UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Fisica e Astronomia “Galileo Galilei”

Corso di Laurea in Fisica

Tesi di Laurea

Calculation of Binding Free Energy via Alchemical Transformations

Relatore

Prof. Denis Bastieri

Denis Bastieri

Correlatore

Prof. Francesco Zonta

Francesco Zonta

Laureando

Giacomo Bertin

Anno Accademico 2019/2020

Contents

Protein-ligand binding is essential to almost all biological processes, and the underlying physical and chemical interactions determine the specific biological recognition at the molecular level. In drug discovery, one tries to find a molecular ligand that either inhibits or activates a specific protein target through ligand binding. However, finding a ligand that binds a targeted protein with high affinity is a major challenge in early-stage drug discovery. Therefore, improving the accuracy of free energy calculation for estimating protein-ligand binding affinity is of significant interest as well as practical utility in drug discovery. As a matter of fact, it has been estimated that a preliminary computational screening that can reach precision of the order of 1 kcal/mol can speed up by several times the process of drug discovery (Mobley et al., 2012), as we can see from fig. 1.

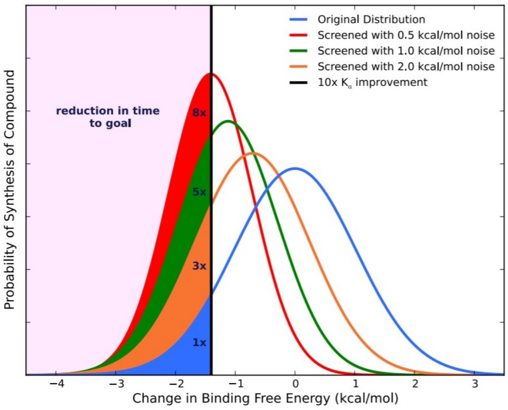

Figure 1: Preliminary computational screening help to drastically reduce the number of molecule that have to be carried on in the experimental phase, thus importantly reducing the time necessary to develop a new drug.

Computation can help speed up the drug discovery process through simulations and could be effectively used for the design of small molecule drugs for the treatment of diseases. Even though the problem can be considered theoretically solved, computation of the binding free energy still remains a challenging, as the binding process can involve complex structural rearrangement of the protein which are not easy to simulate with current computational power (this is the case of the HIV-1 protease, see Ghosh et al., 2011).
In this thesis we will discuss the most effective methods for the in vitro and *in silico* virtual screening, the basics of molecular dynamics and finally an example of scoring for some ligands used to inhibit the Bromodomain-4 (a common domain targeted by anti-tumoral drugs).

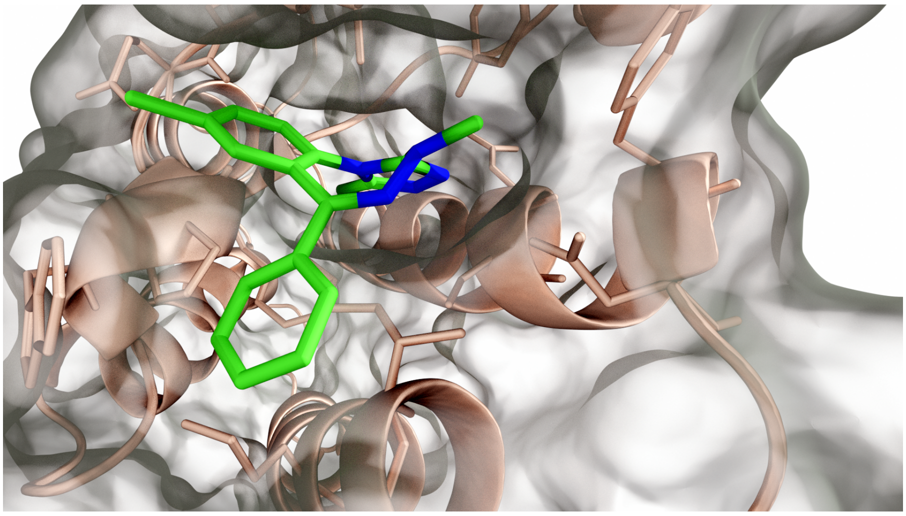

Figure 2: Bromodomain-4 binding site

Chapter 1

Introduction to Drug Discovery

One of the most successful ways to find promising drug candidates is to investigate how the target protein interacts with randomly chosen compounds, which are usually a part of compound libraries. This testing is often done in so called high-throughput screening (HTS) facilities. Compound libraries are available in sizes of up to several millions of compounds. The most promising compounds obtained by screening such libraries, i.e. the compounds that show binding activity towards the target, are called hits. Some of these hits are then promoted to lead compounds which are further refined and modified in order to achieve more favorable interactions and less side-effects. Beside being able to screen molecules using experimental methods it is also possible to use virtual screening methodologies based on the computationally inferred or simulated real screening; the main advantages of these methods compared to laboratory experiments are:

- low costs, indeed no compounds have to be purchased externally or synthesized by a chemist;
- it is possible to investigate compounds that have not been synthesized yet;
- conducting HTS experiments is expensive and VS can be used to reduce the initial number of compounds before using HTS methods;
- huge amount of chemicals to search from. The number of possible virtual molecules available for VS is exceedingly higher than the number of compounds presently available for HTS;

1.1 Binding Affinity

The strength of the interaction between these two molecules is defined by either binding and dissociation constants (K_b and K_d) or Gibbs energy of binding (ΔG_b) and is commonly referred to as affinity. In a simple, reversible one protein – one ligand interaction case, equilibrium exists between the free molecules (P, protein, and L, ligand) and their complex (PL) that associate and dissociate at certain rates (described by rate constants k_{on} and k_{off} , respectively):

$$P + L \leftarrow_{k_{off}}^{k_{on}} \rightarrow PL$$

we can define the binding affinity using equilibrium dissociation (K_d) or (K_b) constants:

$$K_b = \frac{[PL]}{[P][L]} = \frac{1}{K_d} = \frac{K_{on}}{K_{off}}$$

At equilibrium under standard conditions, the Gibbs energy of binding describes the energy difference between the two states:

$$\Delta G_b = RT\ln([P][L]) - RT\ln[PL] = -RT\ln\frac{[PL]}{[P][L]} = -RT\ln K_b = RT\ln K_d$$

$$\Delta G_b = \Delta H_b - T\Delta S_b$$

1.1.1 Experimental approaches to evaluate affinity

The number of viable methods for the determination of the binding affinity of two molecules is quite big, so in this section we will introduce some of the most popular (Kairys et al., 2019).

