Advancing the Use of Simulation Methods in Structure-Based Drug Discovery

PhD Thesis

Lucia Fusani

April 2020

GlaxoSmithKline and University of Strathclyde

collaborative PhD programme

This thesis is the result of the author’s original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

The copyright of this thesis belongs to GSK in accordance with the author's contract of engagement with GSK under the terms of the United Kingdom Copyright Acts. Due acknowledgement must always be made of the use of any material contained in, or derived from, this thesis.

Signed: ………………………… **Date**: …………………………

Abstract

The role of simulation methods in industrial Structure-Based Drug Design (SBDD) is here investigated with focus on the bromodomain-containing protein 4 (BRD4). New tools were developed and applied, in conjunction with existing techniques, to study four problems in SBDD. Firstly, the ability to predict the water network at the bottom of the BRD4 protein was studied with 3D-RISM. The difficulties identified with the available tools led to the development of a new algorithm, GAsol. The results show that GAsol can accurately predict the most probable locations of the individual water molecules in the cavity of BRD4 protein. Secondly, since an incorrect interpretation of the ligand-binding mode in a crystal structure could compromise the overall SBDD process, Binding Pose MetaDynamics (BPMD) was investigated as a quantitative tool to assess the stability of protein-ligand complexes in solution. BPMD clearly separates ligands whose binding pose is supported (stable) and not (unstable) by the electron density. Thirdly, simulation methods were used to guide the design of a probe for the BRD4 protein with selectivity for the BD1 domain over BD2. The generated hypotheses were explored by means of Molecular Dynamics (MD) and MetaDynamics (MetaD) simulations. Insights on the role played by specific water molecules and residues enabled the design of several compounds which culminated with the identification of the desired probe. Finally, the interconversion between two binding modes seen in the X-ray crystal structures of a GSK-ligand in complex with the BRD4 protein is studied with MD, MetaD and a strategy of docking and BPMD. While MD was partially successful in reproducing the multiple ligand binding modes and MetaD with the funnel restraints was not, the docking and BPMD protocol showed promising results.

Acknowledgments

I wish to thank all the people whose assistance was a milestone in the completion of this project. First of all both my supervisors Dr. Ian Wall and Dr. David Palmer. They guided and encouraged me during all my journey. Dr. Alvaro Cortes who helped me in shaping my interests and made me even more passionate about the field of drug discovery & design. Dr. Chun-wa Chung for providing the crystallographic data in most of my projects and Dr. Don Somers for his help in analysing X-ray crystal structures, for his patience and support. Dr. Richard Lonsdale, Dr. Paul Bamborough and James Woolven for their precious advice in drug design. Dr. Peter Pogany for his invaluable help in programming. Dr. Rosa Cookson and Erin Bradley for synthesising compounds for the BET-BD1 project. Prof. Harry Kelly and Prof. Billy Kerr that believed in me and took me on this amazing PhD Programme. Prof. Francesco Gervasio for allowing to visit his group at UCL and Giulio Mattedi for his help with the funnel. Dr. Marie Dreanic, Dr. Sandeep Pal, Dr. Baptiste Canault, Dr. Damien Coupry for their support and interest in my research and for making the lunch break more interesting. The whole Computational Chemistry Group that welcomed me and taught me a lot. All the Palmer Lab members in particular Benjamin Smith and Samiul Ansari and all the PhD colleagues in the PhD Programme. A special thanks to Maxime Rouah and Sam Holman who listened me and made my coffee breaks more enjoyable.

I would like to thank my parents, Maurizio and Nadia, who have always been there for me no matter what. And lastly, I want to recognize the invaluable support, assistance and love that my caring husband, Marco, provided me during my studies.

Thank you!

Table of Contents

Abbreviations

Å	Angstrom
Asn	Asparagine
Asp	Aspartic Acid
BPMD	Binding Pose Metadynamics
BET	Bromodomain and Extra Terminal domain
BM1	Ligand A in Mode 1
BM2	Ligand A in Mode 2
BPTI	Bovine Pancreatic Trypsin Inhibitor
BRD	Bromodomain
BRD2	Bromodomain containing protein 2
BRD2-BD2	Bromodomain 2 of bromodomain containing protein 2
BRD4	Bromodomain containing protein 4
BRD4-BD1	Bromodomain 1 of bromodomain containing protein 4
BRD4-BD2	Bromodomain 2 of bromodomain containing protein 4
COM	Center of Mass
CompScore	Composite Score
Cmpd	Compound
CV	Collective Variable
ED	Electron Density
EpoA	Epothilone A
FES	Free Energy Surface
FF	Force Field
FM	Funnel Metadynamics

GBSA	Generalized-Born
GCMC	Grand Canonical Monte Carlo
Gln	Glutamine
GPU	Graphics Processing Unit
GSK	GlaxoSmithKline
FM	Funnel MetaDynamics
HAT	Histone Acetyltransferase
HDAC	Histone Deacetylase
His	Histidine
HNC	Hyper Netted Chain
IC_{50}	Concentration that reduces the effect by 50%
IET	Integral Equation Theories
IFD	Induced-fit Docking
Ile	Isoleucine
IST	Inhomogeneous Solvation Theory
KAc	Acetyl-lysine
KH	Kovalenko Hirata
Lys	Lysine
µs	Microsecond
MD	Molecular Dynamics
MetaD	MetaDynamics
MM	Molecular Mechanics
nm	Nanometres
ns	Nanosecond

NVT	Canonical ensemble with constant number of particles, volume and temperature
NPT	Isobaric-isothermal ensemble with constant number of particles, pressure and temperature
PBC	Periodic Boundary Conditions
PBSA	Poisson-Boltzmann
PDB	Protein Data Bank
PersScore	Persistence Score
PES	Potential Energy Surface
Phe	Phenylalanine
pIC_{50}	The negative logarithm to base 10 of the IC_{50}
Pro	Proline
PTM	Post-Translational Modifications
ps	Picosecond
QM	Quantum Mechanics
QT	Quality Threshold
RISM	Reference Interaction Site Model
RMSD	Root Mean Square Deviation
RMSF	Root Mean Square Fluctuation
RSCC	Real Space Correlation Coefficient
SAR	Structure Activity Relationship
SBDD	Structure-based Drug Design
SIFt	Structural Interaction Fingerprints
Trp	Tryptophan

Tyr	Tyrosine
Val	Valine
W	Water Molecule in the ligand-binding pocket of BRD protein
WPF	Tryptophan-Proline-Phenylalanine
χ (chi)	Side chain torsion angles of amino acids

List of Figures

List of Tables

GSK CONFIDENTIAL INFORMATION – DO NOT COPY

Chapter 1 Introduction

1.1 Structure-Based Drug Design

Drug discovery is a multi-disciplinary process that aims to identify new chemical entities with the potential to stop or reverse the effect of the disease under investigation. In general, the process of drug discovery is composed of multiple steps (Figure 1-1) starting from the identification and validation of the target that is relevant for the disease to be treated. Then, compound libraries can be used to identify hits that will be optimized to generate a lead compound with adequate potency and selectivity towards the biological target *in vitro* and also efficacy in animal models. The lead compound can then progress towards the drug development phase (pre-clinical and clinical stages). Together, the discovery and development of new innovative drugs is time and cost intensive and it has been estimated to take about 10-15 years and cost \$1.8 billion[1, 2]. Nevertheless, despite all the efforts, only a limited number of drug discovery projects will succeed in identifying a new drug.

