9}$$

One of the drawbacks with metadynamics is that only a few significant CVs can be used during the simulations, as the simulation time will increase exponentially with the number of CVs [15]. Selecting these few CVs can be difficult if the knowledge of the mechanism is limited [6].

2.3.2 Reconnaissance metadynamics

Reconnaissance metadynamics, RMD, is a machine–learning approach to biased MD simulations, where the algorithms tune the applied bias using clustering information gathered at a set interval. Compared to metadynamics, RMD can be more efficient with a larger number of collective variables. This is beneficial when the mechanism is not well known, as one can use a larger set of CVs without it being too computationally expensive.

The bias potential in RMD can be seen as a patchwork of basins, where each basin corresponds to a local free–energy minimum [15].

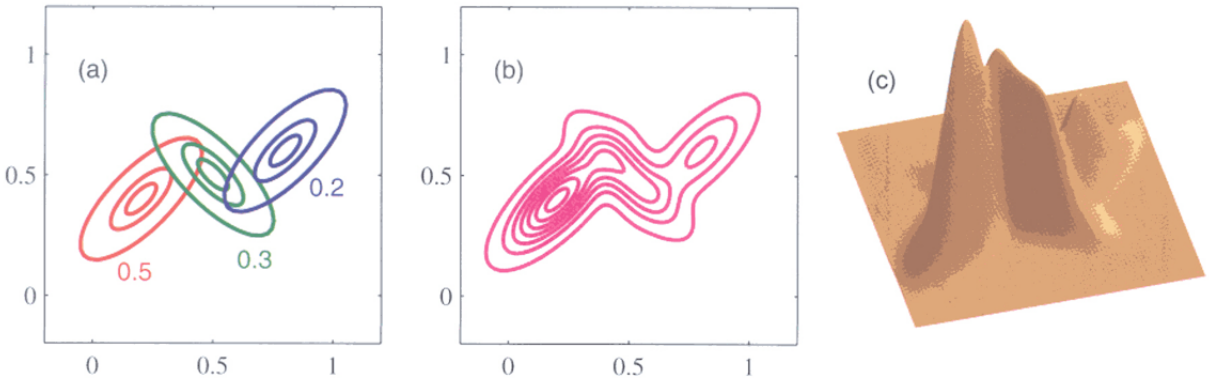


Figure 3: The individual basins are depicted in 3A and how they together describe the CV landscape in 3B. Furthermore, the amount of bias added to the basins can be seen in 3C.

Consider a system with two CVs: $\mathbf{S}_1$ and $\mathbf{S}_2$, as illustrated in Figure 3 and equations 10–11. During the simulation, configurations of the system are sampled and these configurations are then clustered into basins. Covariance matrices describe the dependence between the two CVs, i.e. the shape and size of basins. The covariance matrix $\mathbf{C}$, see eq. 10, is used to project a position in the 2–dimensional CV representation of a basin down to a scalar radial collective coordinate, $r(\mathbf{S})$, see eq. 11.

$$\mathbf{C}_{\mathbf{S}_1,\mathbf{S}_2} = \langle (\mathbf{S}_1 - \boldsymbol{\mu}_1)(\mathbf{S}_2 - \boldsymbol{\mu}_2) \rangle \, ; \quad \begin{matrix} \boldsymbol{\mu}_1 = \langle \mathbf{S}_1 \rangle \\ \boldsymbol{\mu}_2 = \langle \mathbf{S}_2 \rangle \end{matrix} \tag{10}$$

$$r(\mathbf{S})^2 = (\mathbf{S}_1 - \boldsymbol{\mu}_1)^T \mathbf{C}^{-1} (\mathbf{S}_1 - \boldsymbol{\mu}_1) \tag{11}$$

This projection from a higher to a lower dimension is the reason behind why RMD can be used with a larger set of CVs. A bias potential composed of a sum of Gaussian functions is added along $r(\mathbf{S})$ to discourage the system from visiting the same state. All points $\mathbf{S}$ at a certain scaled distance $r(\mathbf{S}) = R$ form an ellipse in the two-dimensional CV space (or, more generally, a hyperellipsoid in a higher–dimensional space). Thus, the added bias in three dimensions can be seen as an onion layer. Therefore, the term "onions" is commonly used to describe the layers of added bias potential in RMD. Unlike metadynamics, the amount of added bias in RMD cannot directly be equated to the free–energy. However, the depth of the basin is connected to the kinetic stability for the configuration, and the kinetic stability is often (but not always) correlated with the thermodynamic stability (free energy).

3 Material and Methods

This section will go through the *three* major steps *Docking*, *Molecular Dynamics* and *Reconn–aissance metadynamics* which were used for the pose predictions, the preparations of input files and the respective settings.

3.1 Data

One of the difficulties with the D3R Challenge 2016 was it being a "blind challenge". The information regarding the experimentally determined crystal structure of the ligands bound to FXR were initially withheld, to be released after a deadline for submission of the pose predictions. Only the SMILES–strings for the ligands and an apoprotein crystal structure of FXR were given at the start of the challenge. With the given files, and any already published crystal structures, one of the goals was to predict the binding pose for the ligands to the receptor. In our study, all crystal structure entries of FXR with ligands bound (as of the 5^{th} of September 2016) in the Protein Data Bank, PDB, were downloaded and prepared alongside the apoprotein provided by the *D3R Grand Challenge 2016* (see Table 2 in the appendix for a list of the receptors PDB IDs). These entries were used to account for any conformational variations in the protein.

3.2 Docking

In order to get a starting point for the Molecular Dynamics simulations a structure with ligand bound to the receptor is needed. With the molecular docking tool AutoDock Vina, the ligand is "docked" into the active site and the lowest free energy dock conformation is calculated [11]. The output is the predicted pose that will be used as a starting point for MD. However, in order to perform the Molecular Docking, both the receptors and the ligands first need to be prepared.

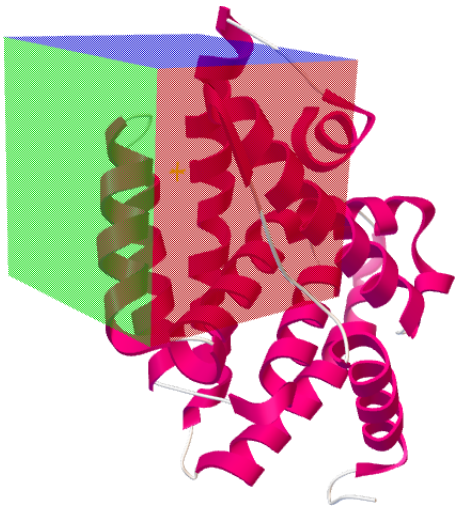


Figure 4: The construction of the grid box, in AutoDock Tools, which covers the active site where the ligand will be docked to. The *green*, *red* and *blue* color represents the box dimensions in x, y and z directions. The FXR receptor seen, in pink, is the one with PDB ID 1OSH.

3.2.1 Preparation of the Receptors

For all crystal structures, provided by D3R or downloaded from PDB, Chain A was isolated from the pdb–files by removing all water molecules as well as any redundant protein chains. The proteins were protonated with the web based software H++ which predicted the protonation state for the receptor at pH 7.4 [16].

The docking software uses a grid box wherein it tries to dock the ligand to the receptor. This box could be covering the entire receptor (blind docking), if the active site is not known or determined, however a more specific grid box covering only the active site would be a more optimized way to perform the docking. Thus, the downloaded crystal structures with bound ligands were used as a reference for the gridbox, as they all indicate that there is only one active site for these types of ligands, present in the receptor. The

AutoDock Tools were used to create a grid box for docking in AutoDock Vina. The grid box was defined with a grid spacing of 1.0 Å and the size for x, y and z at 15 with the respective center coordinates for the box (x, y, z) for all of the crystal structures. A representation of the grid box can be seen in Figure 4.

3.2.2 Preparation of the Ligands

First, hydrogen atoms were added at pH 7.4 to the 2-dimensional molecular structures, provided by Roche in the D3R Grand Challenge, with the software Open Babel's `-h` option [17]. The 36 ligands were then geometry–optimized in vaccum through a conformational search with Open Babels `obconformer` tool and the MMFF94 force field. The conformer tool generates a set of random conformations using a Monte Carlo search and the best out of the set is further optimized in the following optimization steps. After the optimization a 3-dimensional ligand structure is obtained for each of the ligands in the set.

The tool `antechamber` was utilized to get an AMBER prep-file for each ligand to be used further on in `tleap` with the simple generalized Amber FF, GAFF, [18, 19, 20]. All the ligand pdb-files were converted to pdbqt-files with AutoDock Raccoon [21].

3.2.3 Docking Run

After the pdbqt–files for both receptors and ligands were prepared, they were docked with AutoDock Vina. All ligands were docked with the crystal structures from the PDB and the apoprotein. The docking was performed with the setting 8 for exhaustiveness, i.e. the time spent looking for a global minimum of the scoring function in Vina. Increasing the exhaustiveness would increase the probability of finding a global minimum but the computational time would increase. However, as the docking is heuristic it may not be beneficial to increase the exhaustiveness. The AutoDock Vina will output a number of pose binding predictions and calculate a binding affinity prediction in $kcal/mol$. The pose with the most negative affinity will be the top pose and the following poses are ordered in descending order of affinity.

