Predicting the mutagenecity of DNA intercalators through binding affinity calculations

Kristof Farkas-Pall

August 22, 2018

Contents

Chapter 1

Introduction

1.1 DNA

1.1.1 DNA intercalation

Shortly after the unraveling of the complex structure of the double stranded DNA that self-assembles into a helical form started the exploration of ligands that form complexes the DNA in a diverse set of binding forms. Certain molecules could bind to DNA grooves, or intercalate between DNA basepairs. Of these two main categories of binding, the former is less of interest or studied. Intercalators on the other hand have a wide range of interest and use cases, including cancer treatment, due to their varied binding properties to DNA. The existence of intercalators was first predicted and later confirmed by Lerman et al. whereby a small planar aromatic molecule inserts between consecutive DNA-base pairs, perturbing the local structure of the DNA, but still having a stabilising effect through base pair stacking. Intercalation can potentially explain, and aid in understanding a number of DNA specific phenomena including transcription, mutation, and the mechanism of certain anti-tumour agents, like the anti-cancer agent small molecule acridine. Intercalator can be neutrally charged or cationic, they disrupt the continuous and predictable flow of the genetic information encoded in the base-pairs. In contrast other forms of DNA binding, major or minor groove, or electrostatic binding, does not deform the structure, and has a smaller effect of the function of DNA. Intercalation greatly changes the structure of DNA, including the base pair distance, decreasing the twist angle by up to 50%, essentially unwinding the DNA to some extent in the local neighbourhood of the intercalation site. In order to accommodate the molecule, the inter-basepair distance has to be increased, sometimes up to double its original size. The increase in separation elongates the whole DNA helix. Moreover while the DNA helix unwinding compensates for the insertion maximising the base stacking and hence the stability, the backbone structure, for example the distance between consecutive phosphate groups remains nearly identical. This level of perturbation and the fact that it is reversible (as the

binding is non-covalent) provides a basis for a wide range of applications.

The importance of intercalators in drug design and the potential application of these to treat a wide range of cancer has already been shown by certain compounds and FDA approved treatments. Two similar compounds, doxorubicin and daunomycin, are currently used as anti cancer treatment for more forms of cancer than any other compound. Intercalation of these compounds disrupts replication and transcription of DNA and leads to the death of cancerous cells. Once the crystal structure of this complex was determined with X-ray diffraction methods, the mode of action was reavealed to be that of intercalation, as the fused ring system is indeed sandwiched between basepairs. The mechanism was theorised to be a ring-insertion mechanism with the inter-basepair distance increasing, but certain structural features staying the same for example the hydrogen bonding.

The basepair distance opening and molecule insertion nonetheless is too simple of mechanism, and certain high quality experimental measurements indicate a more sophisticated binding mechanism. For example, kinetic measurements indicate a three step intercalation process, with the first step being a fast binding to the outside of the DNA chain, and the second and third slower conformational adjustments. However the experiments are low resolution and the hypothesis of the intercalations is not fully elucidated. Certain computational techniques have been used to investigate this mechanism. Metadynamics and umbrella sampling can provide atomic-scale resolutions and free energy estimates of the intercalation mechanism. They confirm the earlier studies, that the process is composed of three steps, an initial binding in the minor groove of the DNA, followed by the rotation of the drug and DNA deformation into a seconds intermediate metastable state, and finally separated by only a small barrier complete rotation of the molecule into the space between basepairs to establish a full basepair stacking interaction.

1.1.2 Structure

1.1.3 Mutations

1.2 DNA modelling and Molecular dynamics

1.2.1 Atomistic modelling

The structure of biomolecule, like DNA is complex and interesting. How can we gain insight into the atomic composition of structures like this on an atomic level? Experimental techniques, like cryo-em, x-ray crystallography or NMR provide information about the composition at a resolution small enough to distinguish atoms, and some methods also provide dynamical properties too. However, certain phenomena or physical property is simply out of the reach of these experimental techniques. In silico techniques, such as molecular dynamics, can provide insight otherwise not accessible to experimentalists. The focus of

this thesis is applying such in silico methods to bio-molecular systems, with a particular emphasis on molecular dynamics.

Nonetheless, as a starting model, even in silico methods take data from experiment, therefore it is briefly described here.

1.2.2 Experimental models

The structure of DNA and other biological molecules can be determined in a number of ways including X-ray, NMR and Cryo-EM. Unfortunately, it is not always possible to determine the structure of the desired system under investigation experimentally. In that case it might be the case that we can use our knowledge of similar structure and modelling to generate new structures that are likely to resemble reality. Here we briefly describe these techniques.

X-ray crystallography

The most notable method, especially from a historical point of view, to analyse the structure of DNA is by X-ray diffraction. This method lead to the discovery of the basic double helical structure of the DNA strand. The resolution is determined by the wavelength of the radiation used to investigate the structure. In case of biologically relevant molecules we want to distinguish between atoms, therefore a sub 1 angstrom resolution is required. This resolution corresponds to a wavelength of X-ray radiation. To produce a high quality image a regular array of units to scatter the radiation is required. This means that we have to crystallise the molecule, which places it into a state somewhat different from the biological conditions it was in originally.

1.2.3 Computational modelling

Simulating a system via computational methods requires an initial model, where by model, we refer to the coordinates of every particle in the structure of interest. If, for example, we want to simulation nucleic acid structures, then the model is the three dimensional coordinate of every atom in the DNA structure. In general the model can be build in one of two ways: by hand de novo, or using another source (experimentally derived structures for example). The quality of the inital starting structure can have a large effect on the computational results and the validity of the simulation. This will become apparent in the results section, where two similar systems were simulated with different levels of initial structure validity.