GSK CONFIDENTIAL INFORMATION – DO NOT COPY

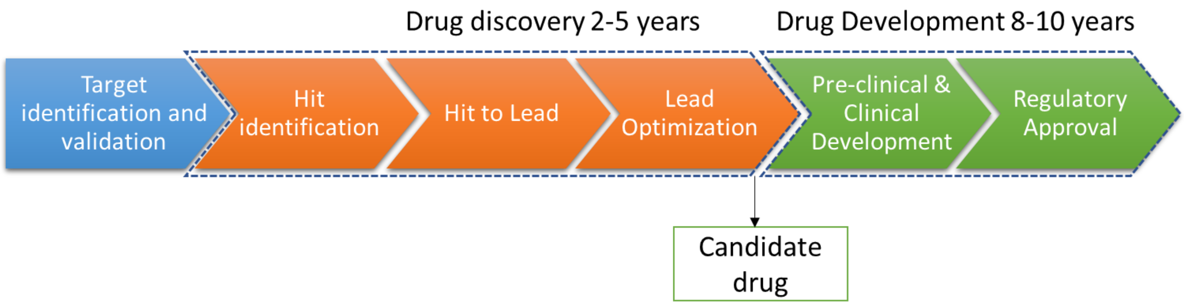


Figure 1-1 Process of drug discovery and development from target identification and validation through to the regulatory approval and the approximate timescale for those processed. The drug discovery process, in orange, can greatly benefit from Structure-based Drug Design techniques.

Structure-based drug design (SBDD) or rational drug design refers to a specific approach of drug design and represents a significant component of most industrial drug discovery projects as well as being the topic of research for many academic laboratories[3]. SBDD is an iterative process that starts from the known structure of a target molecule, typically a protein, followed by *in silico* studies to identify potential binders (ligands). It gives insight in the interaction of a specific protein-ligand pairs, allowing medicinal chemists to devise chemical modifications around the ligand scaffold[4]. The most used SBDD *in silico* strategies are molecular docking, structure-based virtual screening and Molecular Dynamics[5] (MD). As a result of the computational studies, the most promising compounds are synthesised, then biological activity data such as potency, affinity and efficacy are measured using experimental platforms and correlated to the structural information. In this way, the SBDD cycle starts again with the aim to incorporate molecular modifications designed to increase the affinity of new ligands for the binding site[6] (Figure 1-2).

GSK CONFIDENTIAL INFORMATION – DO NOT COPY

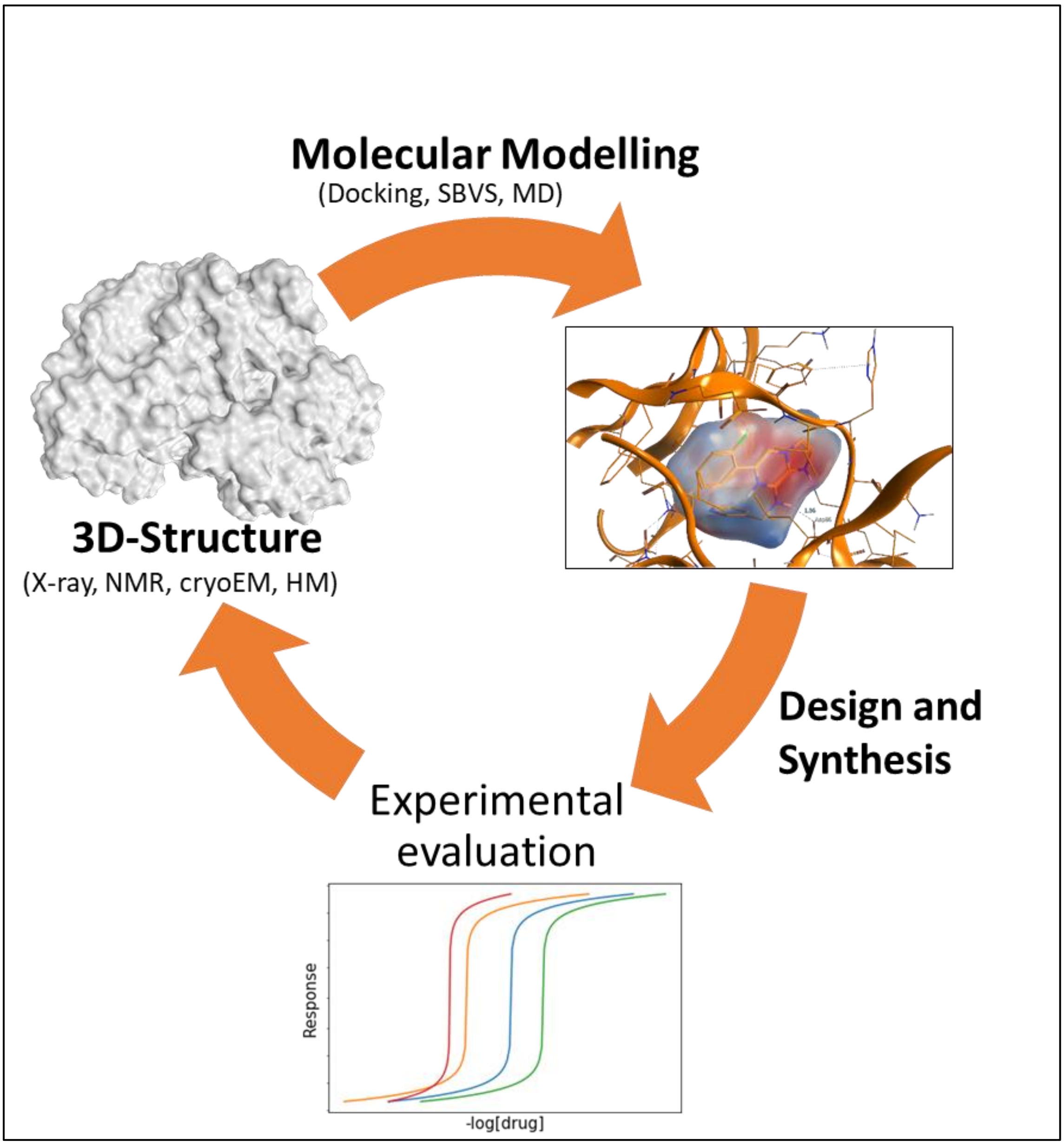


Figure 1-2 The SBDD cycle. The cycle can start either with the chemical lead, which is then crystallized with the target protein or with the structure of the target protein alone for *de novo* design. A number of iterations is usually necessary before arriving to a suitable candidate.

Among the most important successes in SBDD, there is the HIV-1 (Human Immunodeficiency Virus 1) inhibitor Saquinavir[7], which made it possible for many

GSK CONFIDENTIAL INFORMATION – DO NOT COPY

HIV infected individuals to live longer than they could without treatment, and also the FDA approved Dorzolamide which is currently used for the treatment of glaucoma[8].

1.2 Structures in Structure-Based Drug Design

Structural information about the target is essential for SBDD and is obtained experimentally with the use of X-ray crystallography, NMR, cryo-electron microscopy or in extreme cases through computational homology modelling[3]. The number of three-dimensional macromolecular structures has grown exponentially thanks to several factors such as better molecular biology techniques, improvement of protein purification methods, the arrival of powerful synchrotrons, better X-ray sources and the development of cryo-cooling techniques[9]. Nowadays, the Protein Data Bank (PDB) contains more than 150,000 macromolecular structures and this has significantly contributed to the progression of SBDD projects.

X-ray crystallography is well-suited for drug discovery because it allows the acquisition of high quality structures and it provides direct evidence of the binding mode of a small molecule to the target. Crystals can be obtained either with co-crystallization where the ligand is added to the protein in solution and then crystallised or by soaking the ligand into the existing protein crystal[10]. The crystal is then placed in an intense beam of X-rays and the diffracted rays are collected. The diffraction pattern, a collection of intensities and positions of the reflections, allows the calculation of the electron density in the crystal. The resulting 3D electron density map is then used to model the atomic coordinates of the protein and the ligand. The X-ray

GSK CONFIDENTIAL INFORMATION – DO NOT COPY

crystal structure will constitute the model containing the protein-ligand information necessary for subsequent SBDD approaches. It is important to remember that after the initial crystal structure is obtained, multiple refinement steps are done to generate the best possible fit between the observed and calculated electron density. Therefore, X-ray crystal structures are one of the crystallographer's subjective interpretations of the electron-density. In fact, the interpretation of the three-dimensional macromolecular structures from X-ray experimental data remains an undisputed challenge and suffers from limitations and uncertainties[11, 12]. Three of the major concerns in the structure determination are correlated to: 1) uncertainty in the identity/location of protein or ligand atoms; 2) the difficulties in the identification and location of water molecules; 3) the assumption that a single crystal structure could represent the full and true picture of a flexible three-dimensional macromolecule. In fact, crystal structures cannot take into account all of the structural fluctuations that a macromolecule adopts under physiological conditions and this could limit research into new molecules. Moreover, among the ambiguities to solve a three dimensional structure it is important to mention the effect of crystallization conditions on the observed ligand and protein conformation and the detection of hydrogen atoms[9, 11, 12].