3.3 Molecular Dynamics

3.3.1 Preparation of the Complexes for MD

The top docking pose predictions, for all of the ligands, i.e. the prediction with the highest binding affinity, were used as the starting point for the Molecular Dynamics simulations. The output file from AutoDock Vina contains all the predicted poses and therefore the AutoDock Vina tool `Vina Split` was used to obtain the top pose from the output file. Open Babel was used to convert the top pose pdbqt-file back to pdb with the `-xrp` option.

The pdb–files of the top pose predictions, containing the conformational data from AutoDock Vina's docking, were added to the end of the receptor's pdb-file to form a file with the predicted intermolecular complex. AMBER's `tleap` was used to create the AMBER topology and coordinate files to be used in GROMACS, together with the prepared AMBER prep-files constructed earlier, see 3.2.2, as well as the FF14SB force field [22]. Since `tleap` did not have parameters for some of the dihedrals for GAFF, the parameter data set used in `tleap` had to be expanded with the parametrization information for the missing dihedrals. AmberTool's `parmchk2` was used to construct the missing parameters from parameters for similar atom types.

The tool `acpype` was used to convert the ligand–protein complex files prepared previously into GROMACS topology and coordinate files. With the GROMACS' `editconf` tool the complex was placed in a truncated octahedron box, with the distance, `-d`, from the complex to the box boundary of 8.0 Å, to serve as a container to solvate the complex in. The GROMACS `genbox` tool solvated the previously prepared octahedron box and the complex with TIP3P waters. These steps were performed for each protein–ligand complex.

An energy minimization step, using the GROMACS tools `grompp` and `mdrun`, was performed for each complex by using a steepest descent integrator with verlet cut-off scheme for 200 steps.

3.3.2 NPT Equilibration with Restraints

A NPT equilibration (a fixed amount of atoms, N, a set pressure, P, and temperature, T) with positional restraints on the C_α backbone of the receptor and the heavy atoms of the ligand, of nanosecond was performed to allow the water to relax around the complex, without distorting the conformation of the complex. The Coulomb type was Particle mesh Ewald with a cut–off at 10 Å, the pressure coupling was Berendsen isotropic at 1 bar. Integrator for the `mdrun` was a leap–frog integrator with 500 000 steps and a time step, dt, of 2 femtoseconds. The temperature was set to 310 K to mimic the temperature condition in the human body. All calculations were run on the HPC2N Abisko cluster.

3.3.3 NPT Molecular Dynamics Run

The production MD simulation was performed with the same settings as the previous NPT equilibration (3.3.2), however without the restraints on the C_α backbone of the receptor and heavy atoms of the ligand. MD is used to see whether the predicted pose is stable. If the ligand leaves the predicted pose, it is an indication that the pose is not stable. Therefore, after each nanosecond of MD simulation, a RMSD calculation was done to see whether the ligand position deviated too much from the starting position. If the RMSD of the ligand in relation to the C_α backbone of the receptor was >2.5 Å, the simulation was terminated. If the RMSD constantly is <2.5 Å, the simulations continue for 50 nanoseconds.

3.4 Reconnaissance metadynamics

Because of time limitation only a few ligands from the MD simulations were extended with Reconnaissance metadynamics simulations. First, all the rotable bonds in the ligands were determined manually and the corresponding dihedrals added to the *collective variables*, CVs, list in the PLUMED plugins input file [23]. The RMD simulations were performed with the same settings as the previous NPT MD simulations (3.3.3), although now with added bias to the dihedral angles of the ligands. The onion deposit stride was 1 ps, the onion width 1.5 and the onion height 1.0 kJ/mol. Basin tolerance was set to 0.2, the basin expand parameter 0.3 and the basin initial size was 1.5. The RMD internal

Table 1: The side-chain residues that were chosen to have additional dihedral rotational bias in the Reconnaissance metadynamics simulations.

Residue	
LEU	44
MET	47
HIE	51
MET	85
SER	89
LEU	105
ILE	109

clustering frequency was 100 picoseconds and 1000 points were collected in this period. Reconnaissance simulations were run for 30 nanoseconds and the results were interpreted by clustering.

For some ligands, simulations with added CVs to the dihedral angle between the C_α–C_β, for a few of the active site's side chains (see Table 1) in addition to the ligand CVs were run.

3.5 Analysis of Results and Data

3.5.1 Method for RMSD Analysis

To analyze the top pose predictions from docking and the Molecular Dynamics simulations, as well as the reconnaissance metadynamics simulations, the predicted ligand conformations were first aligned to the apoprotein and then compared with the D3R experimental ligand crystal structures (chain A). For the Molecular Docking, all poses for a given ligand were ranked by their binding affinity (high to low), and any duplicate poses were removed. The duplicates were removed by comparing the RMSD between the poses with a cut–off at 2 Å, which would allow for only truly different poses to be ranked. The top five poses for each ligand were then compared with the experimental data by calculating the RMSD between the poses.

The RMSD calculations were performed by the software `rmsd.py` which is bundled with the Schrödinger 2016-3 Maestro 10.7 package [24]. `rmsd.py` first converted the pdb–files into Maestro's mae–files and then calculated the RMSD between the predicted poses and the crystal structure pose, with an "in–place" RMSD calculation. The advantage with using Maestro, is that it takes into account the symmetry of the molecules.

3.5.2 Cluster Data Analysis

Cluster analysis were performed with GROMACS' `g_cluster` using the GROMOS algorithm after the trajectories were sparsed with `trjcat`, `-dt` 20 (i.e. only every 20 picosecond were used for the analysis in order to reduce the computational time). The RMSD was calculated for the heavy atoms of the ligand after alignment of the C_α of the protein [25]. All structures, i.e. the conformations at the each time step, under the chosen RMSD cut–off were clustered together and the structure within the cluster with the most neighbors, i.e. most similar structures, was set as the preliminary cluster center. Then the cluster, and its structures, were removed from the structure pool and a new cluster was determined and analyzed. By removing the structures as the cluster analysis progresses, the use of the same structure within several clusters was avoided. The structure corresponding to the center of a cluster was determined and used as the predicted binding pose for the ligand. The cluster sizes were then ranked against each other and the cluster containing the most structures became the top–pose cluster. The reasoning was that the largest cluster's corresponding structure is probably one of the more energy favorable poses, as it is the most sampled pose during the MD simulation, and thus probably the most accurate prediction with the lowest free energy. The RMSD cut–off used was set to 2.0 Å, and if there were several small clusters only the ones containing $\geq$10 structures were considered as significant clusters.

For the first set of ligands studied with RMD, namely ligands 22, 27, and 32, the poses submitted in the D3R challenge were obtained in the following way. First, the MD and RMD trajectories were merged and clustered together. To confirm the stability of these poses, the

cluster centers of the top 4-5 clusters (4 for ligand 22 and 5 for ligands 27 and 32) were used as starting points for additional MD simulations, each of length 20 ns, except the cluster center corresponding to the MD pose, for which no extra simulation was run as it was already confirmed to be stable in MD. Similarly to the original MD simulations, these simulations were terminated if the ligand pose deviated significantly from the starting pose. Finally, each extra MD simulation was clustered separately and the top cluster center was used as the submitted pose in the RMD submission. In addition, the MD pose was also included in the RMD submission to conform with the picture that in the combined clustering, the MD pose was always dominant, so it should be included even if no refinement was needed. The ranking of the poses was determined by a manual procedure taking into account:

- The size of the clusters in the combined clustering (which is related to the kinetic stability of the pose, as discussed below).
- The stability of the pose in the extra MD simulation (i.e. how long it stayed near the starting pose)
- The decision to avoid submitting the MD pose as the top-ranked pose in the RMD submission (instead the MD pose was submitted as rank 4, 5, and 2, respectively, for ligands 22, 27, and 32).

4 Results and Discussion

4.1 Docking

In the information regarding the challenge, it was stated that the protein has flexible α–helices adjacent to the active site. Therefore, in order to include all structural variation in protein structure, PDB crystal structures with FXR bound to different ligands were used.

In the molecular docking, each ligand was docked to all the crystal structures and for each of these complexes the binding affinity (in *kcal/mol*) was calculated with AutoDock Vina. The top predicted pose, for each of the ligand–receptor combination, can be seen in Table 2. The affinities which are highlighted in Table 2, are the ligand–protein complexes which had the highest binding affinities according to Vina's prediction and scoring functions.

Receptor 3OMM had the most instances of highest docking affinities (15/36), whereas the apoprotein had none. Interestingly, the apoprotein had the overall poorest docking affinities in comparison with the other FXR crystal structures. 3OMM also had the highest affinity for 14 out of the 21 benzimidazole ligands. In the cases where it did not have the highest affinity, the values of 3OMM were still close to the receptors with the highest predicted affinities. This is perhaps explained by the fact that in the PDB entry for 3OMM, it was crystallized with a benzimidazole containing molecule. Therefore, the protein conformation would be configured towards these types of ligands which could explain the high bidning affinities from docking.