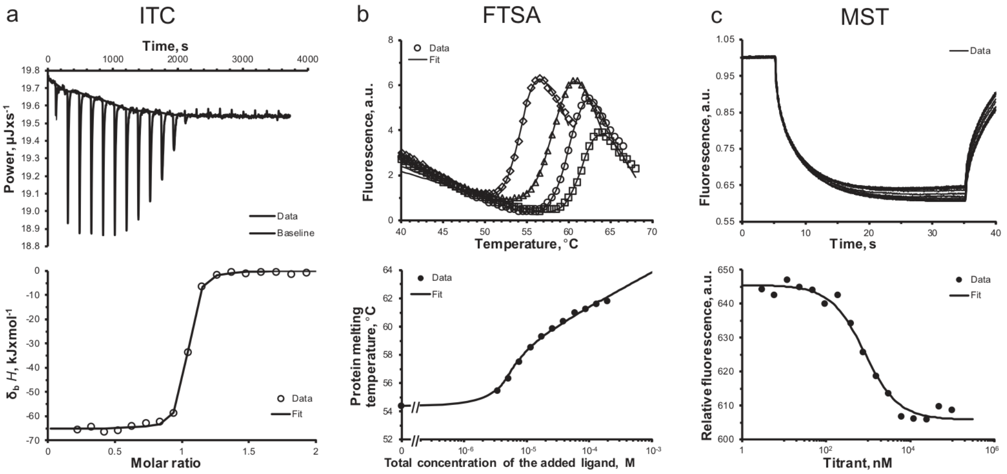

Figure 1.1: Same protein-ligand pair was studied using three different approaches: (a) isothermal titration calorimetry, (b) fluorescence thermal shift assay, and (c) microscale thermophoresis

Isothermal titration calorimetry (ITC) can measure thermodynamic energetics of binding directly, without any need of labelling, immobilization, or any other modification of the interactors. During the measurement, one of the binding partners is titrated with aliquots of the other under constant temperature, and the released or absorbed heat is measured. Construction of a binding isotherm of interaction heat as a function of titrated (Figure 1.1(a)) ligand yields ΔH_b , K_b (and therefore ΔG_b) with ΔS_b calculated as before.

Fluorescence thermal shift assay This technique is based on the hypothesis that drug lead stabilize the target protein, so the aim is to track the stability shift when the protein is heated, however, complex multi-domain proteins and their assemblies are a challenge for any stability shift assay, as unfolding of different domains and other substructures often happens independently and thus yields multiple denaturation signals that can be hard to interpret.
The protein samples with varying ligand concentrations are being subjected to a constantly increasing temperature. Protein denaturation is tracked indirectly, by measuring the fluorescence of a solvatochromic dye molecule, which reports on unfolded protein regions (Figure 1.1(b)). Thus instead of the full thermodynamic denaturation profile, it yields only K_b and T_m but is significantly less time and reagent consuming.

Microscale thermophoresis To determine the binding affinity this kind of assays physically separate the ligand or protein from the protein-ligand complex and quantify these fractions (Figure 1.1(c)). To this end, they exploit the differences in size, charge, hydrodynamic or other properties between these molecules and their assemblies. The ligand (or protein) is usually detected using UV, fluorescence signal, or MS. Determined bound fractions are plotted against target concentrations to calculate the K_d. During the Microscale thermophoresis experiment, molecules move through infrared laser-induced temperature gradients depending on their size, charge, and hydration shell. Fluorescent methods can be used to track the protein.

Nuclear magnetic resonance NMR is the only technique capable of obtaining atomic resolution structures in solution. During an NMR experiment, a magnetic field is applied to the sample, which affects the spins of the nuclei. The energy released by these nuclear spins coming back to their original states is detected and analyzed. The energy differs for the same atom nuclei placed in different chemical environments between different compounds or between the bound and unbound states. K_d determination usually requires a competitor and is limited by binding kinetics. Overall, the usage of structural biology techniques in drug design has been on the rise over the last decade due to significant technological advances and gaining popularity of computational approaches that rely on initial structural data.

1.2 High-Throughput Screening (HTS) process

Advanced methods of combinatory chemistry have made it possible to quickly synthesize vast quantities of compounds for testing. These compounds are then tested in a rapid method of evaluation called high-throughput screening. High-throughput screening (HTS) allows researchers to quickly and cost-effectively process thousands and even hundreds of thousands (ultra-high-throughput screening, or uHTS) of compounds, which enables them to increase the probability of finding an 'hits' (compounds that display the desired characteristic) that will advance into the next stages of drug discovery and development. It has been estimated that the construction of a conventional library, containing 10^6 individual molecules in sufficient quantity and quality for pharmaceutical screening campaigns, may cost between $400 million and $2 billion, so this is a very expensive technique (Cronk et al., 2013).

1.2.1 Steps in a usual HTS experiment

There are multiple steps in any HTS experiment, which can take weeks to complete. However, these steps can be generalized into three categories:

1. Sample preparation.
2. Sample handling: dedicated liquid handling robots can precisely add and mix liquid reagents to multiple wells, which allows scientists to simultaneously screen for thousands of drugs, toxins, chemicals or bioactive compounds.
3. Readouts and data acquisition: many HTS are interpreted through optical measurements – color changes in cells and in liquid reactions, the turbidity of liquid culture, or fluorescence signals.

1.3 DNA encode library

DNA-encoded chemical libraries (DECLs) are collections of compounds, individually coupled to DNA tags serving as amplifiable identification barcodes. Since individual compounds can be identified by the associated DNA tag, they can be stored as a mixture, allowing the synthesis and screening of combinatorial libraries of unprecedented size (around 10^{15} members against the $10^3 - 10^6$ compounds of screening-based method), in addition a DNA-encoded library of 800 million compounds, costs about $150,000 for materials to create and screen (Favalli et al., 2018).

We normally distinguish between two main types of DNA-encoded chemical libraries: (A) "single-pharmacophore libraries", in which individual compounds (no matter how complex) are attached to one DNA fragment; and (B) "dual-pharmacophore libraries", in which pairs of compounds are coupled to the extremity of the complementary strands of the DNA heteroduplex (Fig. 2.4). In this second example, binding fragments are identified, which need to be chemically linked at a later stage, in order to yield organic molecules which can be used in the absence of DNA.

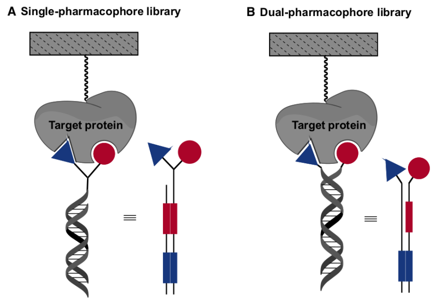

Figure 1.2: Schematic representation of DECLs process

The standard procedure for applying this method is to connect the target proteins on a solid support and subsequently incubated with a DNA-encoded chemical library, allowing the physical separation of preferential binders from the other library members, which can be washed away. After affinity capture, the barcodes of preferential binders are PCR amplified (a polymerase chain reaction in order to create millions of clonal copies of each DNA bead) and submitted to a high-throughput sequencing procedure.