1.3 Molecular Dynamics in Structure-Based Drug Design

A static set of atomic coordinates is the closest approximation of macromolecular structures that crystallography can determine and represent. A crystal structure cannot accurately capture all of the structural fluctuations that a macromolecule adopts under

GSK CONFIDENTIAL INFORMATION – DO NOT COPY

physiological conditions. By contrast, Molecular Dynamics simulations can provide atomistic details and insights into relevant macromolecular conformations.

The analogy used to describe the interaction between an enzyme (receptor) and its substrate (ligand) by Emil Fisher[13] in 1894 as a rigid lock-and-key is nowadays superseded as receptor and ligand flexibility are crucial for drug binding. Several alternative theories were proposed. Koshland[14] proposed the “induced fit” model in which the initial interaction between the ligand and receptor induces a conformational change in each of them. Later on, the conformational selection model was proposed[15] in which the receptor is assumed to exist as an ensemble of conformations in equilibrium and the binding partner (ligand) will preferentially bind to one of those causing a shift in the equilibrium toward the bound state. Later, it was suggested that to better describe the complex mechanism of protein ligand binding a coexistence of all the models must be accepted[16].

Computational techniques like flexible docking and in particular MD are used to address the flexibility issue and overcome the limitations that the use of a static structure for SBDD can have. MD simulations are able to describe the dynamic aspects of protein structures by tracking the time-dependent positions of all atoms in the systems. From the analysis of MD simulations, a variety of properties can be calculated including free energy and other macroscopic quantities, which can be compared with experimental observables.

The MD method was first introduced by Alder and Wainwright in the late 1950's[17] and the first MD simulation of a biological macromolecule was run in 1977 by McCammon, Gelin and Karplus who obtained 10 ps for the protein BPTI (Bovine Pancreatic Trypsin Inhibitor) in vacuum[18]. Since then, the field of biomolecular simulation has progressed enormously thanks to improvements in algorithms, software and computer hardware and thanks to the advances in force field development[19-21] and parallelization schemes. In 1998, the first microsecond simulation of a biomacromolecule in aqueous solution was performed by Duan and Kollman[22] from which it was observed the folding of the 36-residue villin headpiece subdomain from a fully unfolded state. The simulation required 2 months of CPU time with the usage of parallel supercomputers and parallelized codes. A machine designed particularly for MD simulations called ANTON was developed in 2008 by the DE Shaw group[23]. The use of the ANTON supercomputer allowed the ms timescale to be reached for reproducing the folding of bovine pancreatic trypsin inhibitor (BPTI)[24]. Most importantly, what has dramatically impacted the access to higher timescale for academic laboratories and pharmaceutical companies is the use of graphics processing units (GPUs)[25] and of codes able to exploit the speed of the GPUs. In pioneering work Buch et al.[26] investigated binding of benzamidine to trypsin using multiple replica of μs-long MD simulations; another similar work was proposed by Shan et al.[27] where they showed how dasatinib and the kinase inhibitor PPI find the binding site on the Src kinase. Although we have access to simulations with a time scale of the microsecond, computational chemists still express their concerns relative to the reproducibility and statistical significance of MD simulations[28]. In fact, several different trajectories are

GSK CONFIDENTIAL INFORMATION – DO NOT COPY

recommended to obtain adequate statistics and exhaustive sampling of the conformational space. Thus, even when working with only one compound, the process is quite demanding. On this subject, in a recent work by Knapp et al.[29] it was reported that more reliable and reproducible results can be drawn from multiple shorter replicas as opposed to single longer simulations.

Biomolecular simulations are increasingly used in SBDD as they provide significant contributions to better understand molecular recognition principles which are difficult to observe with experimental techniques, such as the dynamics of protein-ligand interactions, the role of solvent molecules as well as conformational rearrangements and flexibility. It is clear that MD can assist and inform rational drug design but robust, fully automated and validated protocols are not yet available so efforts should be made in this direction to allow the complete integration of MD in drug design workflows.

1.4 Water molecules in Structure-Based Drug Design

The role of the water molecules in the active site of the proteins has become of considerable interest over the years. Water molecules contribute to the stabilisation, dynamics and functions of the biomolecules. Furthermore, water mediates protein-ligand interactions and has a key role in the ligand binding process[30, 31]. A study on 392 protein-ligand complexes revealed that in 85% of the cases at least one water molecule is found to bridge interactions between the ligand and the protein[32]. For this reason, when evaluating the binding affinity of a protein-ligand complex, water molecules need to be taken into proper consideration[33, 34].

GSK CONFIDENTIAL INFORMATION – DO NOT COPY

Water molecules have a significant role in the energetic effect during the ligand-protein binding *i.e.* through their contribution to the hydration/dehydration and hydrophobic effect that is the process of self-association of water molecules[35]. The former relates to the process of displacing all the water molecules at the interface between ligand and protein followed by protein-ligand reorganization. However, during the protein-ligand binding event, not all the water molecules are going to be displaced; the strongly bound ones might be retained in the binding site and help in the ligand-protein complex formation through additional bridging interactions[30, 36, 37]. Therefore, the presence of water molecules in the binding site of proteins has different effects on the energy, enthalpy and entropy of the system which could in turn favour or disfavour the overall binding process[35]. For example, there could be an energy cost if a water molecule is trapped in a hydrophobic cavity without any partner molecule. Conversely, energy could be gained when water molecules are released into the bulk solvent or engaged in forming additional hydrogen bonds. Understanding the thermodynamic properties of a water and more precisely being able to predict which water molecule could be favourably displaced could impact SBDD. However, it is not always guaranteed that the replacement/displacement of a water molecule will decrease/increase the binding energy. In fact, while a water molecule is displaced by a portion of the ligand, other water molecules might as well be influenced by this modification in a non-additive manner such that the overall energetic changes might be misleading. Therefore, trying to rationalize the multifactorial behaviour of water molecules is challenging.

GSK CONFIDENTIAL INFORMATION – DO NOT COPY

In SBDD and drug-design in general, developing a ligand with high binding affinity for its target is one of the main goals. Hence, considering the effect of bound water molecules *e.g.* their location and their associated binding energy is extremely important.

To include water molecules in drug design models, the position of the potential water sites needs to be known. The presence or absence of water molecules cannot always be determined with certainty by solving X-ray crystal structures because of their inherent mobility but also because the resolution might be too low or their presence could be strongly influenced by the experimental conditions[38]. In addition, adding an excessive number of water molecules to a model can artificially increase some quality related parameters employed in structural refinement[11] and for this reason, it could become a subjective matter whether a feature in the electron density should be ignored as noise or modelled as a water molecule.

In order to complement experimental techniques and model solvation effects of individual water molecules, a spectrum of various computational approaches have been developed to identify the likely position of water molecules in binding sites and to evaluate their energetic stability[39]. The most popular computational approaches to locate waters molecules are listed below and grouped on the basis of the solvent models (implicit or explicit) employed.

GSK CONFIDENTIAL INFORMATION – DO NOT COPY

WaterMap[40, 41], GIST[42], Grand-Canonical Monte Carlo (GCMC)[43], Just Add Water molecules (JAWS)[30] and WATsite[44] all use explicit water models to describe the solvent around the solute. WaterMap, GIST and WATsite are MD-based tools whereas GCMC and JAWS are Monte Carlo (MC) methods. SZMAP is defined as a hybrid implicit/explicit probe approach and lastly the Three-Dimensional Interaction Site Model[45] (3D-RISM) is an implicit continuum model that includes a description of local variations in solvent density around a solute (see details in Section 3.2).