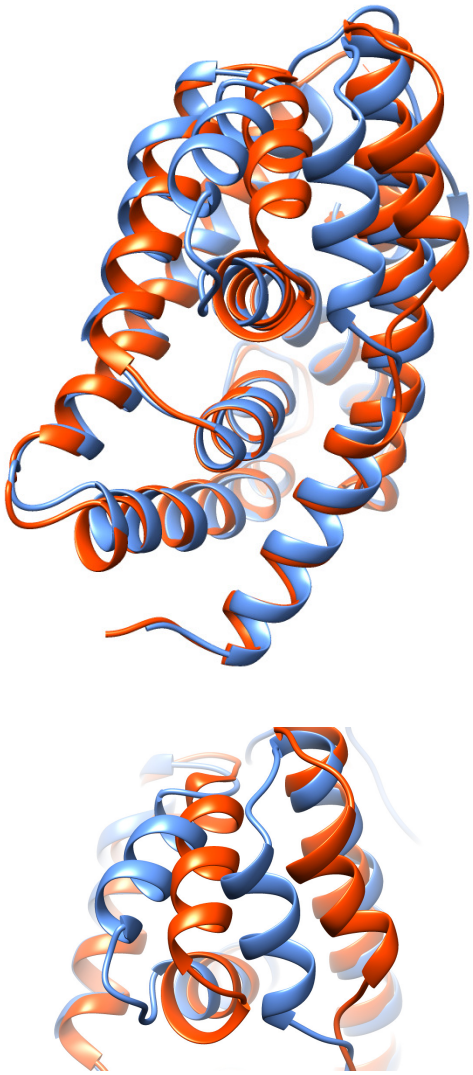


Figure 5: In the top figure the 3OMM (blue) is aligned with the apoprotein (orange). The bottom figure, is a zoomed in version of the above and shows more clearly the difference in positions for the two α–helices.

By comparing the apoprotein with the 3OMM structure, any conformation differences could perhaps explain the binding affinity results. When aligning the structures, there is a clear difference in the position for two α–helices adjacent to the binding site, see Figure 5. This was a trend for all of the receptors (3OMM, 3OMK, 3OLF, 3FLI and 3OOF) that the docking were able to predict a ligand pose correctly with, see Figure 7. All the other receptors were more similar to the apo in terms of the α–helices positions near the active site.

For all three ligands (10–12) from the spiros chemical series, the 1OSH had the highest binding affinity. These ligands were also the only ones that 1OSH had the highest affinity with. However, none of the predicted poses were accurate and the lowest RMSD for these ligands was 2.76 Å, see Table 3. When comparing the experimentally determined protein comformation (for ligand 10) with the 1OSH receptor a difference can be seen in the position for the α–helices, see Figure 6. For one of the helices there is a small shift in position and the other "helix" is not even in a helix conformation. The structure for 3L1B also had this lack of defined helix and both did

Table 2: The docking affinities, from AutoDock Vina, for all ligand and receptor combinations. The values are from AutoDock Vina's top pose predictions in *kcal/mol*. The names of the receptors are their corresponding PDB ID code and "Apo" is the apoprotein provided in the data set from the *D3R Grand Challenge 2016*. Highlighted affinities (in **bold**) are the ligand–protein complex predictions that were chosen to be used as starting points for further molecular dynamics simulations.

	Receptors																		
Ligand	1OSH	3FLI	3FXV	3L1B	3OKH	3OLF	3OMK	3OMM	3OOF	4QE6	4QE8	3DCT	3DCU	3HC5	3HC6	3P88	3RUT	3RUU	Apo
1	-8.9	-8.4	-7.9	-8.0	-8.0	-9.1	**-9.7**	-8.8	-8.9	-8.7	-6.0	-9.1	-9.2	-9.0	-8.7	-8.7	-8.7	-8.5	-4.3
2	-8.6	-7.6	-6.6	-8.1	-10.4	-9.6	-9.7	**-10.5**	-8.7	-10.3	-5.3	-9.2	-9.6	-9.8	-9.3	-9.6	-10.1	-9.3	-4.2
3	-10.0	-9.8	-8.4	-9.0	-8.5	-8.4	-8.4	-9.9	-9.4	-9.6	-8.8	-9.6	-9.9	-9.9	-9.4	**-10.0**	-10.0	-9.6	-7.7
4	-9.4	-10.2	-8.7	-9.9	-9.9	-9.7	-10.3	-10.1	-10.5	-7.6	-7.6	-10.5	-10.4	-9.4	-7.7	**-11.2**	-10.9	-8.8	-5.4
5	-8.7	-10.5	-7.7	-9.5	-9.1	-10.0	-9.0	**-10.5**	-9.8	-9.6	-9.1	-8.8	-9.1	-9.4	-5.8	-8.8	-9.2	-8.3	-6.2
6	-7.8	-9.0	-4.8	-9.2	-11.9	-11.8	**-12.2**	-12.1	-12.0	-7.7	-6.7	-7.2	-8.3	-7.5	-5.1	-7.8	-8.6	-6.1	-2.0
7	-9.4	-8.6	-6.6	-10.7	-10.5	-12.5	-12.6	**-12.8**	-12.8	-9.7	-7.1	-7.9	-8.8	-8.2	-7.3	-9.0	-9.5	-6.3	-1.7
8	-7.2	-8.0	-4.5	-8.5	-10.1	-10.4	-10.7	**-10.9**	-10.8	-9.2	-5.7	-7.5	-8.6	-8.2	-6.9	-8.3	-9.1	-7.2	-3.0
9	-7.1	-6.6	-5.6	-9.2	-9.8	-12.3	-12.9	**-12.9**	-12.3	-8.6	-7.1	-7.4	-9.3	-8.0	-6.8	-9.3	-8.6	-8.6	0.2
10	**-10.4**	-7.8	-6.9	-8.6	-9.0	-8.8	-7.5	-8.1	-8.3	-9.8	-6.8	-3.8	-7.4	-6.3	-4.7	-5.6	-7.7	-6.3	2.1
11	**-9.6**	-7.8	-4.2	-9.3	-8.5	-8.3	-7.2	-8.5	-9.1	-8.3	-8.0	-6.6	-8.1	-7.3	-6.0	-6.3	-8.4	-7.3	9.1
12	**-10.9**	-10.4	-8.5	-8.5	-9.9	-9.4	-9.2	-9.3	-7.9	-9.9	-6.3	-4.0	-8.7	-6.9	-5.8	-5.9	-8.4	-6.7	3.7
13	-9.4	-10.2	-6.8	-10.0	-10.4	**-12.5**	-12.2	-11.9	-9.8	-9.1	-7.6	-8.5	-8.7	-7.0	-9.0	-10.2	-8.4	-7.9	3.7
14	-7.1	-6.8	-4.1	-8.0	-11.0	**-12.0**	-12.0	-11.3	-9.2	-5.3	-2.2	-6.0	-6.2	-4.6	-6.1	-7.2	-7.9	-4.8	1.5
15	-10.6	-8.4	-7.7	-10.5	-9.4	-10.6	-10.1	-11.1	-10.0	-8.5	-10.2	-9.8	**-11.5**	-11.4	-10.7	-11.0	-11.0	-10.9	-0.3
16	-8.7	-9.1	-7.4	-8.5	-8.2	-8.0	-8.0	-8.3	-7.8	-8.3	-7.7	-7.1	**-9.5**	-8.7	-7.9	-9.5	-9.5	-8.9	-0.1
17	-5.5	**-11.2**	-6.6	-8.4	-5.8	-8.7	-7.8	-7.9	-7.3	-3.9	-4.9	-2.9	-7.1	-5.0	-3.2	-5.6	-7.7	-5.0	8.7
18	-6.7	-8.8	-7.2	-7.4	-10.8	-11.7	-10.9	-12.3	**-12.4**	-4.5	-3.4	-6.8	-8.5	-7.2	-8.2	-7.9	-9.4	-6.2	-1.9
19	-7.0	-9.2	-6.1	-8.7	-11.1	-11.6	-11.5	**-11.7**	-11.3	-6.5	-5.0	-7.4	-7.8	-9.0	-5.9	-9.2	-10.4	-7.3	-0.1
20	-8.3	-10.4	-6.0	-9.3	-11.9	-11.8	-12.3	**-12.8**	-11.9	-8.6	-7.2	-8.0	-10.2	-9.0	-6.7	-8.7	-9.8	-7.7	-3.3
21	-8.4	-10.1	-6.2	-8.0	-11.6	-11.5	**-12.2**	-12.1	-11.6	-8.0	-5.8	-8.4	-9.1	-9.7	-7.1	-10.0	-10.5	-8.9	-1.0
22	-9.0	-10.0	-6.8	-8.4	-9.3	-11.0	-10.9	**-11.3**	-10.6	-7.3	-5.1	-8.4	-8.4	-8.5	-7.3	-9.8	-9.9	-7.1	-1.2
23	-10.4	-11.4	-10.1	-10.4	-9.7	-11.2	-9.7	-11.0	-11.2	-9.6	-9.4	-10.4	-11.3	-11.4	**-11.7**	-11.5	-11.7	-11.0	-6.3
24	-8.6	-10.6	-6.1	-8.9	-11.6	-11.3	**-12.1**	-11.9	-11.7	-8.0	-5.3	-8.1	-9.7	-9.8	-6.4	-10.0	-10.0	-8.2	-1.6
25	-5.9	-9.4	-4.1	-8.5	-10.9	-13.1	-13.0	**-13.3**	-12.3	-8.0	-4.1	-7.7	-7.7	-7.3	-6	-11.3	-8.1	-5.6	0.6
26	-9.1	-6.4	-4.7	-8.4	-9.8	-10.9	**-11.5**	-11.4	-11.2	-9.0	-7.4	-5.3	-5.0	-6.4	-3.9	-8.9	-7.8	-5.6	0.4
27	-5.8	-8.5	-6.6	-9.8	-11.0	-13.7	-13.4	**-13.7**	-12.7	-9.4	-3.3	-5.7	-7.6	-7.1	-4.3	-5.8	-7.2	-6.1	1.7
28	-9.2	-8.8	-8.1	-9.6	-10.2	-12.5	-12.4	**-12.8**	-12.6	-8.8	-5.9	-8.5	-9.4	-8.0	-6.9	-9.0	-9.5	-8.1	-4.1
29	-8.0	-8.6	-7.8	-9.1	-9.9	-12.3	-12.2	**-12.9**	-12.4	-8.9	-7.1	-8.0	-8.8	-9.0	-7.1	-8.9	-9.8	-8.4	-4.4
30	-5.5	-8.6	-6.4	-7.4	-8.4	-10.5	**-10.6**	-10.4	-10.3	-5.6	-5.9	-6.4	-7.7	-7.0	-6.9	-9.2	-9.2	-5.4	4.4
31	-7.8	-9.1	-6.5	-7.8	-10.6	-11.8	-12.0	**-12.3**	-11.6	-7.1	-5.5	-5.7	-5.8	-6.3	-4.1	-6.0	-6.4	-4.5	1.3
32	-5.5	-9.5	-6.7	-9.2	-9.4	-11.0	-10.4	**-11.6**	-10.3	-9.2	-4.7	-9.1	-9.7	-9.1	-5.9	-9.6	-10.2	-7.8	0.1
33	-6.5	-5.5	-11.9	-8.3	-7.0	-7.7	-5.5	-8.7	-7.6	-10.4	-7.6	-12.2	**-12.9**	-12.2	-12.6	-12.5	-12.6	-12.4	-0.4
34	-8.6	-8.4	-8.2	-10.0	-9.5	**-11.2**	-10.5	-9.6	-7.6	-8.8	-7.9	-9.4	-9.2	-10.3	-9.9	-11.8	-11.4	-10.1	0.0
35	-7.7	-7.7	-6.2	-9.4	-10.6	-12.0	-12.4	-13.3	**-13.5**	-10.0	-4.4	-7.4	-7.6	-7.9	-5.2	-6.8	-8.2	-7.7	0.5
36	-6.3	-4.7	-5.8	-8.8	-7.6	-11.4	-10.9	**-12.6**	-11.8	-4.6	-4.7	-2.1	-6.1	-5.9	-5.5	-6.6	-7.5	-4.7	5.7