QM

In a quantum mechanical description of a system, the electrons are explicitly accounted for as they are the fundamental particles of this level of description. The energy of the system is calculated as the interaction energy between electrons based on electronic structure calculations. There are few (if any) empirical

or fitted parameters in these models, but the underlying physical description is sometimes approximated for tractability. Usually QM calculations are very expensive and scale cubic (or worse) with the number of electrons in the system.

MM

In molecular mechanics the description is based on particles corresponding to atoms. Every atom in the system has a number of parameters defining the interaction between other particles, including a partial charge, bond, angle and torsion parameters, and non-bonded interaction parameters. These parameters are all empirically determined (fitted to experimental observables, quantum mechanical calculations, or other data). MM simulations are faster than quantum calculations, hence why this study was conducted using MM methods to simulate our DNA based systems. Some of the limitations of MM methods, including lower accuracy and the lack of the ability of bond rearrangement was considered, and determined to be a non issue for the systems at hand. Furthermore, QM simulations are too expensive to run for adequate amount of time (nanoseconds of simulations) that is required to reach and sample the equilibrium state. The size of the system is also an important decision factor in choosing between QM and MM. Roughly speaking, QM simulations have on the order of a couple of hundreds of atoms, while MM simulations are routinely run with tens of thousands of atoms.

1.3 Theoretical details of molecular simulations

Simulating molecular systems is, at a high level, very simple and straight forward. First, build a particle based description of you systems, where particles are the only entities that can interact with each other or the environment. Then, based on rules, propagate the system through time starting from some initial state (as described before the state of a system is the position and velocities of every particle) for a given number of steps or iterations. The rules by which you propagate the system can be either deterministic or random. As you build up a collection of consecutive states of the simulated system, certain properties can be calculated per frame (frame being a state in the context of a simulation) or averaged over multiple states to get ensemble estimates over time.

There are two main categories of molecule simulations based on the nature of the propagation rules: Molecular Dynamics (MD) or Monte Carlo (MC). Molecular dynamics refers to the numerical integration of the equations of motion. They are used to investigate thermodynamic properties as well as structural and dynamical phenomena of the system. Monte Carlo type simulations on the other hand, generate new states based on the current state via probabilistic rules. The set of states generated this way can be used to predict thermodynamic and structural properties, but *not* dynamical ones. This is because a MC simulation has no concept of time by which states are ordered and evolved.

1.3.1 Equation of motion

1.3.2 Forcefield

The molecular mechanics model described in section **??** is based on a number of interatomic potential energy functions. All these potential energy functions have a number of constants, the values of which are fitted to higher level calculations or experiment.

In order to have credibility in the theoretical models that are used to predict experimental measurements, or more drastically, substituting them, we have to validate the force fields that are used in the molecular dynamics simulations. In particular, the last generation of force-fields for DNA systems specifically has raised questions about their validity especially at longer simulation timescales. It was found that while not all theoretical simulations are equally reliable, the latest generation of force fields of the AMBER family designed to model DNA (BSC1 and OL15) can be safely used to reproduce the global structure of DNA, and predict observable quantities of interest, like the free energy of binding. They provide the most accurate results when comparing MD trajectory simulation data to experimental observables like NMR structure of X-Ray resolved structures. Furthermore, these forcefields reproduce experiments in an unbiased environment, and a common technique to force DNA trajectory to align with experimental data, termed restrained MD, is not essential for this study.

1.3.3 Main steps in a molecule dynamics simulation protocol

Every system that is investigated via simulation method required specific care, but there are four steps that are present in most molecular dynamcs workflows. They are not hard written rules, but following these steps assures that the system is correctly equilibrated and sampled during simulations, and also minimizes the chance that the energy of the system gets out of control.

The four steps are: system preparation, minimisation, equilibration and production.

Minimisation

During minimization or relaxation, we want to find a local minimum of the biological system so that during the start of the molecule dynamics simulation the system does not end up in an unstable state (sometimes termed “blow up”) This can happen due to forces on certain atoms being too large at the start, and they were to move too much during the first couple of timesteps of the simulation. To achieve this standard minimisation algorithms are employed, sich as steepest descent or the fire minimiser.

Assignment of velocities

As discuss before, minimisation takes the system to a state where we can start the numerical integration of the equation of motion without the possibility that any subsequent displacement will be too large or physically unrealistic, however, minimisation only takes into account the positions of every particle, and does not change the velocities, therefore we have to set the velocity vectors of every particle separately at the start of the simulation. Thus, to set the starting velocities one usually assigns random initial velocities to atoms in a way such that the correct Maxwell-Boltzmann distribution at the desired temperature is achieved as a starting point. The actual assignment process is typically unimportant, as the Maxwell-Boltzmann distribution will quickly arise naturally from the equations of motion. The last particle in the simulation box is typically given a velocity such that the centre of mass momentum is zero, and the simulation box itself as a whole does not drift with time. This is because in Newtonian dynamics the momentum of the centre of mass is conserved.

Single simulation trajectories are somewhat meaningless, and in most cases, we want to run a set of independent simulations, of different replicase or realisations of the a particular system. This is to assess the error on the particular observable, and collect statistical descriptors like the spread, mean and standard deviation, or help understand how the starting structure effect the results. It is important to note, that even the smallest difference in the inital conditions of the system (for example the position of a single atom, or magnitude of velocity) can lead to completely divergent results after some simulation time, meaning that simply running different simulations starting with different initial velocities will lead to completely different time evolution over a long enough simulation. An even better approach to generating different initial replicas, is to start from completely different conformations such as different conformations of the molecules simulations, different binding pose in a receptor ligand simulation, as this will lead to divergence right from the beginning of the simulation.