WaterMap is the most common algorithm for water placement that belongs to Schrödinger's suite of programs. Its first usage goes back to 2007 where the method was first validated and used retrospectively to explain the affinity of known ligands[41]. Efforts have been made and are still in progress to make WaterMap a routine tool to use in ongoing medicinal chemistry programs. In some cases, WaterMap has been shown to be helpful for the progression of projects, as for example for the identification of spirocyclic sulphonamides for the inhibition of β-secretase 1[46], development of inhibitors of platelet-derived growth factor receptor β[47], understanding the PI3K-Beta/Delta selectivity and SAR of a series of pyrimidone-indoline amide inhibitors[48], and understanding the selectivity and affinity of indirubin analogues towards the kinase DYRK (Dual Specific Tyrosine Phosphorilation-regulated Kinase)[49]. Nevertheless, there are some limitations to the method. The protocol suggests the use of multiple starting structures, as well as apo structures, therefore, it could become computational demanding and results might arrive with delay in respect to the project timeline. Moreover, it should be noted that in WaterMap the system is simulated for

GSK CONFIDENTIAL INFORMATION – DO NOT COPY

short time and with restraints on the receptor, therefore, if the binding site is subjected to high flexibility this method might not be appropriate. Lastly, if the binding site is already occupied by a ligand or is occluded, WaterMap finds difficulties in reproducing the full picture of solvation due to poor sampling[39].

To circumvent the sampling problem Monte Carlo methods have been used such as GCMC and JAWS, where water molecules can be created, destroyed and moved according to the chemical potential of the simulation allowing a complete sampling of the binding site and also of solvent-inaccessible pockets.

GCMC and WaterMap are able to predict the positions of water molecules within a network and while in WaterMap the binding free energy is calculated for one molecule at a time[39], in GCMC it is calculated for the entire water network. In GCMC it is important that the number of water molecules has equilibrated during a simulation otherwise problems can arise for water molecules that for example are dependent on the movement of protein side chains. One of the main limitations of GCMC that probably has limited its applications, was defining the correct chemical potential that yielded equilibrium between the bulk water and the system of interest but recently this has been successfully implemented and there is hope to expand the scope of the method[43].

Water prediction and analysis is a “hot topic” in medicinal chemistry as there is more awareness of the importance that water molecules have in the active site of target proteins. Use of water modelling tools is not yet completely incorporated within drug

GSK CONFIDENTIAL INFORMATION – DO NOT COPY

design projects but many successes on retrospective project have been obtained through the use of such techniques to explain SAR profiles and binding kinetics. It is true that the interpretation of the results from water modelling methods (i.e. WaterMap, GCMC, SZMAP, and 3D-RISM) can be challenging but at the same time, when possible, those methods can provide additional information that conventional approaches might miss. Therefore, with the increasing use of such methodologies as well as the number of publications, water modelling tools are becoming a non-negligible part of modern drug discovery.

1.5 An introduction to Bromodomains

DNA, the biological macromolecule that encodes all the information in the human genome, is packed into chromatin by wrapping around histone proteins to form a nucleosome (histone octamer). Nucleosomes are further compacted in chromatin, whose level of compactness depends on the post-translational modification (PTM) present on the histones, especially on their terminal residue[50] (Figure 1-3). The combination of modifications on histones establish a code (“the histone code”) that relates to the transcriptional properties of the genes nearby.

GSK CONFIDENTIAL INFORMATION – DO NOT COPY

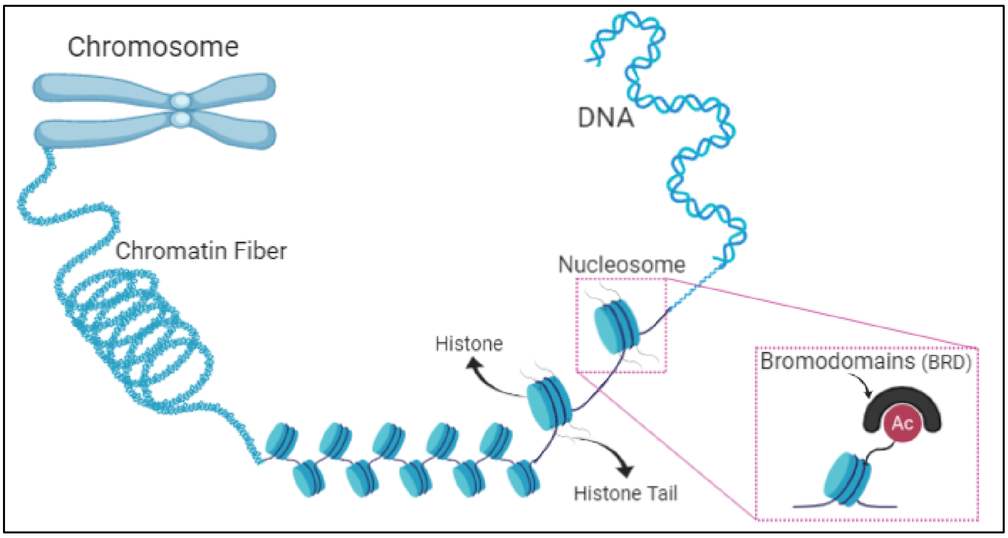


Figure 1-3 DNA associate with histone proteins to form nucleosomes. Chromatin state is influenced by PTM of histone tails such as acetylation. The protein family that recognize and read acetylated groups are the Bromodomains, shown in the inset, YEATS and DPF, not shown.

A frequent epigenetic PTM occurring on histone proteins and implicated in the regulation of the chromatin structure and transcription is the lysine acetylation. The neutralisation of the charge of the lysine with the addition of an acetyl moiety has a profound effect on protein conformation and protein-protein interactions. Usually, in the histone such modifications promote an open architecture of chromatin with resulting transcriptional activation[51] but they have also been linked to compaction of chromatin, DNA repair, protein stability and regulation of protein-protein interactions. In many diseases, aberrant lysine acetylation leads to alterations in gene expression, causing for example the activation of pro-survival and proliferation-promoting pathways and inactivation of tumour suppressor functions.

Cellular acetylation levels are controlled by histone acetyltransferases (HAT) and histone deacetylases (HDAC) that respectively write and erase acetylation marks. The

GSK CONFIDENTIAL INFORMATION – DO NOT COPY

complex pattern of acetylation marks is interpreted by readers which include the double plant homeodomain finger (DPF), YEATS[52] and the bromodomain (BRD) family of proteins (Figure 1-4).

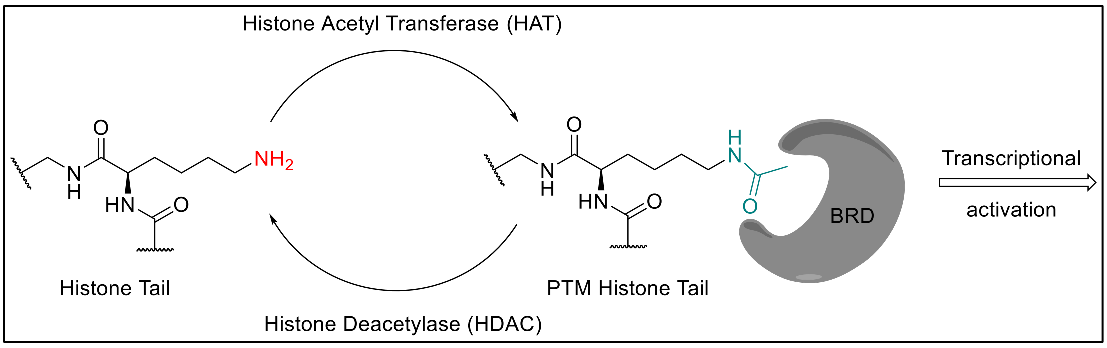


Figure 1-4 Lysine Acetylation (KAc). The lysine ε-nitrogen, red, is acetylated by HAT to give the acetylated lysine, in green. The KAc is read by the bromodomain (BRD) family of protein which leads to transcriptional activation. Lysine residues are deacetylated by HDACs.

BRDs have been named after the *Drosophila* gene *brahma*, for which the core bromodomain sequence motif was first identified[53]. BRDs are evolutionary conserved protein-interaction modules of about 110 amino acids that recognise ε-N-lysine acetylation motifs and are crucial for the regulation of transcription[54]. In an analysis of the human genome, 61 bromodomain found within 46 proteins were categorised in 8 families (Figure 1-5) based on sequence and structural similarities[55, 56].