not perform well in their respective pose predictions.

For the isoxazoles, and the other miscellaneous ligands, the docking method was not able to predict any pose accurately.

By calculating the RMSD between the predicted docked poses and the crystal structures, the accuracy in predicting the poses for the method could be analyzed. The RMSD calculations show that only 19 out of 36 ligands were predicted quite well, see Table 2. The RMSD for these were $\leq$2 Å and 11 of the poses were also among the top ranked docking predictions. The best pose predicted, with an RMSD at 1.05 Å was for ligand 16 with 3FLI, see Figure 8. Also, the second best pose, at 1.27 Å can be seen in Figure, 9

Considering the flexibility of the α-helices, it was of interest to see whether the conformation of the receptor was more important for the outcome of the docking than previously thought. To test this, cross dockings between the ligands with their respective D3R experimental protein crystal structure were performed. The cross docking was able to predict the correct pose for 32 out of 36 ligands, see Table 4. The docking method thus appears to be functioning properly, at least for most cases when docking is performed with the correct structures.

Looking at the helices, there were cases where their positions were more similar to the

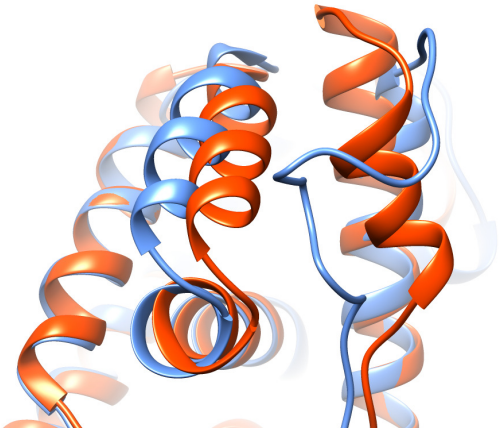


Figure 6: A focused view on the helices adjacent to the active binding site for 1OSH (blue ribbon) and the protein crystal structure (orange ribbon) for FXR bound to ligand 10 (ligand not shown).

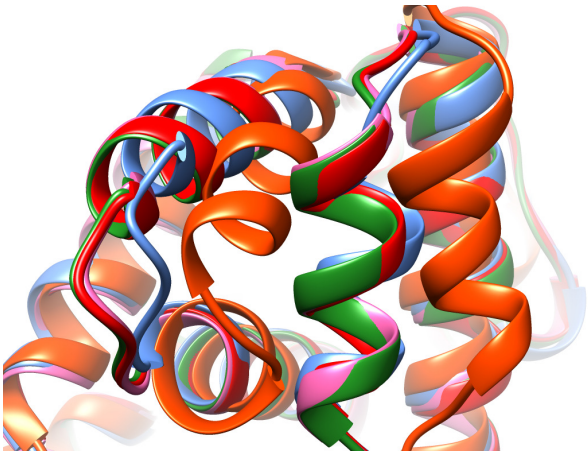


Figure 7: A focused view on the helices adjacent to the active binding site for: 3OMK (red), 3OLF (green), 3FLI (blue), 3OOF (pink) and the apoprotein structure (orange).

3OMM and cases where they were more similar to their positions in the apoprotein. For the benzimidazoles the positions of the helices were along the positions in 3OMM, with a RMSD of ∼1.0 Å when compared with the apoprotein. The isoaxoles, sulfonamides and the spiros however, were more similar to the apoprotein structure, ∼0.5 Å.

Of the four cases (ligands 2, 3, 5 and 22) where the cross docking did not predict the correct structure, the benzimidazole ligand 22 was the most interesting. It was cross docked with a receptor that had its helices shifted from the positions in the apoprotein; positions previously shown to be favorable for benzimidazole ligands. However, ligand 22 would be the only benzimidazole that the cross docking did not predict the correct pose for. Moreover, ligand 32, which is almost an identical copy of ligand 22, was able to be cross docked with good results. This was odd considering their similarities. When comparing these with the other ligands docked from the benzimidazole series, unlike all others they do not have an aromatic ring bonded to the 2^{nd} carbon in the benzimidazole group, see Figure 10. But since the cross–docking was able to dock ligand 32, one cannot draw any conclusions at this point.

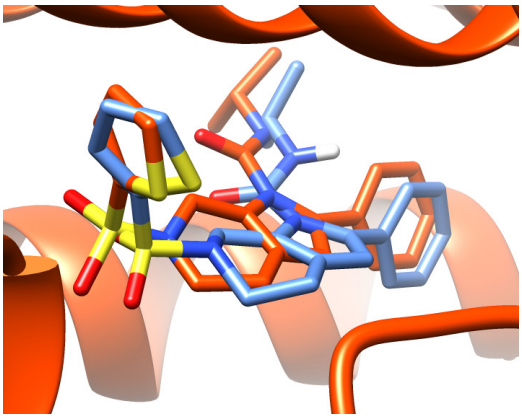


Figure 8: Pose prediction 2 for ligand 16 (blue)(docked with 3FLI) and the experimental pose (orange) aligned to the apoprotein (orange ribbon). RMSD of 1.05 Å between the poses.

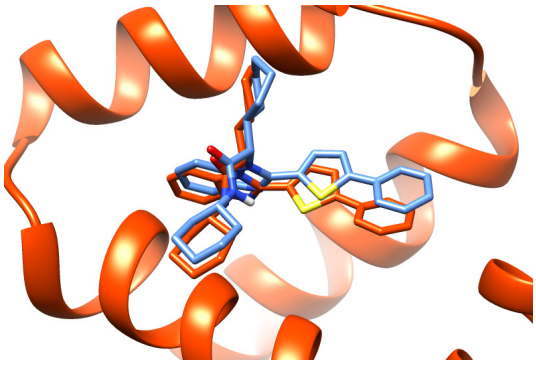


Figure 9: Pose prediction 1 for ligand 7 (blue)(docked with 3OMM) and the experimental pose (orange) aligned to the apoprotein (orange ribbon). RMSD of 1.27 Å between the poses.