This technique has been theorized in 1992 by Sidney Brenner and Richard Lerner (Brenner et al., 1992), but only the recent progresses in the DNA sequencing technologies have cut the cost and allowed to use this in real drug discovery process. High throughput methods use very different approaches in order to sequence the DNA breads, and the performance of those methods differs widely from each other. For example, one of the most popular HTS methods is the Sequencing by synthesis where the DNA molecules attached on a slide of flow cell and amplified via PCR, so that DNA clusters are created, later four type of terminator bases are added and a camera takes images of the fluorescently labeled nucleotides. This technique can solve 2.5 billion of sequence in 1 to 11 days at the cost of 5 - 150 $ / billion bases, while isolating the compound is more expensive.

The potential of DECLs can be understood by looking at some of the recent drugs discovered, such as novel Bruton tyrosine kinase inhibitors, isolated from a library of over 110 million of compounds. This protein is a target for the treatments of certain forms of lymphoma and autoimmune conditions, being key regulator in B-cell development. The drug developed compete with the ATP and inhibit the BTK (Neri et al., 2017).

Chapter 2

Computational Methods

In this section we will introduce some of the most common and efficient method for the *in silico* drug discovery.
Virtual screening methods can be subdivided in two main class: Structure-Based and Molecular Dynamics methods. The members of the first class identify the target's binding pocket and dock the ligand into it. The quality of the docking configuration is measured by a scoring function and this value is used to determine the goodness of the ligand. Usually this procedure take few minutes on a standard workstation.
Molecular Dynamics methods are more expensive in terms of computational power but they can reproduce experimental results with a precision around 1.0 kcal/mol. Methods such as the Alchemical Transformation or the Free Energy Perturbation are based on the interaction energy ligand-target during multiple and very short simulations ($\sim 1-10$ ns each).

2.1 Autodock Vina

The number of scoring function and docking software is increasing rapidly so here we will focus on one of the most precise and fast: Autodock Vina (Trott & Olson, 2010). This software uses a semi-empirical free energy force field to evaluate conformations during docking simulations. The force field was parameterized using a large number of protein-inhibitor complexes for which both structure and inhibition constants, or K_i , are known.

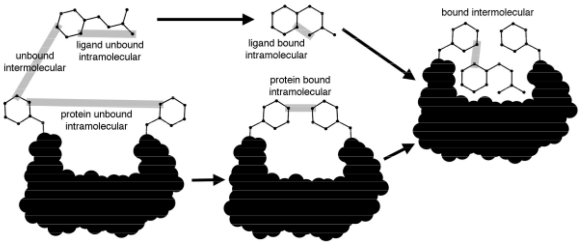
ligand unbound intramolecular
unbound intermolecular
protein unbound intramolecular
ligand bound intramolecular
protein bound intramolecular
bound intermolecular

Figure 2.1: Binding procedure

2.1.1 Scoring Function

The scoring function includes six pair-wise evaluations (V) and an estimate of the conformational entropy lost upon binding (ΔS_{conf}):

$$\Delta G = (V^{L-L}_{bound} - V^{L-L}_{unbound}) + (V^{P-P}_{bound} - V^{P-P}_{unbound}) + (V^{P-L}_{bound} - V^{P-L}_{unbound} + \Delta S)$$

where L refers to the "ligand" and P refers to the "protein" in a ligand-protein docking calculation. Each of the pair-wise energetic terms includes evaluations for dispersion/repulsion, hydrogen bonding, electrostatics, and desolvation:

$$V = W_{vdw} \sum_{i,j} (\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}) + W_{hbond} \sum_{i,j} E(\theta)(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}}) +$$

$$+ W_{elec} \sum_{i,j} \frac{q_i q_j}{e(r_{ij}) r_{ij}} + W_{sol} \sum_{i,j} (S_i V_j + S_j V_i) e^{(-r_{ij}^2/2\sigma^2)}$$

The weighting constants W have been optimized to calibrate the empirical free energy based on a set of experimentally determined binding constants. The function $E(\theta)$ provides directionality based on the angle θ from ideal H-bonding geometry. The third term is a screened Coulomb potential for electrostatics. The final term is a desolvation potential based on the volume of atoms (V) that surround a given atom and shelter it from solvent, weighted by a solvation parameter (S) and an exponential term with distance-weighting factor $\sigma = 0.35$ nm .

2.2 Molecular Dynamics

2.2.1 Introduction

Molecular dynamics simulations solve Newton's equations of motion for a system of N interacting atoms:

$$m_i \frac{\partial^2 \mathbf{r_i}}{\partial t^2} = \mathbf{F_i}, \ i = 1...N$$

The system is treated classically and the forces are obtained from the negative derivative of a potential function (force field) $V(r_1, r_2, ..., r_N)$:.

$$\mathbf{F}_i = -\frac{\partial V}{\partial \mathbf{r}_i}$$

The set of 6N differential equations for positions and velocities is integrated using small timesteps (on the order of the fs) and the dynamics is followed for several nanoseconds (up to millisecond in very long simulations) in order to compute thermodynamic properties of the system from averages along the trajectories. Due to accumulation of errors from numerical integration of the equations of Newton, coupling with thermostats and barostats are necessary to keep temperature and pressure constant. This is usually done by rescaling the velocities at regular interval during the simulation.

When we perform a molecular dynamics simulation we should always mind the different simplifications we take :

1. The simulations are classical;
2. Electrons are in the ground state;
3. Force fields are approximate;
4. The force field is pair-additive;
5. Long-range interactions are cut off;
6. Boundary conditions are unnatural;

2.2.2 Force Field

2.2.3 Non-bonded interactions

Non-bonded interactions are assumed pair-additive and centro-symmetric:

$$V(\mathbf{r}_1, ..., \mathbf{r}_N) = \sum_{i<j} V_{ij}(\mathbf{r}_{ij})$$

$$\mathbf{F}_i = -\sum_j \frac{dV_{ij}(r_{ij})}{dr_{ij}} \frac{\mathbf{r}_{ij}}{r_{ij}} = -\mathbf{F}_j$$

The non-bonded interactions contain a repulsion term, a dispersion term, and a Coulomb term. The repulsion and dispersion term are combined in either the Lennard-Jones. In addition, (partially) charged atoms act through the Coulomb term.