Equilibration

In order to compare the data generated by simulation to experiment, we usually want to bring the simulation conditions to experimentally comparable values (including, temperature, pressure, energy) which requires simulations in a specific ensemble (NVE, NPT, etc.). Data collected should not be biased by the initial conditions on the simulation either. Therefore we need to run the simulations for some period of time until the the system is at the appropriate state, as well as relaxed from any metastable state introduced by the starting state of the simulation. In other words we are interested in sampling the most likely, equilibrium state of the system in the given ensemble. In case of biomolecules it is important to bring them to a configuration that is relevant and could take a long simulation time to reach such state.

The first step in the equilibration phase is to bring the system to the desired thermodynamic state. By thermodynamic state we mean a given temperature

are pressure. Assigning velocities with the correct distribution does set the inital temperature, but the thermostat will still have to partition the kinetic and potential energy of the system correctly. To this end, a thermostated equilibration period is run before collecting data even in the case of the NVE ensemble. To monitor the equilibration phase, we look at the temperature, pressure, kinetic and potential energy of the system over time and monitor their fluctuations. A simulation is siad to have equilibrated when these parameters reach a stead state and only fluctuate around that value with minial drift. This definition while not perfect, is a good measure of equilibration of a system to a given pressure, temperature, simulation box (i.e. volume) or energy.

Other properties of the system of interest should not change too during simulation. This is harder to monitor then the thermodynamic properties discussed above. Even at equilibrium the system might undergo slow fluctuations especially if it has slow degrees of freedom, but the values should stay around a specific value. The bio-molecular system it is common to investigate the root mean squared deviation of the particles over time. and potentially other properties like the number of hydrogen bonds between the biomolecules present and water, as these may be slower to equilibrate than system-wide properties like the temperature and pressure.

Once the energy component and other properties fluctuate around constant values with minimal or no drift we have reach the end of equilibration. If some of the properties of interest exhibit a systematic drift with time, that the system has probably not equilibrated sufficiently.

The target ensemble introduces a difference in the equilibration protocol. In the NVE ensemble the thermostat should be removed and frame of the system should be selected that is as close to the desired kinetic and potential energy as possible. The frame containing the positions and velocities of every particle is used to start the production simulation at the correct temperature. This is because the temperature fluctuates over the simulation when coupled to a thermostat.

If the target is the NPT ensemble (that is the most common for biological simulation) then the system should first be relax to the desired temperature in the the NVT ensemble, then the barostat turned on to relax the simulation box to the desired pressure.

Production

Once the system has equilibrated we can step into the final stage where we collect the data about the system. This phase of the workflow is called production. The main difference between the end of the equilibration and the production run is that in production we keep and analyse the data produce as opposed to discarding it. Production always comes after a rigorous equilibration phase in the correct ensemble, and it should never be collected after a sudden change of any of the properties of the simulation, like the simulation box size, temperature, pressure (unless this is the goal of the simulation).

If the ensemble changes from equilibration to production, then it is advised to

discard the initial data collected even in the production simulation (for example switching from NPT to NVT) with the usual considerations about equilibrium still applicable here.

Analysis of the production work required special care as some of the observable are not trivial to calculate, and the error on these values must be correctly estimated. Usually, analysis involves calculating the expectation values of certain observables, with the important consideration that these values are converged, meaning that they no longer depend on the further length of the simulation (or replica of simulations) or on the initial conditions that the simulation was started from. This is very much related to the equilibration discussed above, with the difference being what is the subject at hand. Depending on the time to relaxation, it is common to realise that the system was not equilibrated after analysis of the production run.

A key consideration is the storage and frequency of storage of the simulated data. Storing data very frequently (for example storing at every timestep) can be tempting, but limits in the physical availability of storage space is usually a limiting factor, but also the additional information contained in snapshots close together diminished drastically as the time separation become small. In particular in molecular dynamics simulations consecutive frames are correlated, so storing and analysing each frame is essentially redundant and adds no value or insight to the calculation of a specific observable. Therefore, storing data more frequently than the autocorrelation time is not necessary. Disk space should also be taken into consideration. Depending on the size of the system (the number of particles) the size of a single frame in a trajectory can be large. But even if disk space is not an issue, storing snapshots with an autocorrelation frequency should be enough.

There are various strategies to reduce size of the data. One can store reduced precision of the coordinates, or chose to save with different frequency the component of interest. For example because energy is much smaller to store, it can be storied with a higher frequency then trajectory coordinates. This depends on the analysis at hand.

1.3.4 Thermostat

Thermostats are ofter used in molecular dynamics simulation to control the temperature. MD simulations are used to understand, observe and predict the properties of interest for a given system. If we want to compare to results done in laboratory conditions (where temperature is usually constant) we have to sample the canonical ensemble. Therefore the temperature of the system must be maintained at a certain level, and a thermostat should be used.

To determine the temperature of a molecular dynamics simulation we can use the equipartition theorem involving the kinetic energy of the system

$$\frac{3}{2}Nk_bT = \langle\sum_{i=1}^{N}\frac{1}{2}m_iv_i^2\rangle \tag{1.1}$$

The temperature is a time averaged quantity indicated by the brackets, meaning that it is an average over many snapshots of the simulation. Applying the equipartition theorem to just a single snapshot of the simulation, we get the instantaneous temperature. The instantaneous temperature is usually not equal to the desired ensemble temperature, but it should fluctuate around that value, the target temperature.

In general, thermostat work by altering the Newtonian equations of motion, which inherently conserve the energy. Therefore one should not use a thermostat if the desired property is dynamical (for example the diffusion coefficient); but instead, the thermostat is used to bring the system to the desired temperature then turned off. While thermostats give non-physical dynamics, some have less effect on particular dynamical properties and are save to use in molecular dynamics simulations.

There are two ways to categories thermostats: (i) they can be either deterministic or stochastic depending on whether they use random numbers to alter the equations of motion or (ii) global or local based on the extent of the effect, whether its on the full system or just a subset of the particles. Thermostats also differ in the algorithm used to control the temperature. Certain thermostats operate by rescaling the velocities of the particles outside the dynamics, by applying the change after a simulation timestep has happened and the coordinates and velocities have been updated. Others include implicit or explicit collisions with an external heat bath.