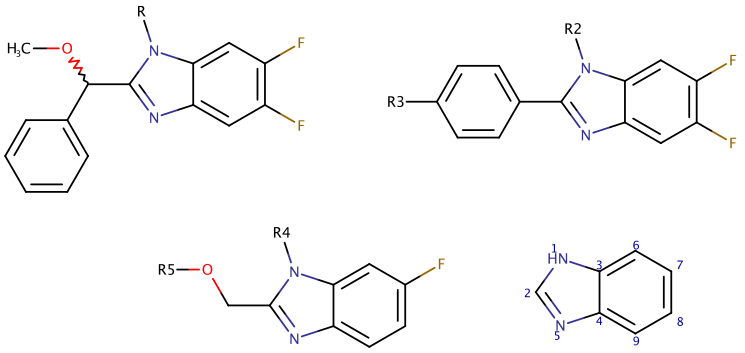


Figure 10: A simplified overview (some functional groups may differ slightly) for the major structural difference between the ligands from the benzimidazole chemical series. Top left, the branching is seen at the carbon that is bonded to the C_2 in the benzimidazole group which is found in ligand 22 and 32. In the top right corner, the most common structure for the ligands from this chemical series. Here, some aromatic ring is bonded to the C_2. In the bottom left, the structure for ligand 20 is shown, where the C_2 is followed by an ether bridge. The last structure, in the bottom right, is the benzimidazole group with notation for the carbons. R is the notation for different side–chains.

Table 3: The RMSD, in Å, for the five top predicted poses from AutoDock Vina versus the D3R experimental protein crystal structures with the respective ligands. Ligands with a number followed by A or B indicate that two different binding poses have been seen in the crystallographic data. Ligand 33, given in the initial D3R package, was not correct and therefore the RMSD calculations could not be performed. The highlighted values and crystal structures correspond to accurately predicted poses.

	Ligand																																					
Pose	1	2	3	4	5	6A	6B	7A	7B	8	9	10	11	12A	12B	13	14A	14B	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
1	5.47	6.48	4.62	3.81	6.25	1.96	1.51	1.27	1.78	5.62	1.29	5.56	4.82	4.59	4.56	7.84	2.23	1.67	9.44	8.18	1.44	8.60	1.64	7.07	6.50	1.60	7.61	6.92	2.00	7.88	1.96	1.92	2.00	6.69	2.13	4.41		11.00	1.59	8.23
2	7.20	6.73	8.08	3.96	7.06	6.71	6.75	7.54	8.25	1.86	7.72	9.60	8.76	2.76	2.86	7.64	6.81	7.04	9.58	1.05	2.32	8.59	5.11	1.84	1.78	6.93	7.83	5.17	6.88	1.82	7.93	6.28	6.31	7.77	7.22	6.30		11.10	5.56	1.55
3	6.00	7.59	5.07	6.78	6.57	5.73	5.86	3.31	3.58	6.58	6.61	9.09	5.16	9.64	9.52	7.07	6.63	6.80	9.36	8.56	4.68	8.90	6.51	5.09	5.16	2.74	4.48	4.49	6.78	7.05	5.18	7.65	7.71	5.80	6.08	2.63		10.80	5.29	7.01
4	3.14	6.44	5.37	5.51	5.79	5.26	5.15	6.67	7.43	6.29	4.83	9.15	4.25	9.78	9.72	8.11	5.69	5.68	7.39	3.93	8.30	9.35	4.54	6.42	4.39	6.68	6.86	1.81	1.82	7.80	5.54	4.70	4.69	4.51	5.42	4.60		5.79	7.85	8.47
5	7.20	5.62	6.89	6.68	7.22	2.13	2.84	6.45	6.72	7.11	6.78	9.72	9.31	9.18	9.19	8.84	5.17	5.00	4.38	8.34	7.40	8.79	6.94	4.49	6.86	4.96	3.48	6.61	6.62	7.67	7.32	6.93	5.96	2.40	6.07	5.68		10.80	5.04	6.81
1	3OMK	3OMM	3P88	3P88	3OMM	3OMK	3OMK	3OMM	3OMM	3OMM	3OMK	1OSH	1OSH	1OSH	1OSH	3OLF	3OLF	3OLF	3DCU	3DCU	3FLI	3OOF	3OMM	3OMM	3OMK	3OMM	3HC6	3OMK	3OMM	3OMK	3OMM	3OMM	3OMM	3OMK	3OMM	3OMM		3P88	3OOF	3OMM
2	3DCU	3OMM	1OSH	3DCT	3FLI	3OMM	3OMM	3OMM	3OMM	3OMM	3OMM	4QE6	3L1B	3FLI	3FLI	3OMM	3OLF	3OLF	3HC5	3FLI	3FLI	3OKH	3OLF	3OMK	3OMM	3OLF	3RUT	3OMM	3OMM	3OMM	3OOM	3OLF	3OLF	3OLF	3OMM	3OMM		3P88	3OOF	3OMM
3	3OMK	3OKH	3RUT	3OOF	3OOF	3OOF	3OOF	3L1B	3L1B	3OMM	3OOF	3OKH	3OOF	3OKH	3OKH	3OKH	3OLF	3OLF	3DCU	3RUT	3OLF	3OKH	3OMK	3OMM	3OMK	3OMK	3FLI	3OOF	3OOF	3OMK	3OLF	3OMK	3OMK	3OMM	3OOF	3OMM		3P88	4QE6	3OOF
4	3OLF	4QE6	3OMM	3OOF	4QE6	3OLF	3OLF	3OLF	3OLF	3OMK	3OMK	3OLF	3OKH	4QE6	4QE6	3FLI	3OLF	3OLF	3OMM	1OSH	3L1B	3OOF	3OLF	3OOF	3OKH	3OOF	3OOF	3OMM	3OMM	3OMM	3OMM	3OLF	3OLF	3OOF	3FLI	3OMM		3OLF	3L1B	3L1B
5	3HC5	3RUT	3FLI	3OMK	3L1B	3OMM	3OMM	3OKH	3OKH	3OLF	3OLF	3L1B	3OMM	3OKH	3OKH	3P88	3OMK	3OMK	1OSH	3L1B	3L1B	3OMM	3OLF	3OOF	3OKH	3FLI	3OLF	3OMK	3P88	3OLF	3OMM	3OMM	3OMK	3OLF	1OSH	3OMM		3P88	3OKH	3L1B

Table 4: The RMSD, in Å, for the top five AutoDock Vina cross-docking predictions of the D3R experimental protein crystal structures with their respective ligands. Ligands with a number followed by A or B indicate that two different binding poses have been seen in the crystallographic data. Here, the crossdocking is performed with the same ligand but with both of the two binding conformations (A and B) of the receptor shown.

	Ligand																																					
Pose	1	2	3	4	5	6A	6B	7A	7B	8	9	10	11	12A	12B	13	14A	14B	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
1	0.90	2.59	5.14	1.10	2.63	0.37	1.62	0.59	1.70	0.89	0.49	1.52	1.94	1.79	1.90	0.32	0.75	1.66	8.07	0.71	0.50	4.10	1.00	0.80	0.42	3.64	0.58	1.26	6.60	0.39	0.70	1.22	0.70	2.52	1.57	0.74	0.92	1.88	1.99	0.81
2	0.85	4.19	3.58	2.58	7.16	1.65	0.34	0.88	1.79	6.55	7.47	1.47	3.11	2.20	1.86	4.65	1.65	0.58	8.82	1.85		1.28	5.74	7.16	6.35	5.37	6.84	6.50	1.56	7.41	7.45	5.61	7.44	1.96	6.96	5.27	1.89	3.70	2.65	8.08
3.	1.94	2.55	6.88	4.04	7.41	6.57	6.55	7.45	8.09	5.04	6.86						6.54	6.75	1.96	7.36		5.92		4.14	5.73	6.51	7.07	5.73	6.53	7.02	5.52	1.73	5.62	7.33	5.83	3.77	3.77	3.72	5.30	
4	6.39	3.82	6.89	4.11	5.53	5.57	5.63	4.81	5.07	1.24									2.67	1.88				5.35	1.47	5.53	4.26		5.85	4.72		7.25	5.14		2.26	5.22	2.51	2.03	5.53	
5	6.42	6.66	7.78	4.12	6.94					6.43									2.21					7.37	5.83	5.85	4.47		1.28			5.65				2.17		3.26	7.74	

Table 5: The RMSD, in Å, for the predicted pose of the largest MD simulation cluster for the respective ligand-complexes versus the D3R experimental crystal structures.