The Lennard-Jones interaction

The interaction between two atoms is the sum of the Lennard-Jones interaction:

$$V_{LJ}(\mathbf{r}_{ij}) = 4\epsilon_{ij}((\frac{\sigma_{ij}}{r_{ij}})^{12} - (\frac{\sigma_{ij}}{r_{ij}})^{6})$$

The σ_{ij} and ϵ_{ij} parameters could be constructed following two different combination rules:

$$\sigma_{ij} = \frac{1}{2}(\sigma_{ii} + \sigma_{jj})$$

$$\epsilon_{ij} = (\epsilon_{ii}\epsilon_{jj})^{\frac{1}{2}}$$

or an geometric average for both parameters can be used:

$$\sigma_{ij} = (\sigma_{ii}\sigma_{jj})^{\frac{1}{2}}$$

$$\epsilon_{ij} = (\epsilon_{ii}\epsilon_{jj})^{\frac{1}{2}}$$

Coulomb interaction

The Coulomb interaction between two charge particles is given by:

$$V_c(r_{ij}) = f\frac{q_j q_i}{\epsilon_r r_{ij}}$$

where $f = \frac{1}{4\pi\epsilon_0} = 138.935485$

2.2.4 Bonded interactions

Bonded interactions are based on a fixed list of atoms. They are not exclusively pair interactions, but include 3- and 4-body interactions as well. There are bond stretching (2-body), bond angle (3-body), and dihedral angle (4-body) interactions. A special type of dihedral interaction (called improper dihedral) is used to force atoms to remain in a plane or to prevent transition to a configuration of opposite chirality (a mirror image).

Harmonic bonded potential

The bond stretching between two covalently bonded atoms i and j is represented by a harmonic potential:

$$V_b(ij) = \frac{1}{2}k^b_{ij}(r_{ij} - b_{ij})^2$$

Harmonic angle potential

The bond-angle vibration between a triplet of atoms $i - j - k$ is also represented by a harmonic potential on the angle θ_{ijk}

$$V_a(\theta_{ijk}) = \frac{1}{2}k^{\theta}_{ijk}(\theta_{ijk} - \theta^0_{ijk})^2$$

The numbering i, j, k is in sequence of covalently bonded atoms.

Proper dihedrals: periodic type

ϕ is the angle between the ijk and the jkl planes,

$$V_d(\phi_{ijkl}) = k_{\phi}(1 + cos(n\phi - \phi_s))$$

Improper dihedrals: harmonic type

The simplest improper dihedral potential is a harmonic potential:

$$V_{id}(\xi_{ijk}) = \frac{1}{2}k_{\xi}(\xi_{ijkl} - \xi_0)^2$$

2.2.5 Restraints

Special potentials are used for imposing restraints on the motion of the system, either to avoid disastrous deviations, or to include knowledge from experimental data. In either case they are not really part of the force field and the reliability of the parameters is not important. Usually they are implemented as an harmonic potential.

2.2.6 Integrators

The leap-frog integrator

For the integration of the equations of motion the most used algorithm is the so-called *leap-frog.* The leap-frog algorithm uses positions $\mathbf{r}$ at time t and velocities $\mathbf{v}$ at time $t - \frac{1}{2}\Delta t$ and updates the positions and velocities using the forces $\mathbf{F}(t)$ determined by the positions at the time t using:

$$\mathbf{v}(t + \frac{1}{2}\Delta t) = \mathbf{v}(t - \frac{1}{2}\Delta t) + \frac{\Delta t}{m}\mathbf{F}(t)$$

$$\mathbf{r}(t + \Delta t) = \mathbf{r}(t) + \Delta t\mathbf{v}(t + \frac{1}{2}\Delta t)$$

The algorithm is of third order in r and is time-reversible.
The equations of motion are modified for temperature coupling and pressure coupling, and extended to include the conservation of constraints, all of which are described below.

The Langevin integrator

Stochastic or velocity Langevin dynamics adds a friction and a noise term to Newton's equations of motion, as

$$m_i\frac{d^2\mathbf{r}_i}{dt^2} = -m_i\gamma_i\frac{d\mathbf{r}_i}{dt} + \mathbf{F}_i(\mathbf{r}) + \bar{\mathbf{r}}_i$$

where γ_i is the friction constant $[1/ps]$ and $\mathbf{r}_i(t)$ is a noise process with $< \bar{r}_i(t)\bar{r}_j(t+s) >= 2m_i\gamma_ik_BT\delta(s)\delta_{ij}$.
This algorithm is not used often because require an implicit water models.

2.2.7 Water models

The choice of the water model is one of the most critical point of simulations preparation. The most used and reliable model is the TIP3P, where a water molecule is parametrized by 3 atoms, 2 harmonic bonds and 1 harmonic angle. The force field parameters are determined from quantum mechanics, molecular mechanics, experimental results, and these combinations. For simulations with a small number of atoms (100-1000) or very short time scale (some ps) can be considered more complex and precise water models, such as the TIP4P, where dummy charged atoms gives at the molecule dipole moment and a more realistic charge distribution. Some times, in order to speed up the computation, the 2 bonds are considered rigid and this can generate artifact that give poor thermodynamic results.

2.2.8 Parrinello-Rahman pressure coupling

In cases where the fluctuations in pressure or volume are important per se (e.g. to calculate thermodynamic properties), especially for small systems, it may be a problem that the exact ensemble is not well defined and that it does not simulate the true NPT ensemble. With the Parrinello-Rahman barostat, the box vectors as represented by the matrix **b** obey the matrix equation of motion:

$$\frac{d\mathbf{b}^2}{dt} = V\mathbf{W}^{-1}\mathbf{b}'^{-1}(\mathbf{P} - \mathbf{P}_{ref})$$

The volume of the box is denoted V , and **W** is a matrix parameter that determines the strength of the coupling. The matrices P and P_{ref} are the current and reference pressures, respectively. The equations of motion for the particles also change in the following way:

$$\frac{d^2\mathbf{r}_i}{dt^2} = \frac{\mathbf{F}_i}{m_i} - \mathbf{M}\frac{d\mathbf{r}_i}{dt}$$

$$M = \mathbf{b}^{-1}[\mathbf{b}\frac{d\mathbf{b}'}{dt} + \frac{d\mathbf{b}}{dt}\mathbf{b}']\mathbf{b}'^{-1}$$

The (inverse) mass parameter matrix $\mathbf{W}^{-1}$ determines the coupling strength. Since $\mathbf{W}^{-1}$ depends on the box size we only choose the approximate isothermal compressibilities β and the pressure time constant τ_p; those parameters are linked at the coupling strength by:

$$(\mathbf{W}^{-1})_{ij} = \frac{4\pi^2\beta_{ij}}{3\tau_p^2 L}$$

2.2.9 Berendsen temperature coupling

The Berendsen algorithm mimics weak coupling to an external heat bath with given temperature T_0. The effect of this algorithm is that a deviation of the system temperature from T_0 is slowly corrected according to:

$$\frac{dT}{dt} = \frac{T_0 - T}{\tau}$$

2.3 Alchemical Transformation

As previously mentioned there are different technique to compute the free binding energy in a more precise way. In this work we chosen the Alchemical Transformation because it seems to be a stable and precise method, indeed it allows to compute absolute free energy, while other methods calculate relative free energy and their precision is limited by the similarity of the ligands that you want to test; furthermore absolute energies are easier to confront with experimental values. The idea behind Alchemical Transformations is that we can create an artificial thermodynamic cycle of equilibrium states that decouple the interaction between the ligand and the environment, so instead of computing directly $\Delta G^0_{binding}$ (that is in practice impossible) we compute ΔG^{solv}, ΔG^{prot} and get $\Delta G^0_{binding} = \Delta G^{prot} - \Delta G^{solv}$

2.3.1 Theory of Alchemical transformation

As introduced yet, in an alchemical transformation we want to compute the free energy difference between a state where the ligand and the protein interact with each other, and a decoupled state where the interactions are turned off. To achieve this aim we introduce a parameter λ that is 0 in the decoupled state (that later we will call state A) and 1 in the coupled one (called state B), so our hamiltonian will be function of this parameter $H = H(p, q, \lambda)$ (Trott & Olson, 2010).