GSK CONFIDENTIAL INFORMATION – DO NOT COPY

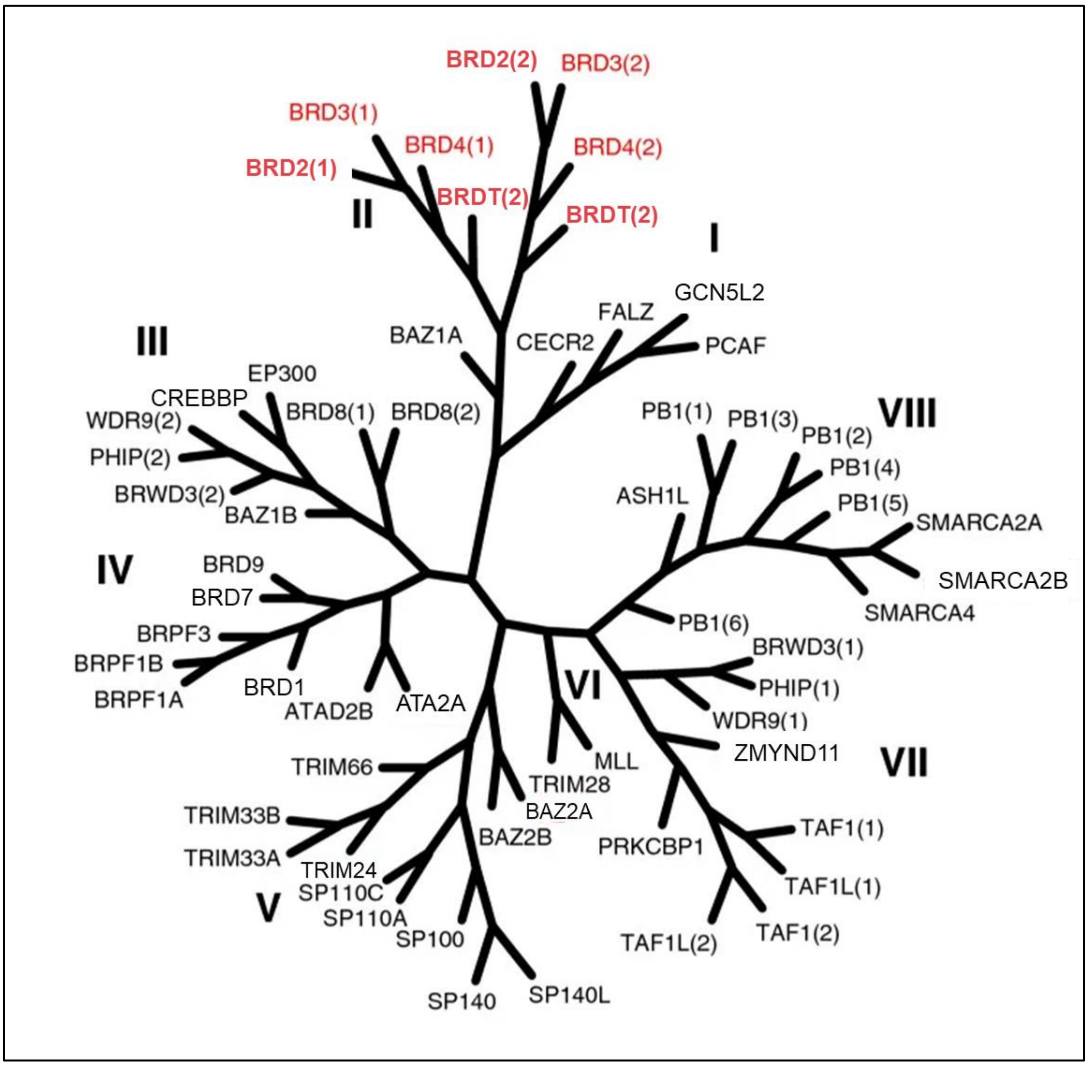


Figure 1-5 Phylogenetic tree of the human BRD family adapted from Filippakopoulos *et al*[56]. The different families are enumerated with Roman numbers (I-VIII). Proteins of the BET family (family II) are highlighted in red.

The most studied family of BRDs is the Bromodomains and Extra Terminal domain (BET) consisting of two tandem bromodomains (BD1 and BD2) and the C terminal extra-terminal (ET) domain (Figure 1-6). The members of this family are BRD containing protein 2 (BRD2), BRD3, BRD4 which are ubiquitously expressed, and BRDT which is only expressed in testis. Inhibition of this family has been suggested to be beneficial in the treatment of cancer, inflammation, immunology and male contraception[57-59].

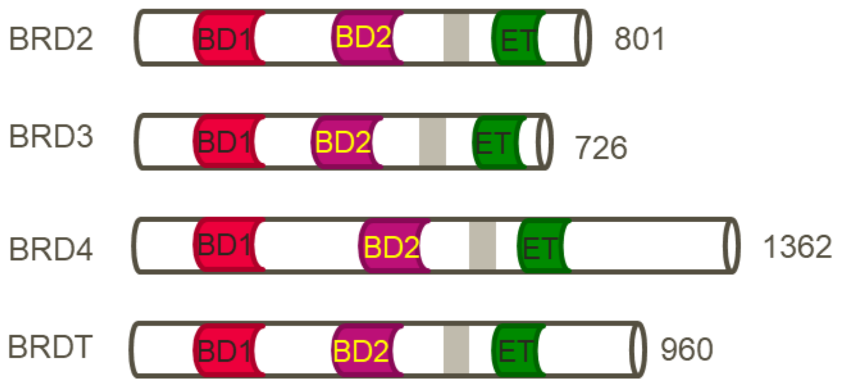


Figure 1-6 Representation of the four BET proteins: BRD2, 3, 4 and T. The BD1, BD2 and Extra-Terminal domains are reported as well as the number of amino acids in the primary structure (on the right).

NMR and X-ray crystal structures revealed that all the BRDs have four left-handed α-helices packed in an antiparallel bundle termed the 'BRD fold' linked with two loops ZA ($\alpha_Z - \alpha_A$) and BC ($\alpha_B - \alpha_C$) of variable lengths that constitute the hydrophobic pocket able to recognize the acetylated lysine residues[56, 60, 61]. In the BRD family, the most common variations are seen in the sequence of the ZA and BC loops but the amino acid residues involved in the acetyl-lysine recognition are among the most conserved residues[62]. The most important feature of the BRDs are described in the following section. Adjacent to the ZA channel there is a lipophilic region [63, 64] that in many BRDs (BRD2,3,4 and BRDT) is formed by Tryptophan-Proline-Phenylalanine (WPF) as in the specific case of BRD4 (Figure 1-7). An additional characteristic feature of the acetyl-lysine recognition pocket in BRDs is the presence of a network of water molecules that forms hydrogen bonds with carbonyl groups of the protein backbone at

GSK CONFIDENTIAL INFORMATION – DO NOT COPY

the base of the pocket. These water molecules, which are an integral part of the acetyl-lysine binding pocket, and the residues in the deeper part of the acetyl-lysine binding pocket are conserved[54, 55] over most of the BRD family.