	Ligand																																					
Pose	1	2	3	4	5	6A	6B	7A	7B	8	9	10	11	12A	12B	13	14A	14B	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
1	5.24	7.67	4.17	4.08	6.76	2.03	1.71	1.44	2.34	5.24	1.50	3.69	3.68	4.51	4.44	7.60	1.98	1.41	6.97	7.82	2.03	8.58	1.17	6.27	6.59	1.72	7.55	6.80	1.77	7.58	1.96	1.47	1.70	6.69	2.28	4.81		5.70	1.84	7.92
2	5.03				6.18									4.95	4.74				7.61	8.16																				

4.2 Molecular Dynamics

Molecular Dynamics was performed to refine the docking poses further, by letting the ligand–receptor complexes relax during a simulation with explicit water molecules. In addition to the pose refinement, the MD would also show the stability of the predicted pose. If the pose would deviate too much from the pose predicted by docking during the MD simulations, it would most likely not be in low free–energy state. Thus, if the pose deviation was $\geq$2.5 Å, the pose would be considered as unstable. In order to be truly unbiased by experimental data, the top pose prediction from AutoDock Vina was always chosen as a starting point for simulations.

For the 36 different Molecular Dynamics simulations, see Table 6, only five (3, 5, 10, 15 and 16) did not continue past the initial 20 nanoseconds. Neither of these were good predictions, when judging by the experimental results, see Table 3. Thus, it was not odd that these simulations were terminated. Yet these were the only 5 out of the 22 ligands that were poorly predicted by docking. One explanation for this could be that the predicted poses for the other ligands were in a local free–energy minimum and therefore stable enough for the MD simulations. However, the predicted pose did not correspond to the global free–energy minimum. Therefore, the simulations could continue even if the predicted poses were not correct.

When comparing the poses after MD with the crystal structures, the 1–2 poses found by clustering did not seem to have improved the overall docking predictions further, see Table 5.

Table 6: The complexes chosen based on the AutoDock Vina results. The table shows which receptor was used with which ligand going into the MD simulations.

Ligand	Receptor
1	3OMK
2	3OMM
3	3P88
4	3P88
5	3OMM
6	3OMK
7	3OMM
8	3OMM
9	3OMK
10	1OSH
11	1OSH
12	1OSH
13	3OLF
14	3OLF
15	3OMM
16	3DCU
17	3FLI
18	3OOF
19	3OMM
20	3OMM
21	3OMK
22	3OMM
23	3HC6
24	3OMK
25	3OMM
26	3OMK
27	3OMM
28	3OMM
29	3OMM
30	3OMK
31	3OMM
32	3OMM
33	3DCU
34	3OLF
35	3OOF
36	3OMM

4.3 Reconnaissance metadynamics

Without the RMSD restraints and with addition of bias to the system, RMD would perhaps be able to overcome these free–energy barriers as well as find and refine the correct pose, even if the starting pose for the simulations was not accurate.

Due to limited amount of time before the release of the *D3R Grand Challenge 2016* crystallographic experimental data, initially only three MD simulations were chosen for further RMD simulations. Later when the crystal structures were released, three additional ligands were chosen for RMD. The reasoning for performing RMD was to see whether addition of bias would overcome the high energy barriers in MD simulations, and if the enhanced sampling would be able to find the correct pose regardless of the starting stucture for the ligand.

Table 7: The RMSD between the poses found with clustering of the RMD simulations (for ligands 22, 27 and 32) and the corresponding D3R experimental crystal structures. The RMSD values are in Å, and correspond to the 4–5 top pose predictions from the clustering of the RMD data.

	Ligands		
Pose	22	27	32
1	1.33	4.63	6.90
2	2.53	5.01	4.81
3	6.73	5.20	5.87
4	1.72	4.53	4.83
5		1.96	5.50

All dihedrals of the ligands were manually determined and chosen as *collective variables*. The three complexes (22, 27 and 32) were chosen randomly from the set of complexes but with certain criteria: they had all completed the full MD simulation time of 50 nanoseconds. These structures would be the starting point for the RMD. The simulation were run for 30 nanoseconds and then a clustering analysis was performed to determine the largest basin clusters. The predicted poses were determined through a combination of clustering and additional MD simulations as detailed in section 3.5.2.

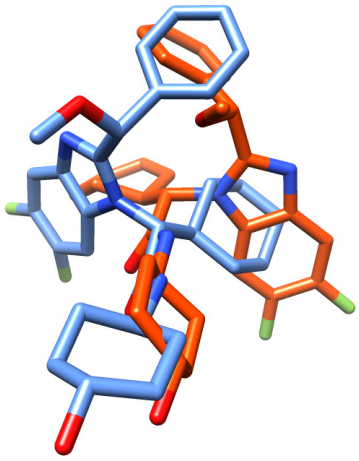


Figure 11: The poses for ligand 32 aligned to the active site of FXR (protein not shown). The crystal pose shown in orange and the MD pose prediction is depicted in blue.

The pose predictions trivally included the MD pose. For ligand 22, RMD did refine the MD pose further. Moreover, for ligand 22 the RMD managed to rank the best prediction (1.33 Å) as the top pose. However, the second best prediction at 1.72 Å (the MD pose), did not stay the second ranked pose. Except for the ranking, the RMD did function as intended for this ligand.

Ligands 27 and 32 did not manage to improve the results further. For ligand 27, the best prediction was the MD pose (1.96 Å) but the internal ranking put it in last place.

Ligand 32 had the MD pose in second place (4.81 Å), but in contrast to ligand 22 and 27 this was not an accurate pose to begin with. Still, RMD did not manage to improve the pose prediction for the ligand. When comparing the MD pose with the correct pose for ligand 32, it is clear that there is a difference which RMD in this case was not able to overcome, see Figure 11.

To see whether more CVs would be able to help find the correct poses, three more RMD runs were performed. They were set up as the three previous simulations but had a duplicate with side–chain CVs in addition to the torsion CVs for the ligand. The duplicate enabled comparison with and without added

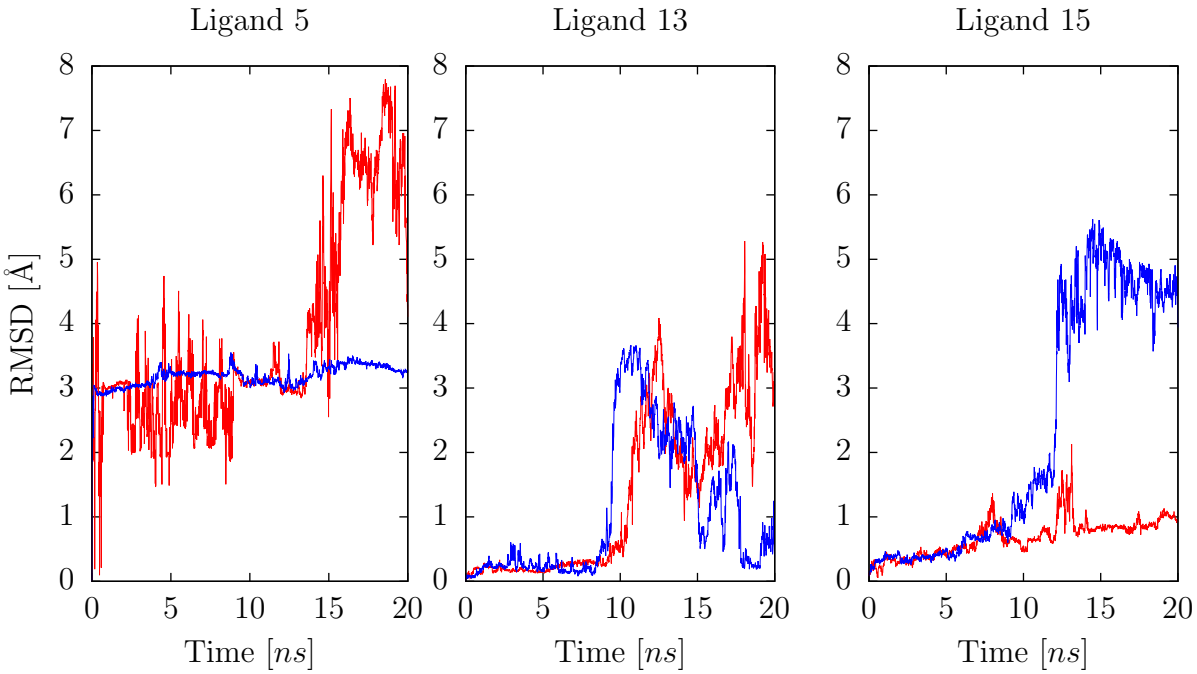


Figure 12: The RMSD between the MD pose and the conformation with time for each of the three ligands' RMD trajectories with and without sidechain CVs. The red line represents the RMSD for the trajectory *without* sidechain CVs and the blue line represents *with* sidechain CVs. All RMSD plots starts at RMSD($t = 0$) = 0, but for ligand 5 there is a large jump from the starting pose in the beginning of the simulations both with and without side–chain CVs.

side–chain CVs. The side–chains with CVs can be seen in Table 1, in section 3.4. Ligand 5, 13 and 15 were chosen for the simulations based on the criteria that they all did not perform well in the docking pose prediction and did not complete the full 50 nanoseconds of MD. The reasoning was to see whether these more complicated cases would be aided by RMD or not.