1.3.5 Barostat

Thermodynamic properties of interest are measure inside laboratory condition, therefore, at short timescale the system is under constant temperature *and* pressure. The pressure is usually at atmospheric pressure, unless some form of pressure control is used, for example a piston, inert gas, etc. This set of conditions is part of the isothermal-isobaric ensemble, probably the most common ensemble in molecular dynamics simulation, especially that of biologically relevant systems. Similar to thermostat, if the pressure in the system needs to be constant, a barostat algorithm should be used.

Barostats control the pressure if the system, so if the aforementioned isothermal-isobaric ensemble is desires, that they must be used in conjunction with a thermostat. Without one, we would be in a isoenthalpic-isobaric ensemble, where only the number of particles, the pressure and the enthalpy of the system is constant.

The background if barostat is analogous to that of thermostats. The virial theorem is used to measure the pressure in a molecular dynamics simulation. In the case of periodic boundary conditions and pairwise interactions, other equations are considered. Nonetheless, these formulas also provide pressure as a time averaged quantity over the simulation range. The pressure calculated just from a single snapshot is termed the instantaneous pressure. Similarly to temperature, the instantaneous pressure is not equal to the target, but fluctuates constantly around that value during the simulation.

To model the idea of a barostat, consider a system that is compressed/expanded by a hypothetical piston from every direction with a uniform force. Since it is acting from all directions it can be considered that the compression is uniform. The mass of the piston can be changes to correspond to stronger compression force. This changes the frequency with which particles interact with the system enclosure. These impacts on the simulation box can be related to the stress the surrounding is applying to the system.

1.3.6 Integrators

Any system with more than three interacting bodies has a solution to their equation motion that cannot be analytically derived. Instead, approximation have to be introduced to the dynamics of a system by numerical methods in a discrete manner, termed numerical integration of the equations of motion. The algorithms that perform such numerical integration are called integrators. A key criteria for any integrator is energy conservation at short and long timescales.

There are certain feature that integrators should have in order to be effectively used in molecular dynamics simulations. To start with, the integrator should introduce little error in the dynamics of the system. By definition integrators use discretisation to approximate the dynamics of the continuous motion, therefore they will introduce an error because of this approximation. Alongside discretisation errors, an additional source of error comes from truncating the numbers as they are stored in computers to a fixed precision. There are many strategies to reduce discretisation error, while truncation error can be reduced by increasing the precision of the floating point numbers used by the computer code.

By minimising the discretisation error, we want to minimise errors in the phase space volume and conserve the energy during simulation. If phase space volume changes over the simulation, then the system is, by definition, is a different ensemble from one timestep to the next. This means that the data collected during simulation is from different ensembles. One simple solution is to guarantee that the integrator is reversible. If the operator that propagates the system preserves phase space, then by definition is also reversible, in other words the operator is applied to a state stepping δt forward, and then applying the same operator $-\delta t$ step, and the system remains the same, than we have reversibility and hence phase space volume conservation.

1.4 Free energy calculations

1.4.1 Kinetic interpretation

Planar aromatic molecules containing a fused ring system scaffold can intercalate between consecutive DNA base pairs. They are termed intercalators. Whilst there are cases when intercalators form covalent bonds with the DNA base pairs this is an irreversible process. If the bonding is facilitated via non-bonded

interactions only (π stacking), then the interaction is reversible. The study here is only concerned with the later intercalation type, as this is the more common form and the tools used (molecular mechanics) are suited to model this type of interaction only. In a solution containing double stranded DNA fragments and the intercalator, they will be constantly interacting, with the intercalator binding and unbinding from the DNA. This process can be modelled by a reversible chemical reaction

$$A + B \rightarrow AB \tag{1.2}$$

where the rate constant of the forward and backward reaction are the intercalation and de-intercalation reaction rates respectively. In equilibrium the two rate equal, therefore the concentration of each component (DNA, intercalator, and complex) is constant. The association constant K_a is determined by the concentrations at equilibrium, and is given by

$$K_a = \frac{k_1}{k_{-1}} = \frac{[AB]_{eq}}{[A]_{eq}[B]_{eq}} \tag{1.3}$$

1.4.2 Thermodynamic interpretation: the binding affinity

One can interpret the process of intercalation as the reaction to minimise the appropriate thermodynamic potential for the given ensemble. The most common ensemble for the experimental conditions is the NPT ensemble, because the number of molecules, pressure and temperature are constant. The associated potential if the Gibbs free energy, G, given by

$$G = U + pV - TS = H - TS \tag{1.4}$$

The intercalation will happen spontaneously if the difference in free energy (ΔG) between the intercalator and DNA in free state and the complex is negative. The intercalation then will happen until the free energy is minimised and equilibrium has been reached. This difference can be calculated as

$$\Delta G = G(complex) - G(DNA) - G(intercalator) \tag{1.5}$$

the numerical value corresponding to the strength of binding, with more negative values being stronger binders. The change in free energy can also be related to the aforementioned equilibrium constant K_a via the van't Hoff equation

$$\Delta G = -RT \log K_a \tag{1.6}$$

This means that the temperature and the free energy difference are closely related. If the temperature increases then the concentration of the complex also increases at the equilibrium. The value of ΔG is often called binding affinity or free energy of binding. The two terms can be used interchangeably. For certain use cases it is meaningful to break down the free energy difference into an enthalpic and entropic component

$$\Delta G = \Delta H - T\Delta S \tag{1.7}$$

1.5 Computational methods for free energy calculations

The aim of this thesis is to develop and use computational methods to calculate the free energy change of intercalation of a molecule between DNA base pairs. To compute the free energy this way offers a number of advantages over investigating the system experimentally. This is because it hard or impossible to study these systems at an atomistic level. Molecular simulations provide this level of detail, but in order to understand the macroscopic properties (like the free energy) we have to use statistical methods to derive them from the microscopic simulations.