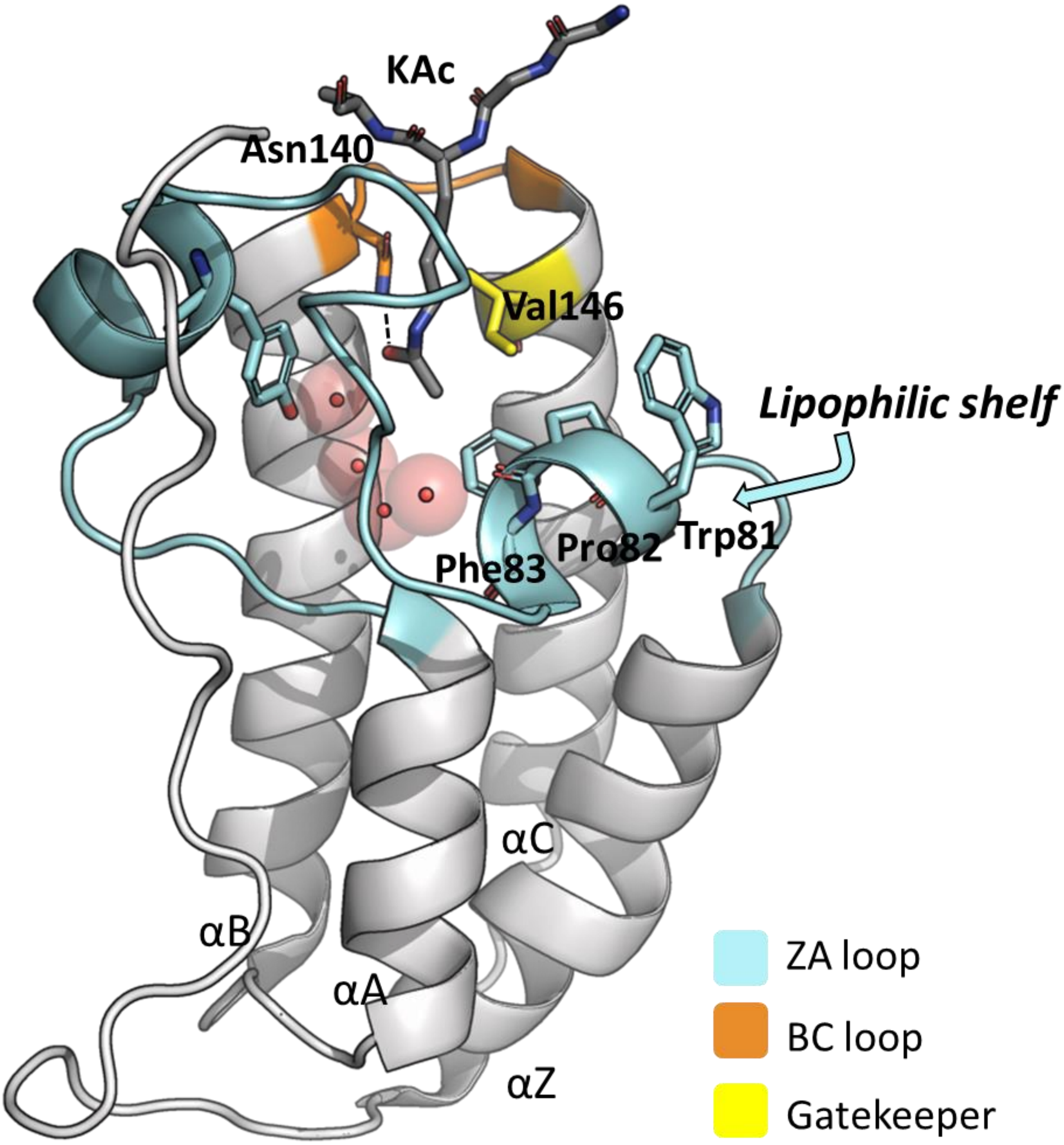


Figure 1-7 Overview of BRD4 protein (PDB code 3JVK) in complex with KAc. The conserved Asn140 residue required for binding acetylated lysine (hydrogen bond in dashed black line), the gatekeeper (Val146) and the residue in the lipophilic shelf (Trp81, Pro82, Phe83) are in stick representation as well as the conserved water network at the bottom of the pocket.

GSK CONFIDENTIAL INFORMATION – DO NOT COPY

1.5.1 In silico MD approaches to the study of Bromodomains

Several *in silico* tools have been used in recent years to advance the knowledge of new small molecules able to bind to BRDs and the dynamical behaviour of BRDs in the presence of inhibitors. In this section a brief overview of selected studies in which primarily MD simulations were employed is reported.

In 2013 Steiner *et al.*[65] studied with MD simulations the flexibility of several apo BRDs showing a heterogeneous flexibility of the ligand binding sites. Multiple replicas were used to reveal that the binding sites, especially the ZA and BC loops, are highly flexible and that the binding pocket can be occluded by some residues although no ligand induced effects were studied. The same group also investigated the spontaneous and reversible binding of KAc for the TAF1 BRD using 24 x 0.5 μs replicas with the intention to provide an atomistic description of the binding pathway and ultimately to enable the design of new small molecules[66]. Interestingly, they identified that the acetyl-lysine can bind in two different modes which are separated by an energy barrier that is smaller than the unbinding one. The first mode corresponds to the one observed in the crystal structure with the formation of a hydrogen-bond to the Asn140 while the second showed a buried KAc that is deeper in the binding site.

MD simulations can be of great value to interpret biological results. For example, short MD simulations were used to inform on design strategies for isoxazole derivatives as BRD4 inhibitors with improved affinity and metabolic stability[67]. From this study, the activity loss observed in the series of the compounds could be attributed to the

GSK CONFIDENTIAL INFORMATION – DO NOT COPY

disruption of a favourable intramolecular hydrogen bond which is stabilizing the bioactive conformation.

The combination of MD and end-point calculations seems to provide additional information when studying and comparing ligands activity. For example, it is common to use an implicit solvent model, either Poisson-Boltzmann (PB) or Generalized-Born (GB), coupled with a surface area term to calculate ligand-binding affinities from MD snapshots. These Molecular Mechanics-based (MM) approaches are called MM-PBSA or MM-GBSA[68].

MD simulations of the μs length and MM-PBSA/MM-GBSA were used to suggest differential dynamic properties between the BC and ZA loop with the aim to describe the selectivity of RVX-208 (see Figure 1-8), towards the BRD4-BD2 domain over BRD4-BD1[69]. At the same time, to study the same enhanced selectivity of RVX-208 towards BRD4-BD2 a QM/MM approach[70] in which a small portion of the macromolecular system is treated quantum mechanically and the remainder with molecular mechanics, was reported. In both cases the same conclusion was reached; that the ZA loop in BD2 displays decreased flexibility when in the presence of the ligand, and two residues, His437 and Val439, were identified as the most important for the observed selectivity. A combination of steered MD and MM-PBSA/MM-GBSA was also used after a virtual screening approach to identify new BRD4 ligands that provided insights on the structural requisite for binding[71].

GSK CONFIDENTIAL INFORMATION – DO NOT COPY

The differential activity between the two enantiomers of the ligand JQ1 (see Figure 1-8) were studied with MD, QM/MM simulations and a combination of umbrella sampling and steered MD[72]. The high flexibility of the ZA loop is found to be important to accommodate the ligand in the pocket and more interestingly that it is the kinetics or better the difficulties in entering the binding pocket that makes (-)-JQ1 enantiomer inactive. In addition, from the umbrella sampling a plausible explanation of unbinding of the ligand enantiomers was given which consisted of several steps involving a closed-to-open rearrangement of the ZA loop.

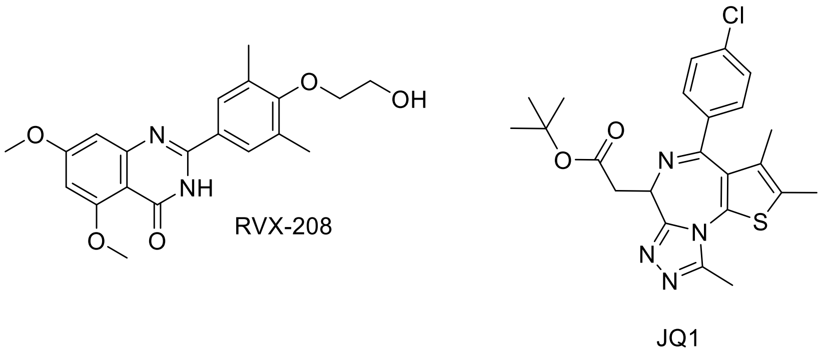


Figure 1-8 Structures of RVX-208 (on the left) and JQ1 (on the right).

Lastly, advances in predicting the binding affinity of BRDs ligands have been made by several groups either by using alchemical transformations to calculate absolute binding free energies[73] or with a combination of MD and MM-PBSA machinery[74].