The pose predictions did not become any better with the RMD as can be seen by the RMSD values in Table 8. The hope was that the addition of side–chain biasing would increase the flexibility and volume of the active site which would give the ligand more room to explore more conformations. However, the additional CVs did not improve the result. In contrast to the RMD simulations for ligands 22, 27 and 32, the MD pose was not included as one of the ranked poses for the new RMD ligands. When studying the differences between the RMD poses and the D3R experimental crystal structures, one can see that

Table 8: The RMSD between the *D3R* crystallographic structure and the prediction from the reconnaissance metadynamics simulations for ligand 5, 13 and 15. Data from both the simulation with CVs on only the dihedrals of the ligand (No SC CVs) and with both CVs for dihedrals and on a few side–chains (SC CVs) are presented. RMSD in Å.

	No SC CVs			SC CVs		
Pose	5	13	15	5	13	15
1	6.77	7.66	7.33	5.93	7.79	7.38
2	5.97	8.26	7.72	5.74	7.42	6.36
3	6.79	7.34	6.44	6.81	8.35	7.05
4	6.32	7.72	7.00	6.85	6.87	6.75
5	7.02	8.18	7.48	5.81	7.71	7.27

the ligands were rotated around their axis. The CVs used were not enough to overcome these starting positions. Introduction of different types of CVs can perhaps overcome this problem. CVs that would help with larger conformational changes in the receptor could perhaps allow the ligand to leave the flipped pose. CVs that would let the whole ligand rotate around its axis in–place could perhaps also be an option.

By comparing the RMSD for the trajectories with and without additional side–chain CVs, one could compare whether the addition of side–chain CVs did lead to more sampling or not. If the RMSD for the simulation with side–chain CVs increased, then more poses would have been sampled as the simulation would have gone further away from the initial pose. This was the intended effect of the introduction of side–chains CVs, but the RMSD calculations did not show this desired outcome for these cases. In Figure 12 it can be seen that the addition of CVs did not clearly increase the sampling. For ligand 5, after the initial jump in RMSD at the beginning of the simulation, the side–chains CVs did not improve the sampling. This differs with the simulation without side–chain CVs, where the sampling was increased. Only in the case for ligand 15 the addition of the side–chain CVs did increase the sampling.

5 Conclusions

The method described in the thesis was not able to accurately predict the binding poses for all ligands in the set. The positions for two of the α–helices adjacent to the active site in FXR proved to be important for the accuracy of the docking prediction. The positions of the α–helices depended on the chemical series that the ligands came from, and it is shown that when the receptor had the helices in the correct position for a ligand, it docked more accurately. However, the positions of the side–chains could also be affecting the binding accuracy. This possibility, in addition to the positions of the α-helices and any interplay between the two, was not studied in the thesis.

For 32 out of 36 ligands, where the ligands were cross docked with the D3R experimental crystal structures, the docking software were able to predict the pose. However, the software was only able to predict the poses for 19 out of 36 cases when docking with the PDB and apoprotein crystal structures. The nineteen protein conformations, that were able to predict the binding pose, were also similar to the experimental structure. Thus, it appears to be difficult to get a good prediction with a rigid protein conformation not close to the crystal structure as the conformation of the receptor affects the docking accuracy. Therefore, the lack of flexibility in the receptor when docking the ligand could possibly be resolved by using a more flexible docking method.

In only 3 out of the 14 cases, where the docking pose were accurate, did Molecular Dynamics further refine the structures to a pose with lower binding free–energy. This could be a consequence of deficiencies in the force field, considering that a simple force field was used (GAFF). Another possibility is that MD was not able to find the lowest free–energy state with this force field, even though the pose was close to that of the crystal structure in these cases.

The collective variables used in Reconnaissance metadynamics simulations for the six ligands, were not able to find poses more similar to the poses found in the crystal structures. A possible explanation is that the CVs on the rotable bonds in the dihedrals in the ligands and side–chains were not enough to allow the ligands to fully rotate and explore the free–energy surface thoroughly for the active site. A better set of CVs for the positions of the helices and on the dihedral angles for the side–chains in the active site, could be key to solve the steric hindrance that the ligands experienced.

6 Further Work

Further work is needed to determine the relative importance of the positions for the side–chains versus the overall configurations of the α–helices for pose predictions. Also, in order to construct a better set of collective variables which takes the flexibility of the protein into account, any interplay between them has to be determined.

If bias potentials are added to the positions of the two α–helices adjacent to the active site, the ligands may be allowed to more freely explore the active site in order to find a configuration state with lower free–energy. Also, by utilizing flexible docking methods, the short–comings of the rigid docking would perhaps be avoided and allow for better pose predictions.

A better force field model for the ligands could also be determined and used in order to possibly improve the refinement of the ligands by MD.

7 Ethics and Conflicting Interests

There are no ethical or conflicts of interest known.

References

[1] D. L. Mobley and K. A. Dill. Binding of small-molecule ligands to proteins: "what you see" is not always "what you get". *Structure*, 17:489–498, 2009.

[2] S. B. Jiang, Q. Zhao, and A. K. Debnath. Peptide and non-peptide HIV fusion inhibitors. *Current Pharmaceutical Design*, 8:563–580, 2002.

[3] R. O. Dror, A. C. Pan, D. H Arlow, D. W. Borhani, P. Maragakis, Y. Shan, H. Xu, and D. E. Shaw. Pathway and mechanism of drug binding to G-protein-coupled receptors. *Proceedings of the National Academy of Sciences of the United States of America*, 108:13118–13123, 2011.

[4] Y. Shan, E. T. Kim, M. P. Eastwood, R. O. Dror, M. A. Seeliger, and D. E. Shaw. How does a drug molecule find its target binding site? *Journal of the american chemical society*, 133:9181–9183, 2011.

[5] D. E. Shaw, R. O. Dror, J. K. Salmon, J. P. Grossman, K. M. Mackenzie, J. A. Bank, C. Young, D. Young, M. Martin, B. Batson, K. J. Bowers, E. Chow, M. P. Eastwood, D. J. Ierardi, J. L. Klepeis, J. S. Kuskin, R. H. Larson, K. Lindorff-Larsen, P. Maragakis, M. A. Moraes, S. Piana, Y. Sh, and B. Towles. Millisecond-scale molecular dynamics simulations on anton. In *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*, SC '09, pages 39:1–39:11, New York, NY, USA, 2009. ACM.

[6] P. Söderhjelm, G. A. Tribello, and M. Parrinello. Locating binding poses in protein-ligand systems using reconnaissance metadynamics. *Proceedings of the National Academy of Sciences of the United States of America*, 109:5170–5157, 2012.

[7] D3R grand challenge 2016. https://drugdesigndata.org/about/grand-challenge-2. Accessed: 2016-09-19.

[8] H. G. Richter, G. M. Benson, K. H. Bleicher, D. Blum, E. Chaput, N. Clemann, S. Feng, C. Gardes, U. Grether, P. Hartman, B. Kuhn, R. E. Martin, J-M. Plancher, and M. G. Rudolph. Optimization of a novel class of benzimidazole-based farnesoid x receptor FXR agonist to improve physicochemical and ADME properties. *Bioorganic and Medicinal Chemistry Letters*, 21:1134–1140, 2011.

[9] A. R. Leach. *Molecular Modelling - Principles and Applications 2nd Edition.* Pearson Education Limited, Essex, England, 2001.

[10] S. F. Sousa, P. Alexandrino Fernandes, and João Ramos. Protein–Ligand Docking: Current Status and Future Challenges. *PROTEINS: Structure, Function, and Bioinformatics*, 65:15–26, 2006.

[11] O. Trott and A. J. Olson. Autodock vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. *Journal of Computational Chemistry*, 31:455–461, 2010.

[12] W. D. Cornell, P. Cieplak, I. R. Gould, K. M. Jr. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*, 117:5179–5197, 1995.

[13] A. Barducci, M. Bonomi, and M. Parrinello. Review: Metadynamics. *Computational Molecular Science*, 1:826–843, 2011.

[14] A. Laio and F. L. Gervasio. Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Reports on Progress in Physics*, 71, 2008.

[15] G. A. Tribello, M. Ceriotti, and M. Parrinello. A self-learning algorithm for biased molecular dynamics. *Proceedings of the National Academy of Sciences of the United States of America*, 107:17509–17514, 2010.

[16] R. Anandakrishnan, B. Aguilar, and A. V. Onufriev. H++ 3.0: automating pk prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic Acids Research*, 40:537–541, 2012.

[17] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison. Open babel: An open chemical toolbox. *Journal of Cheminformatics*, 3:33–47, 2011.

[18] J. Wang, W. Wang, P. A. Kollman, and D. A. Case. Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of Molecular Graphics and Modelling*, 25:247–260, 2006.

[19] 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*, 25:1157–1174, 2004.

[20] 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. Kolossvry, 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*. University of California, San Franscisco, 2014.

[21] S. Forli, R. Huey, M. E. Pique, M. F. Sanner, D. S. Goodsell, and A. J. Olson. Computational protein-ligand docking and virtual drug screening with the autodock suite. *Nature Protocols*, 11:905–919, 2016.

[22] Berendsen et al. Gromacs. *Comp. Phys. Comm.*, 91:43–56, 1995.

[23] M. Bonomi, D. Branduardi, G. Bussi, C. Camilloni, D. Provasi, P. Raiteri, D. Donadio, F. Marinelli, F. Pietrucci, R.A. Broglia, and et al. Plumed: A portable plugin for free-energy calculations with molecular dynamics. *Comp. Phys. Comm.*, 180:1961–1972, 2009.