1.5.1 Sources of error

Modelling the system and running simulations, just like experimental techniques, contain sources of error. There are two major sources or inaccuracy in MD simulations: phase space sampling, and force field inaccuracy.

Incomplete phase space sampling

In order to get reliable and converged results a large part of the phase space has to be sampled. The phase space of large biomolecule, like DNA and proteins, are large, and simulations of the order of a tens of nanoseconds cannot completely explore.

Force field accuracy

Chapter 2

Methods

2.1 Experimental measurements of intercalation

(for example through displacement assays (move to experimental section, experimental difficulties, why is it hard to measure DNA intercalator biding affinity)). Here we looked at a congeneric series intercalators based on a common quinoxaline scaffold with experimental measurements [].(Move to methods)

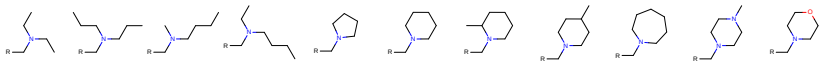

Figure 2.1: A congeneric series of DNA intercalators. All molecules share a common quinoxaline scaffold. The experimental study was conducted by [] producing binding affinity values that we could compare our calculations to. (Mention it in the experimental section)

2.2 Docking and scoring

For certain application, like that of high throughput screening in the pharmaceutical industry, it is important to assess the binding strenght of a large number of drugs or molecules bound to large biological macromolecules like proteins or DNA in a very fast manner. This is done via easily evalutable scoring function, some notable examples including: AutoDock, X-Score, DrugScore, ChemScore, GOLD, FlexX, LigScore and LUDI. Generally speaking, they take into account a single structure and little or no protein movement is taking into account during evaluation. One of the simplest scoring function consists of the empirical surface-area based method that shows that ligand binding reduces the surface are in the receptor that is accessible to the surrounding solvent. The function makes the assumption that the polar and non-polar area on the surface of the

molecule has a linear relation to the free energy. Due to the crude simplifications in these models, they have weak predictive power and are rarely used for accurate free energy estimation. Here, a scoring methods is used for two reasons: (i) to find an initial structure for the intercalator-DNA complex, and (ii) to compare the scores and hence the ranking of this methods with other, more accurate models. Docking and scoring has its place in the drug discovery pipeline. In high throughput scenarios, where speed is an important factor due to the large number of molecules that need to be tested, scoring functions are often used to filter out candidates that have low probability of being good binders.

2.2.1 Starting structures for docking

2.3 Molecular Mechanics Poisson-Boltzmann Surface Area

Molecular Mechanics Poisson-Boltzmann Surface Area (MMPBSA) is one of the theoretically approximate methods to calculate the free energy of system. It is the most accurate of the approximate methods, and it has previously been applied successfully to the free energy calculation of various biological systems including protein-ligand and DNA intercalator complexes. This method is a good choice for comparing and ranking a set of molecule by their binding affinities for two main reasons: MMPBSA is able to handle a wide variety of systems and can the results can be theoretically calculated from a a single trajectory, unlike theoretically exact alchemical methods like FEP or TI.

MMPBSA has become the a popular method to calculate binding affinities because of a balance between the details included in the physical model and the speed of the calculations themselves. As one of the main methods used in this thesis to calculate DNA-intercalator interactions, it will be discussed in more detail.

Applying the MMPBSA method to a system involves calculating the absolute free energy of binding by calculating three separate components: the free energies of the DNA-intercalator complex, the DNA and the intercalator molecules alone. The change is free energy upon intercalation then is

$$\Delta G = \langle G_{complex} \rangle - \langle G_{DNA} \rangle - \langle G_{intercalator} \rangle \tag{2.1}$$

where $\langle \dots \rangle$ represent the average value of a post-processing calculation performed on every frame of the simulation trajectory. Simulations are performed in explicit solvent and with counter-ions to balance the charge in the periodic simulation box. To aid faster convergence of the free energy values the MMPBSA methods employs two strategies: first, all solvent molecules and counter-ions are replaced from the trajectory post simulation but pre-analysis by a continuum solvent approximation, and a thermodynamic cycle is used.

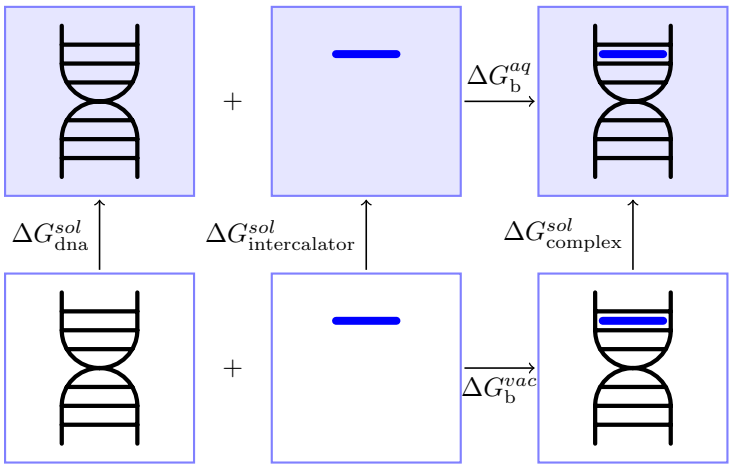

Figure 2.2: Thermodynamic cycle

2.3.1 The thermodynamic cycle

2.3.2 Single and Component trajectories

The different contributions in equation 2.3 can be extracted from a single trajectory of the complex, or from separate trajectories of the DNA, intercalator and complex. If a single trajectory is used, then the during the calculation of the different component, the rest of the system is removed and the analysis is done on part of the trajectory. This approach is effective in most scenarios and saves the additional cost of running multiple simulations for the separate components. The other advantage of the single trajectory approach is that part of the complex that do not affect the intercalation cancel out exactly, as the coordinates are taken from the same trajectory, hence they are exactly the same. If separate trajectories are run for each component, than this error cancellation does not occur, as the component are free to explore different conformations. On the other hand, a single trajectory will not be able to capture dynamical changes to either the intercalator or, more importantly to the DNA, upon intercalation, and will consequently ignore certain contributions to the binding free energy. In the rest of the thesis, the three possible trajectory approaches are termed 1, 2 and 3 trajectory, corresponding to MMPBSA calculations where all three components are from the same simulation (1 trajectory), the intercalator is from a separate simulation (2 trajectory), or all three components, including the DNA are from three separate simulations (3 trajectory).