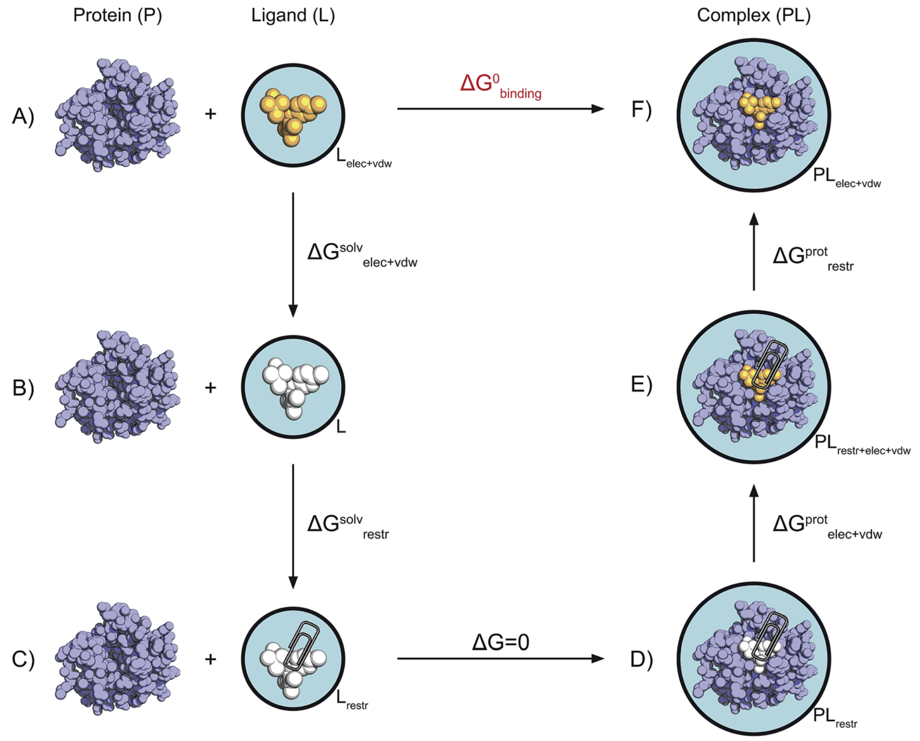

Figure 2.2: Alchemical Path

The Helmholtz free energy A is related to the partition function Q of an NVT ensemble, which is assumed to be the equilibrium ensemble generated by a MD simulation at constant volume and temperature:

$$A(\lambda) = -k_B T ln Q$$

$$Q = c \int \int \exp(-\beta H(p;q;\lambda)) dpdq$$

We will use the Gibbs free energy G, related to the partition function Δ of an NPT ensemble (the equilibrium ensemble generated by a MD simulation at constant pressure and temperature) which better represent the environmental conditions during an experiment than NVT ensemble:

$$G(\lambda) = -k_B T ln \Delta$$

$$\Delta = c \int \int \int exp(-\beta H(p;q;\lambda) - \beta pV) dpdqdV$$

$$G = A + pV$$

where $\beta = 1/(k_B T)$ and $c = (N! h^{3N})^{-1}$. These integrals over phase space cannot be evaluated from a simulation, but it is possible to evaluate the derivative with respect to λ as an ensemble average:

$$\frac{dA}{d\lambda} = \frac{\int \int \int (\partial H/\partial \lambda) exp(-\beta H(p;q;\lambda)) dpdq}{\int \int exp(-\beta H(p;q;\lambda)) dpdq} = \langle \frac{\partial H}{\partial \lambda} \rangle_{NVT;\lambda}$$

with a similar relation for $dG/d\lambda$ in the NPT ensemble. The difference in free energy between $\lambda = 0$ and $\lambda = 1$ can be found by integrating the derivative over λ:

$$G^B(p,T) - G^A(p,T) = \int_0^1 \langle \frac{\partial H}{\partial \lambda} \rangle_{NpT;\lambda} d\lambda$$

In Cartesian coordinates, the kinetic energy term in the Hamiltonian depends only on the momenta, and can be separately integrated and, in fact, removed from the equations. When masses do not change, there is no contribution from the kinetic energy at all; otherwise the integrated contribution to the free energy is $-\frac{2}{3} k_B T \ln(m_B/m_A)$.

2.3.2 Force Field

In this section we will describe the λ-dependence of the potentials used for free energy calculations. All common types of potentials and constraints can be interpolated smoothly from state $A(\lambda = 0)$, that represent the non bonded state, to $B(\lambda = 1)$, which is the bonded state, and vice versa. All bonded interactions are interpolated by linear interpolation of the interaction parameters. Non-bonded interactions can be interpolated linearly or via soft-core interactions.

Harmonic potentials The example given here is for the bond potential, however, these equations apply to the angle potential and the improper dihedral potential as well.

$$V_b = \frac{1}{2}[(1-\lambda)k_b^A + \lambda k_b^B][b - (1-\lambda)b_0^A - \lambda b_0^B]^2$$

Proper dihedrals For the proper dihedrals the equations are:

$$V_d = [(1-\lambda)k_d^A + \lambda k_d^B](1 + \cos[n_\phi \phi - (1-\lambda)\phi_s^A - \lambda \phi_s^B])$$

Coulomb interaction The Coulomb interaction, between two particles of which the charge varies with λ is:

$$V_c = \frac{f}{r_{ij}}[(1-\lambda)q_i^A q_j^A + \lambda q_i^B q_j^B]$$

Kinetic Energy When the mass of a particle changes, there is also a contribution of the kinetic energy to the free energy:

$$E_k = \frac{1}{2}\frac{\mathbf{p}^2}{(1-\lambda)m^A + \lambda m^B}$$

Lennard-Jones interaction For the Lennard-Jones interaction between two particles of which the atom type varies with λ we can write:

$$V_{LJ} = \frac{(1-\lambda)C_{12}^A + \lambda C_{12}^B}{r_{ij}^{12}} - \frac{(1-\lambda)C_6^A + \lambda C_6^B}{r_{ij}^6}$$