[24] Schrödinger Suite 2016-3. *Maestro 10.7*. Schrödinger, LLC, New York, NY, 2016.

[25] X. Daura, K. Gademann, B. Jaun, D. Seebach, W. F. van Gunsteren, and A. E. Mark. Peptide folding: When simulations meets experiment. *Angewandte Chemie International Edition*, 38:236–240, 1999.

8 Appendix

8.1 List of Softwares and Computational Tools

- GROMACS 4.6.2
  - A Molecular Dynamics software.
- PLUMED 1.3
  - A plugin to GROMACS which enables biasing with collective variables, i.e. a plugin for metadynamics and reconnaissance metadynamics.
- AutoDock Tools 1.5.6
  - A toolbox for AutoDock; In this thesis used for the construction of the docking grid box.
- AutoDock Vina 1.1.2
  - A command line molecular docking software.
- AutoDock VS Raccoon 1.0
  - Docking tool which was used for conversion between the file–formats *.pdb* and *.pdbqt*.
- Open Babel 2.3.0
  - File format conversion. Also used for conformational search.
- AmberTools 14
  - Used to build the force field (topology).
- Acpype 0.0.0
  - Conversion of topology from Amber to GROMACS format.
- H++ 3.0
  - A web–based software used to determine the protonation state.
- HPC2N Abisko
  - High Performance Computing Center North, a CPU cluster where most Molecular Dynamics and Reconnaissance metadynamics simulations were performed.
- Schrödinger 2016-3: Maestro 10.7
  - A part of the software package were used, namely the *rmsd.py* tool. It was used for RMSD calculations [24].
- Kubuntu 12.04.5
  - The simulations and scripting were performed on the Kubuntu distribution of Linux.

8.2 Ligands

The SMILES–strings for the ligands can be seen in the following Table 1.

Table 1: The SMILES–strings, given in the D3R Grand Challenge 2016, for the 36 ligands.

Ligand	SMILES–strings
1	COc1ccc(cc1OC)N(CC(=O)NC(C(C)C)C(C)C)S(=O)(=O)c2ccc(C)cc2
2	Cc1cccc(C)c1Nc2c(nc3cccc(C)n23)c4c(F)cccc4F
3	Fc1ccc(\C=C\C(=O)N(C2CCCCC2)C(=O)NC3CCCCC3)cc1
4	Cc1onc(c2ccc(Cl)cc2Cl)c1C(=O)N(C3CCCCC3)C4CCCCC4
5	CCOC(=O)C1=CN(CC(C)(C)c2c1[nH]c3ccccc23)C(=O)c4ccc(F)c(F)c4
6	O=C(NC1CCCCC1)[C@H](C1CCCCC1)n1c2ccccc2nc1c1ccc(OC)cc1OC
7	O=C(NC1CCCCC1)[C@H](C1CCCCC1)n1c2ccccc2nc1c1ccc(s1)c1ccccc1
8	c1(n([C@H](C(=O)NC2CCCCC2)C2CCCCC2)c2c(n1)cccc2)c1ccc(cc1)CO
9	O=C(NC1CCCCC1)[C@H](C1CCCCC1)n1c2ccccc2nc1c1ccc(cc1)c1[nH]nnn1
10	OC(=O)c1ccc(CN2C(=O)C3(CCN(CC3)S(=O)(=O)c4cccs4)c5cc(Br)ccc25)cc1
11	Brc1ccc2N(Cc3ccc(cc3)c4nnn[nH]4)C(=O)C5(CCN(CC5)S(=O)(=O)c6cccs6)c2c1
12	OC(=O)c1ccc(CN2C(=O)C3(CCN(CC3)S(=O)(=O)c4ccccc4Cl)c5cc(Br)ccc25)cc1
13	c1ccc2c(c1)nc(n2[C@H](C(=O)NC1CCCCC1)C1CCCCC1)c1ccc(C(=O)c2ccccc2)cc1
14	COc1ccc(c(OC)c1)c2nc3ccccc3n2[C@@H](C4CCCCC4)C(=O)Nc5c(C)cccc5C
15	Fc1ccccc1S(=O)(=O)N2CCc3cc(C(=O)NCc4ccccc4)n(Cc5ccccc5)c3C2
16	CC(C)NC(=O)n1c2CN(CCc2cc1c3ccccc3)S(=O)(=O)c4cccs4
17	CCOC(=O)c1ccc(NC(=O)c2c3CN(CCc3nn2c4ccccc4)S(=O)(=O)c5cccs5)cc1
18	C(=O)(c1ccc(cc1)Cl)N([C@@H](C1CCCCC1)C(=O)NC1CCCCC1)c1cc(ccc1)NC(=O)C
19	[H][C@]1(CCCCO1)[C@@]([H])(C(=O)NC2CCCCC2)n3c4cc(c(cc4nc3c5ccc(cc5)Cl)F)Cl
20	c1c(cc2c(c1)nc(n2[C@H](C(=O)NC1CCCCC1)C1CCCCC1)COc1ccc(cc1C)Cl)F
21	Fc1cc2nc(c3ccc(Cl)cc3)n([C@@H](C4CCCCC4)C(=O)Nc5ccccc5)c2cc1F
22	c1(c(cc2c(c1)nc(n2[C@H](C(=O)NC1CCCCC1)C1CCCCC1)[C@@H](c1ccccc1)OC)F)F
23	CC(C)c1onc(c2ccccc2Cl)c1C(=O)N[C@H](C)c3ccc4ccccc4c3
24	Fc1ccccc1NC(=O)[C@H](C2CCCCC2)n3c(nc4cc(F)c(F)cc34)c5ccc(Cl)cc5
25	c1(c(cc2c(c1)nc(n2[C@@H](C1CCCCC1)C(=O)Nc1c(cccc1)C#N)c1ccc(cc1)Cl)F)F
26	Fc1cc2nc(c3ccc(nc3)n4cccn4)n([C@@H](C5CCCCC5)C(=O)NC6CCCCC6)c2cc1F
27	OC(=O)c1ccc(NC(=O)[C@H](C2CCCCC2)n3c(nc4cc(F)c(F)cc34)c5ccc(Cl)cc5)c(Cl)c1
28	[H][C@@](COc1ccc(cc1F)C(=O)O)(C2CCCCC2)n3c4cc(c(cc4nc3c5ccc(cc5)Cl)F)F
29	[H][C@@](COc1ccc(cc1)C(=O)O)(C2CCCCC2)n3c4cc(c(cc4nc3c5ccc(cc5)Cl)F)F
30	COc1ccc(c(OC)n1)c2nc3cc(F)ccc3n2[C@@H](C4CCCCC4)C(=O)N[C@@H]5CC[C@H](CC5)C(=O)O
31	c1(n(c2c(n1)cc(c(c2)F)F)[C@H](C(=O)N[C@H]1CC[C@@H](CC1)O)C1CCCCC1)c1c(nc(cc1)OC)OC
32	CO[C@@H](c1nc2cc(F)c(F)cc2n1[C@@H](C1CCCCC1)C(=O)N[C@H]1CC[C@H](O)CC1)c1ccccc1
33	CC(C)c1onc(c1COc2ccc(c(C)c2)c3ccc4c(cn(C)c4c3)C(=O)O)c5c(Cl)c[n+]([O-])cc5Cl
34	C[C@H](CCC(=O)Nc1cc(cc(c1)C(=O)O)C(=O)O)[C@H]2CC[C@H]3[C@@H]4CC[C@@H]5C[C@H](O)CC[C@]5(C)[C@H]4CC[C@]23C
35	C1CCCC(C1)[C@@H](C(=O)N[C@@H]1CC[C@H](CC1)OS(=O)(=O)O)n1c2c(nc1c1ccc(cc1)Cl)cc(c(c2)F)F
36	n1(c(nc2c1cc(c(c2)F)F)c1ccc(c2nccs2)cc1)[C@H](C(=O)Nc1ccc(C(=O)O)cc1)C1CCCCC1

8.3 PDB IDs for the Receptors

The list for the Receptors PDB IDs can be found in the following Table 2

Table 2: The PDB IDs for the receptors.

1OSH
3FLI
3FXV
3L1B
3OKH
3OLF
3OMK
3OMM
3OOF
4QE6
4QE8
3DCT
3DCU
3HC5
3HC6
3P88
3RUT
3RUU

8.4 Additional data

Table 3: RMSD for the protein between the crystal structures and the apoprotein. RMSD in Å.

Ligand	RMSD
1	0.878
2	0.380
3	0.764
4	0.483
5	0.995
6	1.018
7	1.064
8	1.044
9	1.037
10	0.408
11	0.347
12	0.465
13	0.992
14	1.000
15	0.486
16	0.481
17	0.796
18	0.395
19	0.858
20	1.006
21	0.988
22	0.997
23	0.578
24	1.048
25	1.060
26	1.028
27	1.029
28	1.025
29	1.043
30	1.031
31	1.007
32	1.020
33	0.670
34	0.824
35	1.011
36	1.008