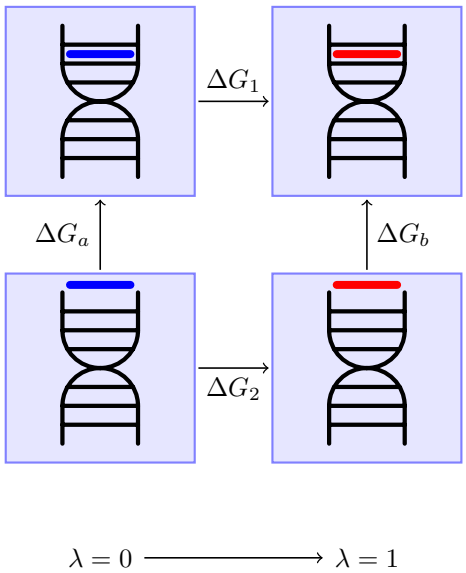

Figure 2.3: Thermodynamic cycle

2.4 TIES

2.4.1 The thermodynamic cycle

2.4.2 Large scale adaptive binding free energy calculations

To enable the running of large amounts of simulations on supercomputers we developed the high throughput binding affinity calculator (HTBAC). The architecture of HTBAC and its uses is described here, and published in more detail in the proceedings of the 2018 eScience IEEE conference.

HTBAC is a software system for running ensemble-based free energy protocols adaptively and at scale on HPC resources. Currently, HTBAC supports protocols composed of an arbitrary number of analysis and simulation steps, and relies on the ensemble management system and runtime system provided by the RADICAL-Cybertools (RCT). HTBAC is designed to be extended to support more types of protocols and alternative runtime middleware.

Design and implementation

HTBAC exposes four constructs to specify free energy protocols: Protocol, Simulation, Analysis, and Resource. Protocol enables multiple descriptions of protocol types, while Simulation and Analysis specify simulation and analysis parameters for each protocol. Resource allows to specify the amount of resources needed to execute the given protocols. Together, protocol instances, simulation and analysis parameters, and resource requirements constitute an HTBAC ap-

plication. In HTBAC, each protocol models a unique protein ligand physical system. Each protocol follows a sequence of simulation and analysis steps, assigning ensemble members to execute independent simulations or analysis. An ensemble member that executes an independent simulation within a simulation step is referred to as a replica. Each simulation is assigned a different initial velocity, which enables simulations to begin in different parts of the ligand's phase space. Individual simulations or analyses with input, output, termination criteria and dedicated resources are designed as a computational task. Aggregates of tasks with dependencies that determine the order of their execution constitute a workflow. In this way, HTBAC encodes NP instances of the Pth protocol as a workflow of computational tasks. Fig. **??** shows the components and subcomponents of HTBAC. HTBAC API enables users to codify protocol descriptions in terms of protocol type, simulation and analysis steps, and computer infrastructure requirements. Descriptor uses two subcomponents to aggregate protocol descriptions into a single application and resource description. Note that Descriptor can aggregate different types of protocols, with different computing and resource requirements. Runner has three subcomponents: Execution Manager, Middleware Connector and Runtime Adaptive Evaluator. Execution Manager communicates with the execution layer via a connector to coordinate the execution of the application. In principle, HTBAC can use multiple connectors for diverse middleware to access different computing infrastructures. Middleware Connector converts the application description of HTBAC into a middleware-specific format. Execution Manager can pass the given application to the connector in full or only in parts. This enables to start the execution of an application before its full description is available or to change those parts of the application that still have to be executed. This will enable future capabilities like, for example, to concurrently execute the application on diverse middleware. Runtime Adaptive Evaluator enables the execution of adaptive applications. This subcomponent can evaluate partial results of an application execution via tailored algorithms. On the base of this evaluation, Runtime Adaptive Evaluator can decide to return the control to Execution Manager or to modify the description of the application that is being executed. In this way, HTBAC implements adaptivity for diverse protocols, allowing users to define arbitrary conditions and algorithms. HTBAC is implemented in Python as a domain-specific library. All components of HTBAC are implemented as objects that communicate via method calls. HTBAC uses two RCT as building blocks: Ensemble Toolkit (EnTK) and RADICAL-Pilot (RP). EnTK provides HTBAC capabilities to execute ensemble-based applications. EnTK exposes three constructs: Task, Stage and Pipeline. Tasks contain information regarding an executable, its software environment and its data dependencies. Stages are sets of tasks without mutual dependencies that can execute concurrently. Pipeline are lists of stages, where stages can execute only sequentially and each pipeline can execute independently. HTBAC uses a Middleware Connector for EnTK to encode a protocol instance as a single pipeline that contains stages of individual simulations and analyses tasks. EnTK uses RP to execute tasks via pilots. RP supports task level parallelism and high-throughput by acquir-

ing resources from a computing infrastructure and scheduling tasks on those resources for execution. Pilot systems execute tasks directly on the resources, without queuing them on the infrastructure's scheduler.