Soft-core interactions In a free-energy calculation where particles grow out of nothing, or particles disappear, using the simple linear interpolation of the Lennard-Jones and Coulomb potentials may lead to poor convergence. When the particles have nearly disappeared, or are close to appearing (at λ close to 0 or 1), the interaction energy will be weak enough for particles to get very close to each other, leading to large fluctuations in the measured values of $\frac{\partial V}{\partial \lambda}$.
Soft-core potentials V_{sc} are shifted versions of the regular potentials, so that the singularity in the potential and its derivatives at $r = 0$ is never reached:

$$V_{sc}(r) = (1-\lambda)V^A(r_A) + \lambda V^B(r_B)$$

$$r_A = (\alpha\sigma_A^6\lambda^p + r^6)^{\frac{1}{6}}$$

$$r_B = (\alpha\sigma_B^6(1-\lambda)^p + r^6)^{\frac{1}{6}}$$

where V_A and V_B are the normal “hard core” Van der Waals or electrostatic potentials in state $A(\lambda = 0)$ and state $B(\lambda = 1)$ respectively, α is the soft-core parameter, p is the soft-core λ power, σ is the radius of the interaction, which is $(C_{12}/C_6)^{1/6}$. For intermediate λ, r_A and r_B alter the interactions very little for $r > \alpha^{1/6}\sigma$ and quickly switch the soft-core interaction to an almost constant value for smaller r. Usually $p = 2$, and $0.001 < \alpha < 0.003$.

2.3.3 Multistate Bennett Acceptance Ratio

In practice, in order to compute the integral $\int \langle \partial H/\partial \lambda \rangle d\lambda$, we compute the free energy difference between contiguous λ states and we sum them all. This is done with a technique called BAR (Bennett, 1976), or his extension MBAR. MBAR is derived from a set of $K \times K$ weighting functions, $\alpha_{i,j}(\vec{q})$, that minimized the variance during the reweighting across the board. Starting from our core free energy equation, we have:

$$\Delta A_{ij} = -\beta^{-1} \ln \frac{Q_j}{Q_i}$$

and for any $\alpha(\vec{q}) > 0$ the following relation is true:

$$Q_i \langle \alpha_{ij} exp(-\beta U_j) \rangle_i = Q_j \langle \alpha_{ij} exp(-\beta U_i) \rangle_j$$

now we can write:

$$\sum_{i=1}^{K} \frac{\hat{Q}_i}{N_i} \sum_{n=1}^{N_i} \alpha_{ij} \exp(-\beta U_j(\vec{q}_{i,n})) = \sum_{j=1}^{K} \frac{\hat{Q}_j}{N_j} \sum_{n=1}^{N_j} \alpha_{ij} \exp(-\beta U_i(\vec{q}_{j,n}))$$

assuming we use the empirical estimator for the expectation values of $\langle g \rangle_i = N_i^{-1} \sum_{n=1}^{N_i} g(\vec{q}_{i,n})$ Choosing the optimal α_{ij} can be done by looking through the literature at extended bridge sampling. We then get an α_{ij} of:

$$\alpha_{ij} = \frac{N_j \hat{c}_j^{-1}}{\sum_{k=1}^{K} N_k \hat{c}_k^{-1} \exp(-\beta U_k)}$$

Finally we can get an expression for an estimated free energy of:

$$\hat{A}_i = -\beta^{-1} \ln \sum_{j=1}^{K} \sum_{n=1}^{N_j} \frac{\exp[-\beta U_i]}{\sum_{k=1}^{K} N_k \exp[\beta \hat{A}_k - \beta U_k]}$$

Here we have a single free energy, not a difference, because the free energies for a given set of states is only uniquely determined up to an additive constant.

2.3.4 Constraints

During the decoupling we annihilate the interaction between the ligand and the protein so if we don't use a set of constraint to keep the molecule in place it will move away, far from the binding site, leading to an incorrect result. The set of constraint used is a bond, two angles and three dihedrals. In total we will need three atoms from the protein and other three from the ligand. Choosing this set of atom is not trivial and a selection that include atoms with high mobility can generate artifacts in the energy distribution and lead to a wrong result. A wise choice is to select the protein atoms nearest to the ligand that are part of an alpha helix or a beta sheet. Compute the contribution of those constraints at the free energy in the system protein-ligand must be done numerically and require at least 10 λ-windows, while it can be performed analytically for the solvation system using the following formula:

$$\Delta G^{solv}_{restr_on} = RT \ln[\frac{8\pi^2 V^0}{r_0^2 \sin\theta_{A,0} \sin\theta_{B,0}} \frac{(K_r K_{\theta_A} K_{\theta_B} K_{\phi_A} K_{\phi_B} K_{\phi_C})^{\frac{1}{2}}}{(2\pi k T)^3}]$$

where: R is the ideal gas constant; T is the temperature in Kelvin; V^0 is the volume corresponding to the one molar standard state (1660 angstrong) r_0 is the reference distance for the restraints; θ_A, θ_B are the reference angles for the restraints; K_x is the force constant for the distance (r_0), two angles (θ_B, θ_B) and three dihedrals (ϕ_A, ϕ_B, ϕ_C) restraints we applied

Chapter 3

Methods Benchmark

In this chapter we will test the precision of the Alchemical transformation, and after we will compare those result with an other function score, Autodock Vina. Here we will use as test set a small group of ligands (showed in fig. 3.1 and 3.2) for the Bromodomain-4, proteins with versatile functions in the regulation of protein-protein interactions mediating gene transcription, DNA recombination, replication and repair and targeted in therapies for cancer and inflammatory disorders.

3.1 Test via Alchemical Transformation

3.1.1 Implementation

The pipeline followed to perform the Alchemical Transformation on a protein-ligand with GROMACS 5.0.7 (Abraham et al., 2015) system is this:

1. System preparation: starting from a crystal structure of the protein-ligand we generate the topology and coordinate files of the complex system and the only ligand in solution. The ligand's charges assignation is done with a software that perform quantum calculation using a semi-empirical hessians in order to compute the electron distribution. We choose three stable atoms from the protein and other from the ligand to set the constraints.
2. Set the λ-path and for each λ-window perform 10 ns of MD:
   - Decouple the electromagnetic interaction using 10 λ-windows;
   - Decouple the Van der Waals interaction using 20 λ-windows;
   - Turn off the constraints using 10 λ-windows (only for the protein-ligand system);
3. Analyze the energy distribution with the MBAR technique and compute the ΔG

Overall the whole procedure requires 700 ns simulation, which is done in 2 days on 5 nodes of a cluster (each node made off 4 GPU Nvidia Tesla P100 and one processor Intel Xeon with 44 cores).