GSK CONFIDENTIAL INFORMATION – DO NOT COPY

1.6 Aims and Objectives

Productivity is one of the most significant challenges that the pharmaceutical industry is facing. It takes on average between 10-15 years together with about $2 billion to bring a new drug on the market[2]. It is interesting to note that the majority of drug discovery projects fail (only about 35 % succeed) for different reasons such as unclear biology, difficulties in identifying a lead compound, poor potency or selectivity, lack of efficacy or animal toxicity. A long standing goal of computational chemistry is to accelerate the drug discovery process and guide medicinal chemists in a rational and quantitative way. A universally best computational technique does not exist and many are used depending on the different stages of the drug discovery process. Similarly, also in this thesis, different computational tools have been used for each of the specific problems to be addressed.

As mentioned in the introduction, water molecules and ligand-protein dynamics play a relevant role in SBDD and cannot be neglected. The present work has been conceived with the aim to contemplate the issue of both water placement and of the existence of different ligand binding modes acknowledging that a single static crystal structure might not always represent the highly dynamic solvated environment.

In this thesis several computational methods have been investigated with the aim to provide solutions and insights into drug design problems such as i) the placement of water molecules in protein cavities; ii) the identification of incorrect ligand binding modes in X-ray crystal structures; iii) the optimization of molecules to achieve

GSK CONFIDENTIAL INFORMATION – DO NOT COPY

selectivity towards a particular target; iv) the investigations on a ligand presenting multiple binding modes in the available X-ray crystal structures. Furthermore, at the beginning of each chapter a more detailed description of the specific aim is provided.

1.7 Outline and Thesis Structure

This thesis is divided into seven chapters; the background and theory are firstly discussed. The rest of the chapters contains the results of novel research in the field of computational chemistry applied to SBDD.

Chapter 3 relates to the problem of correctly predicting the water network in the cavities of the binding site in BRD4-BD1, an important and promising target in drug discovery.

In Chapter 4, the accuracy of crystal structure models is tested with an enhanced sampling method, Binding Pose MetaDynamics, that has the advantage of providing a quantitative assessment from full-atomistic simulations in a computationally efficient manner.

In Chapter 5 an example of how simulations can help to inform the progression of a drug discovery project is reported, specifically towards the identification of a BRD4-BD1 probe.

In Chapter 6, a case of a ligand co-crystallized multiple times showing two alternative binding mode is computationally investigated in detail.

GSK CONFIDENTIAL INFORMATION – DO NOT COPY

Lastly, an overall summary of the results from each chapter and future challenges are discussed in Chapter 7.

GSK CONFIDENTIAL INFORMATION – DO NOT COPY

Chapter 2 Theory

2.1 Force Fields

Theoretical studies of biological molecules aim to study the relationship between structure, function and dynamics at the atomic level. Most of the problems that need to be studied in biological systems involve many atoms and it is not feasible to treat them using Quantum Mechanics (QM). In Molecular Mechanics (MM) the electronic motions are ignored, and the energy of the system is calculated as a function of the nuclear position only (potential energy surface). In this way, in MM several assumptions need to be valid. First, the Born-Oppenheimer approximation which enables the movement of the nuclei to be separated from the movement of the electrons, without which it would not be possible to write the energy as a function of the nuclear coordinates. Then, the atoms are treated as "balls" and the bonds as "springs" such that the potential energy of a molecule can be written as a sum of terms involving bond stretching, angle bending, dihedral angles and nonbonded interactions.

The set of equations and parameters that define the potential energy surface of a molecule is referred to as Force Field (FF).

A typical FF for a system on N atoms with masses m_i and Cartesian position vectors r_i has the following form:

$$V = \sum_{i}^{bonds} \frac{k_{l,i}}{2}(l_i - l_0)^2 + \sum_{i}^{angles} \frac{k_{\alpha,i}}{2}(\alpha_i - \alpha_0)^2 + \sum_{i}^{torsions} \left\{ \sum_{k}^{M} \frac{V_{i,k}}{2}[1 + \cos(n_{ik}\theta_{ik} - \theta_0)] \right\} + \sum_{i,j}^{pairs} 4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6} \right] + \sum_{i,j}^{pairs} \frac{q_i q_j}{4\pi\varepsilon_0\varepsilon_r r_{ij}} \qquad (2.1)$$

the first three terms on the right hand side of the equation are also called the “bonded” terms (Eq.2.1), they describe variations in potential energy as a function of bond stretching, bending and torsions between atoms directly involved in bonding relationships. In the equation, l_i is the bond length, l_0 is the reference energy bond length and k_l is the force constant changing with the particular bond type; α_i is the angle, α_0 is the reference valence angle and k_α is the force constant. The third equation term represents the torsional rotation of four atoms around the central bond which is assumed to be periodic and is described by a cosine function. The last two equations are referred to as “non-bonded” terms (Eq.2.1). More specifically, the fourth and fifth terms are the van der Waals and Coulombic interactions between atom *i* and *j* at a distance r_{ij}. The ε_{ij} and σ_{ij} are the typical constants defining the 12-6 Lennard-Jones potential which represent the depth of the energy well and the distance at which the intermolecular potential between two particles is zero (van der Waals radius). The q_i and q_j are the partial charges of a pair of atoms, ε_0 is the permittivity of free space or electric constant and ε_r is the relative permittivity or dielectric constant (that in the vacuum is 1). In some FF additional terms are included such as the improper torsional terms that is the potential for moving one of the atoms out of the plane spanned by the three other atoms. This term is useful for example in the case of ensuring planarity e.g. aromatic systems or carbonyl groups. Lastly, to reflect coupling between internal coordinates, cross terms can be included.

Some of the most popular FFs that are widely used for biomolecular simulations are Assisted Model Building and Energy Refinement (AMBER)[75], Chemistry Assisted at HARvard Molecular Mechanics (CHARMM)[76], Optimized Potential for Liquid Simulations (OPLS)[77] and GROMOS[78]. The major advantage in using a FF is that it speeds up calculations as compared to quantum mechanics (QM).

2.2 Potential Energy Surface and Energy Minimization

The Potential Energy Surface (PES) describes the energy of a systems in terms of the positions of the atoms. Changes in the energy of the system can be considered as movement onto this surface. For a system with *N* atoms the PES is a function of 3N – 6 internal or 3N Cartesian coordinates. Therefore, it is impossible to visualize it unless only a set of specific coordinates are used to generate a slice of the PES giving a

GSK CONFIDENTIAL INFORMATION – DO NOT COPY

reduced-dimensionality energy surface. A schematic of a typical PES is provided in Figure 2-1.