Adaptivity

The design of HTBAC permits enhancing protocols while continuing to use "static" simulation engines. To this end, we implemented two adaptive methods using HTBAC: adaptive quadrature and adaptive termination. Both of these methods use the features of adaptivity offered in HTBAC to scale to large number of concurrent simulations and to increase convergence rate and obtain more accurate scientific results. The aim of introducing adaptive quadrature for alchemical free energy calculation protocols (e.g., TIES) is to reduce time to completion while maintaining (or increasing) the accuracy of the results. Time to completion is measured by the number of core hours consumed by the simulations. Accuracy is defined as the error with respect to a reference value, calculated via a dense λ window spacing (65 windows). This reference value is used to establish the accuracy of the non-adaptive protocol (which has 13 λ windows) and the adaptive protocol (which has a variable number of λ windows, determined at run time). One of the input parameters of the adaptive quadrature algorithm is the desired acceptable error threshold of the estimated integral. We set this threshold to the error of the non-adaptive algorithm calculated via the reference value. The algorithm then tries to minimise the number of λ windows constrained by the accuracy requirement.

Chapter 3

Results and discussion

This chapter describes and discusses the results. The results for each of the computational approaches to estimate either the binding free energy directly or range the intercalators is presented first, followed by a discussion of their relative merits.

3.1 Docking and scoring

The rigid planar scaffold of the intercalators that bind between DNA basepairs reduces the possible number of conformations that the docked structure can have. We have shown that using an extensive search method for docking, intercalators can be docked with RMSD less than 1 angstrom for a wide range of intercalators compared to crystal structures.

We collected nine crystal structures of DNA-intercalator complex where the intercalator is between two consecutive basepairs. Table 3.1 shows the RMSD of every intercalator docked compared to the crystal structure. For seven of the cases docking is very close to the crystal structure, with average RMSD of 0.76 angstrom. There are two cases where the docked structure is far from the reference.

After further investigating the outlier, the docked intercalators are in the correct place between the two basepairs, but their orientation is different from the one in the reference crystal structure. If there is no symmetry present in the intercalator molecule, then there are four possible orientations once docked. The free energy difference between these orientations is probably low, and the experimentally observed, dominant orientation is dependent of the kinetics of the intercalation too.

The quinoxaline scaffold based intercalators were also docked and scored into the basepair opening of the 1Z3F crystal structure removing the original intercalated molecules.

Table 3.1: Results of docking nine intercalators compared to crystal structures.

PDB ID	RMSD [angstrom]
1Z3F	5.5
1Z3F	0.83
1G3X	0.59
1HX4	1.2
2ROU	3.9
1D11	0.59
1D11	0.87
1D12	0.40
1D12	0.86

3.2 ESMACS

In the past, there has been a number of MMPBSA type calculations for DNA intercalator systems. A large majority of these studies only takes into account just a few (one to three) intercalators binding to the DNA, and does not compare to experimental studies. We obtained a correlation Pearson coefficient of $r_p = 0.70$ across the 10 planar molecules for the 3 trajectory calculation.

Convergence

It is important to assess the sampling efficiency and converge of the predicted free energy values in order for the results to be reliable and reproducible. First, the comparison in convergence between a single trajectory approach and an ensemble of trajectories is compared for the MMPBSA and NMODE analysis methods in the context of free energy calculations. As the improvement in ensemble methods was evident, we investigated the length of each replica, extending the simulations from the original 5 ns simulation to 20 ns for each of the 25 replicas. It was important to sample the system long enough to cover a large part of the phase space, but not to oversample in places where the methods is no longer valid, for example during a de-intercalation pathway.

Replica size and bootstrapped error

Predicting the ranking

The experimental free energy values are between $-3.69\,\mathrm{kcal\,mol^{-1}}$ and $-3.50\,\mathrm{kcal\,mol^{-1}}$, making them hard to differentiate computationally. Approximate methods, like ESMACS, that are based on MMPBSA and NMODE for the free energy and entropy contribution, respectively do not have the resolution to differentiate experimental values at this scale. Nonetheless, correct error control and ensemble based calculations can give meaningful results relative to each other, i.e. ranking the intercalators based on their free energies will be comparable to experimental ranking.

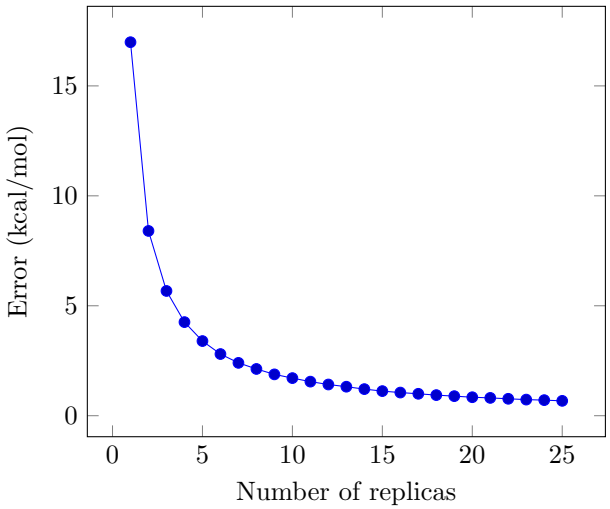

Figure 3.1: Bootstrapped error as a function of replica size.

We have run 1, 2 and 3 trajectory calculations and as seen in Figure 3.2 the correlation improves, with the 3 trajectory being the most accurate with a correlation coefficient of $r_p = 0.70$. The small interval of the experimental values suggests that the intercalators are very closely binding to the DNA.

The correlation achieved with ESMACS in this study suggests that one of the underlying forces (electrostatic, vdW, or internal energy) dominates the interaction gradient between the intercalators, and that signal is picked up by the MMPBSA calculation. Figure **??** shows the spread of the free energy values of the different components. Previous studies [] have shown that vdW interactions contribute most to the binding energy of the intercalators in absolute terms. However, the correlation of the different components to the total binding energy shows that the internal energy (the sum of bond, angle, and torsion terms) correlate surprisingly with the total. This further supports our claim that the intercalation is an induced fit inside the two base-pairs.