Single-Precision	Double-Precision	PCIe x16 Interconnect Bandwidth	Compute Capability
9.3 teraFLOPS	4.7 teraFLOPS	32 GB/s	6.0

Table 3.1: Some data about the GPU used

3.1.2 Results

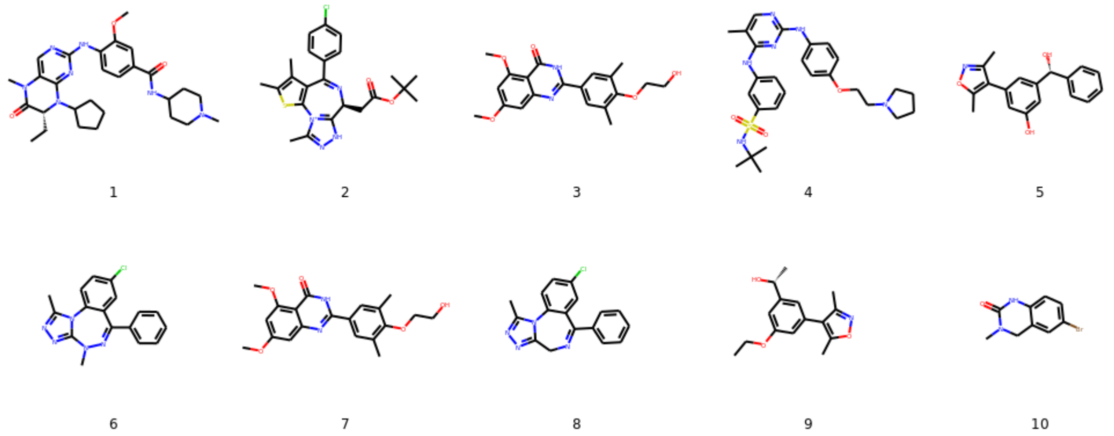

Figure 3.1: Chemical structure of the ligands

1	2	3	4	5
4OGI	3MXF	4MR3	4OGJ	4J0R
6	7	8	9	10
3U5L	4MR4	3U5J	3SVG	4HBV

Table 3.2: Ligands indexes and the pdb name of Bromodomain binded with them

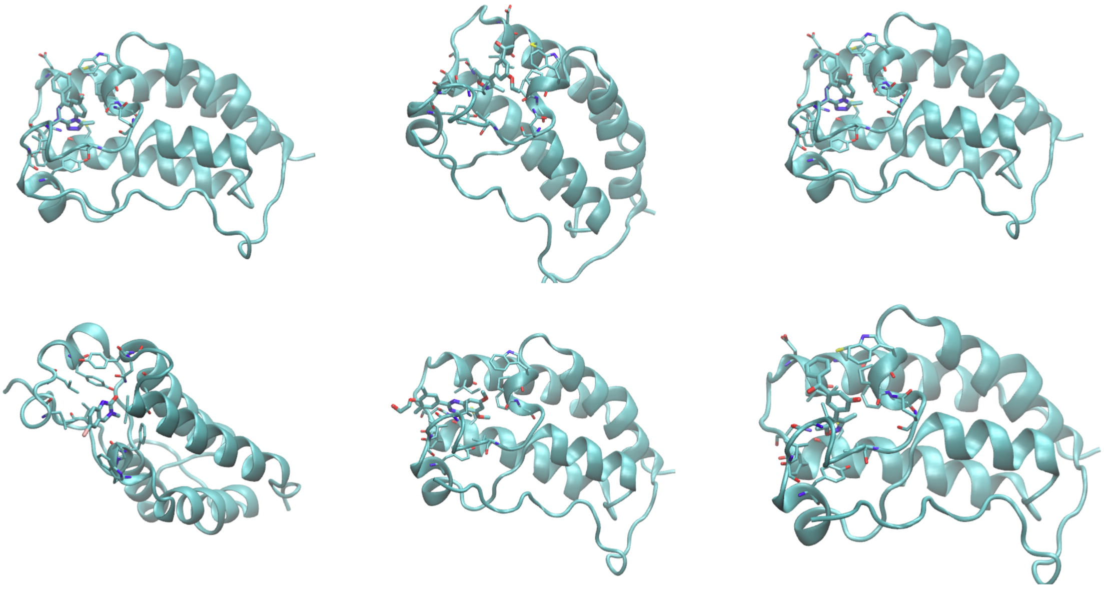

Figure 3.2: Bromodomain with different ligands. from the top left: 3SVG, 3U5J, 3U5L, 4HBV, 4J0R, 4MR4. Here we can observe how the different ligands interact with the protein

We report the results of our calculation in table 3.2 (for comparison we report also the results of same calculation obtained in Aldeghi et al. (2016)).

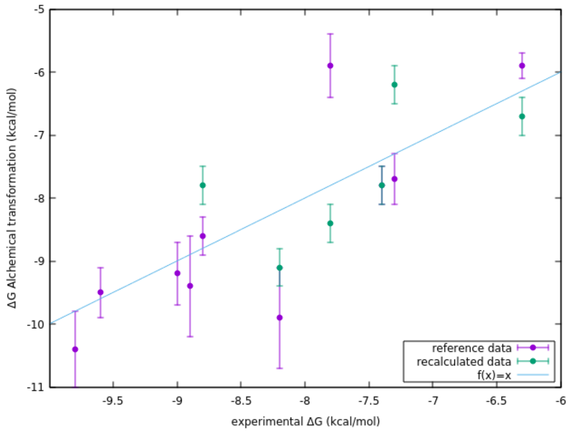

Figure 3.3: Alchemical Transformation results ($r = 0.8$ for the reference data and $r = 0.6$ for the recomputed data)

ΔG_{calc} [kcal/mol]	ΔG_{exp} [kcal/mol]	$\Delta G_{reference}$ [kcal/mol]	PDB name
	-9.8 ± 0.1	-10.4 ± 0.6	4OGI
	-9.6 ± 0.1	-9.5 ± 0.4	3MXF
	-9.0 ± 0.1	-9.2 ± 0.5	4MR3
	-8.9 ± 0.1	-9.4 ± 0.8	4OGJ
-7.8 ± 0.3	-8.8 ± 0.1	-8.6 ± 0.3	4J0R
-9.1 ± 0.3	-8.2 ± 0.1	-9.9 ± 0.8	3U5L
-8.4 ± 0.3	-7.8 ± 0.1	-5.9 ± 0.5	4MR4
-7.8 ± 0.3	-7.4 ± 0.1	-7.8 ± 0.3	3U5J
-6.2 ± 0.3	-7.3 ± 0.1	-7.7 ± 0.4	3SVG
-6.7 ± 0.3	-6.3 ± 0.1	-5.9 ± 0.2	4HBV

Table 3.3: Comparison of values computed (ΔG_{calc}), values from Aldeghi et al. (2016) ($\Delta G_{reference}$) and experimental values (ΔG_{exp})

As a comparison, we report the predicted binding energy calculated using Autodock Vina. The free energy calculation start from the crystal structure of the target protein and the SMILE of the ligand that we want to test, so a software find the flexible substructures of the ligand and a Montecarlo supported by a Genetic Algorithm try to dock it in position. A score is computed for each configuration with the formula previously introduced and the best score will be the free energy estimation.

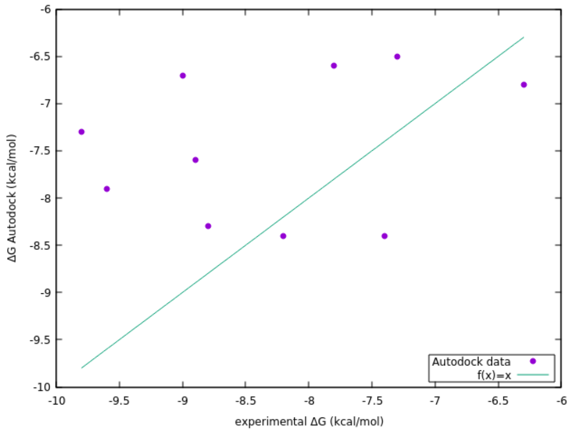

Figure 3.4: Autodock Vina results for the comparison with the Alchemical Transformation ($r = 0.2$)

As we can see the precision of this function score is worse than Alchemical transformation. However it can be used for a fast screening of hundred of compounds. To show this we will compare the docking score and the measured affinity of couples ligand-targets previously chosen. The K_D experimental values were obtained from the 'The Binding Database' (https://www.bindingdb.org). Since the number of experimental results of binding energy on the bromodomain is limited, in order to improve the statistical meaning of this test, we decided to include several other known targets, even though this will create non-homogenous testing set. The data set is composed by the following proteins:

1. Matrix metalloproteinase 12 (MMP12)
2. Bovine trypsin-inhibitor
3. Humanised monomeric RadA
4. HIV-1 Protease (with mutations outside the Active Site)
5. HIV-1 Protease NL4-3 L90M Mutant
6. HIV-1 Protease NL4-3 V82F Mutant
7. Human carbonic anhydrase isozyme II
8. KDM5A Jmj Domain
9. Human Galectin-3 CRD
10. First bromodomain of human BRD4

Those are relevant biological targets studied for their role in cancer or inflammatory processes. For example HIV-1 protease is an enzyme that cleaves proteins to their component peptides, that is essential for the life-cycle of HIV-1 virus. Galectins are a particular kind of proteins that can bind very long chains of sugar but not the shortest one; this peculiarity is very useful to diagnostic cancer, indeed tumor cells generate long sugar chains so a blood test can reveal the presence of cancers even in the earliest stages.

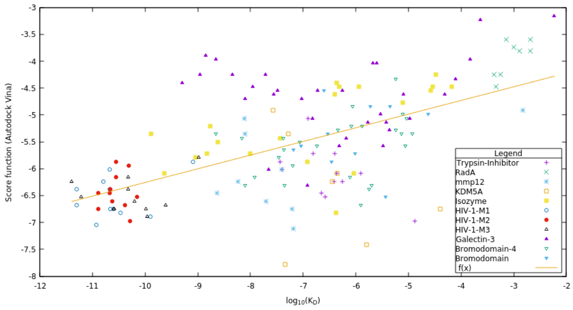

Figure 3.5: Autodock Vina results for the chosen test set ($r = 0.60$)

As we can clearly see from the data correlation this method is useful for a very fast screening but inadequate to discern between ligands whose K_d differs by less than 10^3. One of the most problematic aspect of Structure-based methods is their inability to take in account conformational changes of the protein after the ligand is docked. Other issues derive from the omission of hydrogen bonds with water molecules that can mediate the interaction between two near atoms.

3.2 Conclusions

The purpose of this thesis work was to test alchemical transformation method for computing binding free energy in currently available high performance computing clusters, and in particular to test whether such method can be used for high throughput screening. We set up the pipeline for performing the required molecular dynamics simulations and reproduced published results within the errors. Our benchmarks show that, in order to reach the desired convergence of simulations, it is necessary to follow the dynamics of the complex formed by the protein and the ligand for 700 ns overall. For a system like the bromodomain, which contains 56 000 atoms (if we consider the water molecules), this translate in 48 hours in the SIAIS cluster with specifics reported in table 3.1. Therefore, it is foreseeable to test up to several tens of molecules for a given project. Despite being remarkable higher than what was feasible only few years ago, this number is still much smaller than the hundred of thousands or millions tests required to have a proper preliminary screening. A workaround strategy we propose is to use a less precise, but faster screening method, such as Autodock Vina, to obtain a first screening, and then use alchemical transformations to be able to predict experimental results and produce a preliminary list of promising candidates that can be further analysed experimentally.

Bibliography

David L. Mobley, Pavel V. Klimovich, Perspective: Alchemical free energy calculations for drug discovery, The Journal of Chemical Physics, vol. 137,23 (2012), doi: 10.1063/1.4769292.

Ghosh AK, Osswald HL, Prato G. Recent Progress in the Development of HIV-1 Protease Inhibitors for the Treatment of HIV/AIDS. J Med Chem. 2016 Jun 9;59 (2011):5172-208. doi: 10.1021/acs.jmedchem.5b01697. Epub 2016 Jan 22. PMID: 26799988; PMCID: PMC5598487.

Visvaldas Kairys, Lina Baranauskiene, Migle Kazlauskiene, Daumantas Matulis & Egidijus Kazlauskas, Binding affinity in drug design: experimental and computational techniques, Expert Opinion on Drug Discovery (2019), vol. 14:8, 755-768, doi: 10.1080/17460441.2019.1623202

D. Cronk, Chapter 8 - High-throughput screening, RG Hill, HP Rang, Drug Discovery and Development (Second Edition) , Churchill Livingstone (2013), Pages 95-117, ISBN 9780702042997, doi: 10.1016/B978-0-7020-4299-7.00008-1

Favalli, Nicholas et al. DNA-encoded chemical libraries - achievements and remaining challenges. FEBS letters vol. 592,12 (2018): 2168-2180. doi: 10.1002/1873-3468.13068

Brenner S, Lerner RA. Encoded combinatorial chemistry; PubMed Central; vol. 15;89 (1992):5381-3. PMID: 1608946; doi: 10.1073/pnas.89.12.5381.

Dario Neri, Twenty-five Years of DNA-Encoded Chemical Libraries, ChemBioChem, vol. 18,9 (2017): 827-828, doi: 10.1002/cbic.201700130

Trott, Oleg and Olson, Arthur J., AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading, Journal of Computational Chemistry, vol. 31.2 (2010): 455-461, doi: 10.1002/jcc.21334

M.J. Abraham, D. van der Spoel, E. Lindahl, B. Hess, and the GROMACS development team, GROMACS User Manual version 5.0.7, www.gromacs.org (2015)

Charles H. Bennett, Efficient Estimation of Free Energy Differences from Monte Carlo Data, Journal of Computational Physics , vol, 22:245-268 (1976)

Aldeghi, Matteo and Heifetz, Alexander and Bodkin, Michael J. and Knapp, Stefan and Biggin, Philip C., Accurate calculation of the absolute free energy of binding for drug molecules, Chem. Sci. vol. 7.1 (2016):207-218, doi: 10.1039/C5SC02678D