Each ESMACS calculation was run 25 times, the only difference being the random seed at the start of the simulation.

Variations in the results from the different replica was used to evaluate the error on the predicted free energy. Figure 3.1 shows the bootstrapped error as a function of the number of replicas. There is a leveling off at around 20 replicas, with 25 replicas showing consistently low error bars.

NMODE analysis was performed to include an approximation to the entropy contribution. Results show that including NMODE does not improve the ranking.

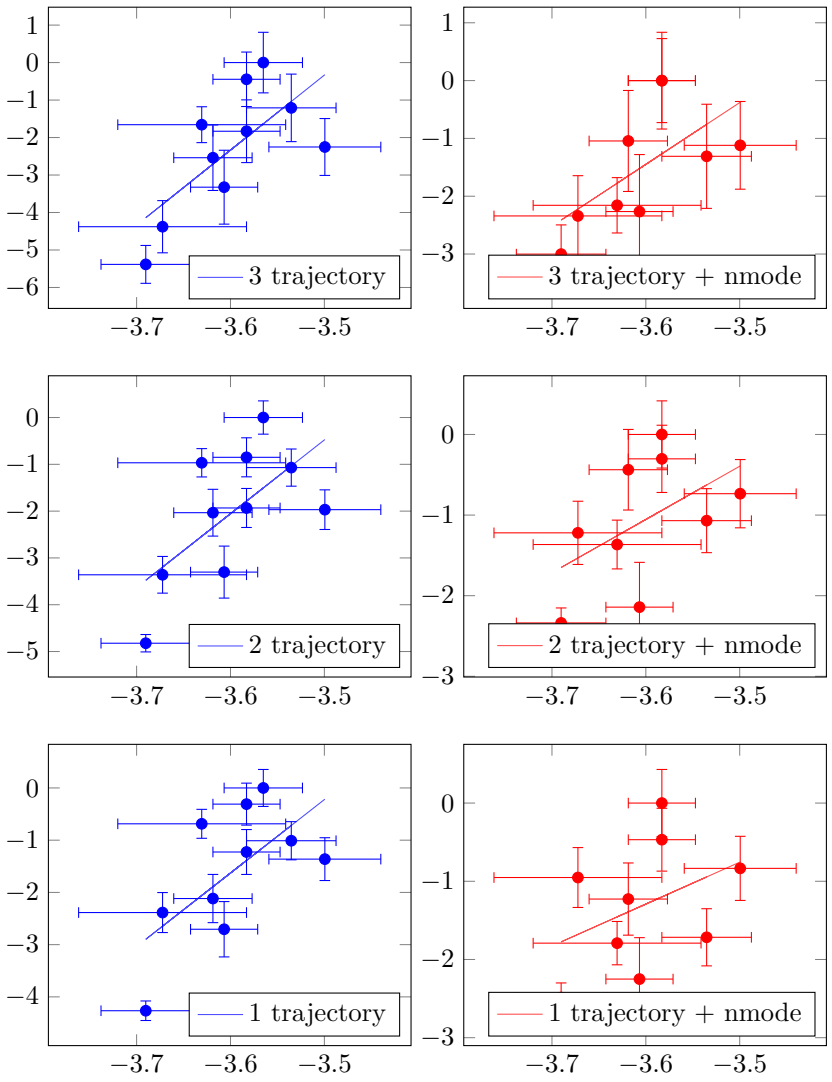

Figure 3.2: Correlation between one (bottom), two (middle) and three (top) trajectory calculations and experiment. Values are presented for NMODE (red) analysis and without NMODE (blue). The best Pearson correlation is for the NMODE-less values starting from 0.64 for the 1, and 2 trajectory and 0.70 for the 3 trajectory. The NMODE calculations have a Pearson correlation of 0.39, 0.5 and 0.62 for the 1, 2, and 3 trajectory calculations respectively.

3.3 TIES

Relative alchemical free energy calculations were done between five pairs of the intercalator from the original experimental study. As discussed above, there

are two groups of intercalator there: the four and five membered ring scaffolds. Pairs were selected based on two criteria: first to maximise the experimental free energy difference and second to sample all three possible combinations of transformation (between the 4 ring systems, 5 ring systems, and a transformation from 4 ring to 5 ring).

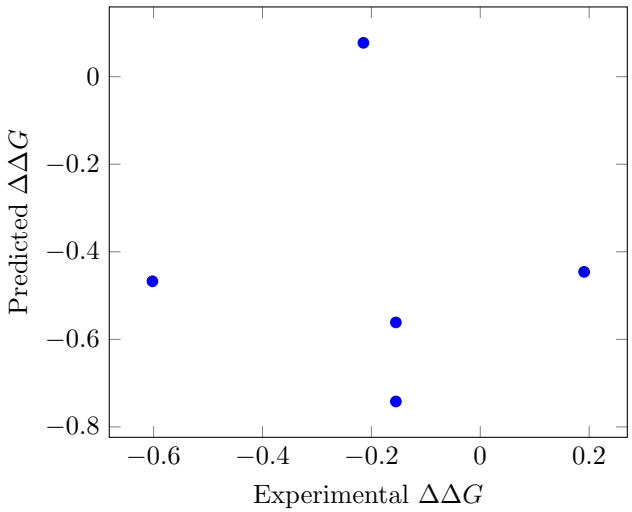

Figure 3.3: Correlation plot comparing the experiment free energy difference with the calculated values via TIES

3.4 Discussion

The three methods we investigated correspond to three theoretically more accurate ways of quantifying the relative or absolute binding affinity of these intercalators.

Chapter 4

Conclusion