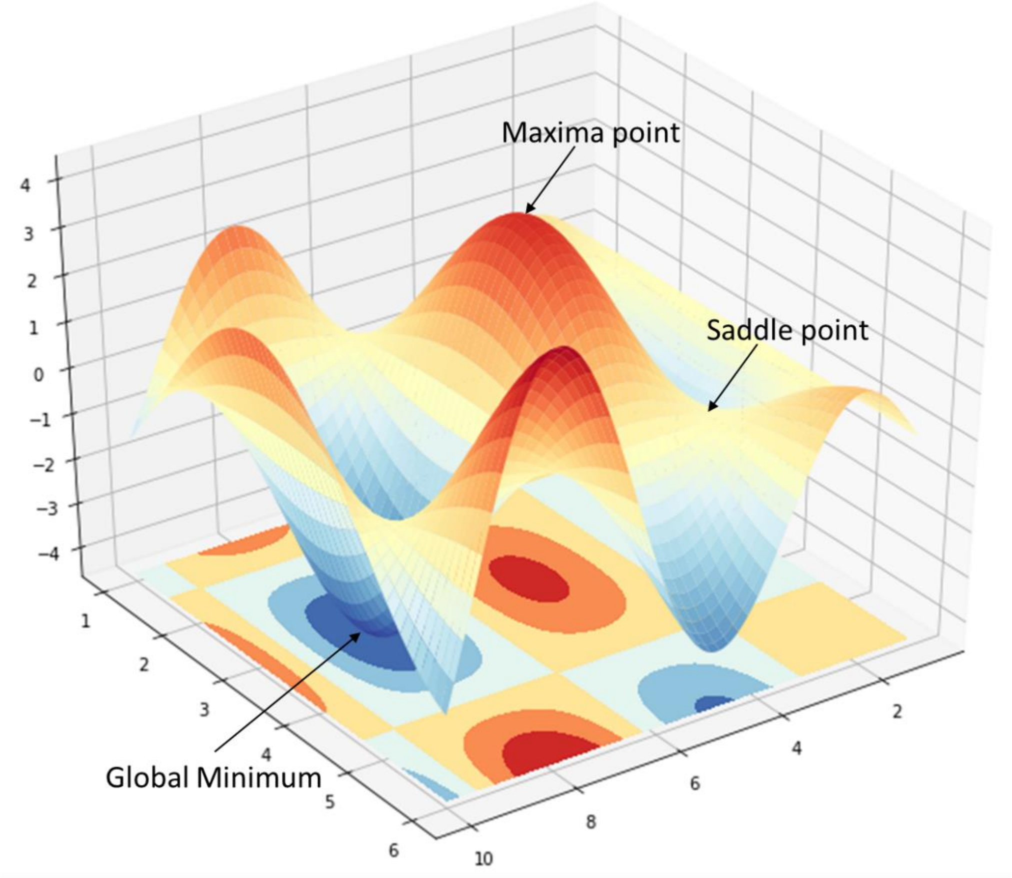


Figure 2-1 Example of a simple Potential Energy Surface.

The most interesting points in a PES are minima and saddle points. The minima points correspond to optimized low energy molecular structures in which the first derivative of the energy is zero with respect to the internal or Cartesian coordinates. In a PES there might be a very large number of minima; the point with the lowest energy is called *global energy minimum*. Minima can be connected through high energy points that can be either maxima or saddle points. The saddle points are stationary points on the energy surface characterized by having no slope in any direction, downward curvature for a single coordinate and upward curvature for all the other molecular

GSK CONFIDENTIAL INFORMATION – DO NOT COPY

coordinates (first-order saddle point). In the PES, the saddle points are defined as the lowest energy barriers on paths connecting minima. To identify the geometries that are in a minimum in the energy surface, energy minimization (or "geometry optimization") algorithms can be applied. During the minimization process, the geometry of the system is modified by small increments, until the total energy of the system reaches a minimum. Energy minimisation is an integral part of conformational analysis techniques which aim to find the low-energy conformations that a system can adopts. Most conformational search methods can be classified as either *systematic* or *stochastic*. Both approaches are applied also in molecular docking programs (Section 2.4). In the *systematic* conformational search new structures are generated by gradually changing the ligand conformation. For example, by using two rotatable bonds each torsional angle is incremented by a user defined value; the dihedral value is then kept fixed while the molecule is energy-minimized with respect to all other degrees of freedom. This results in a potential energy map which informs not only on the energy minima but also on energy barriers and pathways. This method is normally only used for problems involving a few rotatable bonds because it does not scale well to larger systems. In the *stochastic* conformational search, an iterative procedure is followed in which a structure is selected from those previously generated, randomly modified and then minimised. The resulting minimized conformation is kept if considered new compared to those already stored. This procedure can be repeated many times and the completeness of the search can be estimated by the number of times a stored conformation is found, or by whether new low conformers are being found. The process can be made more efficient by applying stochastic search algorithms (e.g. genetic algorithms or Monte Carlo). In general, the number of energy minima might

GSK CONFIDENTIAL INFORMATION – DO NOT COPY

be so large that is impractical to successfully identify all of them with either searching algorithm. Conformational searches are concerned with locating structures at energy minima only, whereas, if the interest is in predicting the behaviour of the system e.g. how the interconversion between closely related minima occurs, then computer simulations method such as Molecular Dynamics should be used.

2.3 Molecular Dynamics

In classical MD, a trajectory which represents the molecular configurations of the systems as a function of time, is generated by simultaneous integration of the second order differential Newton's equations of motion for all atoms in the system[79] (Eq. 2.2):

$$\boldsymbol{F}_i(t) = m_\mathrm{i}\,\boldsymbol{a}_\mathrm{i}(t) = m_\mathrm{i}\frac{d\boldsymbol{v}_\mathrm{i}}{dt} = m_\mathrm{i}\frac{d^2\boldsymbol{x}_\mathrm{i}}{dt^2} \qquad (2.2)$$

where F_i is the force acting on atom i of the system at time t, m_i is the mass, a_i is the acceleration, v_i is the velocity, t is the time and x_i is the position of the atom i at time t.

In MD simulations, the forces acting on the atoms are calculated using the classical mechanics description of forces through the gradient of the potential energy ($V(\boldsymbol{x})$):

$$\boldsymbol{F}_\mathrm{i}(t) = -\frac{\partial V(\boldsymbol{x}(t))}{\partial \boldsymbol{x}_\mathrm{i}} \qquad (2.3)$$

where $\boldsymbol{x}(t)$ is the vector describing the position of the N interacting atoms in the Cartesian space $(\boldsymbol{x} = x_1, y_1, z_1, x_2, y_2, z_2, \ldots x_N, y_N, z_N)$ at time *t*.

The potential energy function or Force Field (FF) works well for describing forces acting on particles such as nuclei, but it cannot describe the electron motion (as mentioned in 2.1). The accuracy of the simulation and of the prediction of the system's properties are directly related to the FF used to describe the interactions between particles.

The potential energy is a function of the atomic positions of all the atoms in the system; under the influence of a continuous potential the motions of all the particles are coupled together and this leads to a many-body problem that cannot be analytically solved. Therefore, numerical methods must be used to split the integration of the equations of motion into discrete time intervals, called time steps, δt. Forces are assumed to be constant during each integration step with the use of a small enough δt. Many numerical algorithms have been developed for integrating the equations of motion[80]: Verlet, velocity Verlet, LeapFrog, Beeman, Gear predictor-corrector.

The Verlet algorithm[81] is the simplest method for integrating the equation of motion. It uses the positions and accelerations at time *t*, and the position from the previous step $\boldsymbol{x}(t\text{-}\delta t)$ to calculate the new positions at $t+\delta t$. The Verlet algorithm can be derived using the following Taylor expansion of the atomic coordinates:

$$\boldsymbol{x}(t+\delta t) = \boldsymbol{x}(t) + \frac{d\boldsymbol{x}(t)}{dt}\delta t + \frac{\delta t^2}{2}\frac{d^2\boldsymbol{x}(t)}{dt^2} + \frac{\delta t^3}{6}\frac{d^3\boldsymbol{x}(t)}{dt^3} + \vartheta(\delta t^4) \quad (2.4)$$

where the first derivative of the position with respect to time is the velocity (*v*), the second derivative of the position with respect to time is the acceleration (*a*), *b* is the third derivative, and so on. Therefore, Eq. 2.4 can be written more succinctly as:

$$\boldsymbol{x}(t+\delta t) = \boldsymbol{x}(t) + \delta t\boldsymbol{v}(t) + \frac{1}{2}\delta t^2\boldsymbol{a}(t) + \frac{1}{6}\delta t^3\boldsymbol{b}(t) + \vartheta(\delta t^4) \quad (2.5)$$

A similar equation can be written for the atomic coordinates at the previous time step, ***x****(t-δt)*:

$$\boldsymbol{x}(t-\delta t) = \boldsymbol{x}(t) - \delta t\boldsymbol{v}(t) + \frac{1}{2}\delta t^2\boldsymbol{a}(t) - \frac{1}{6}\delta t^3\boldsymbol{b}(t) + \vartheta(\delta t^4) \quad (2.6)$$

Adding equations 2.5 and 2.6 gives the expression that is used to calculate ***x****(t+δt)* in the Verlet algorithm:

$$\boldsymbol{x}(t+\delta t) = 2\boldsymbol{x}(t) - \boldsymbol{x}\,(t-\delta t) + \delta t^2\boldsymbol{a}(t) + \vartheta(\delta t^4) \quad (2.7)$$

The velocities are not explicitly calculated but can be calculated by dividing by 2δt the difference in position at time t+δt and t-δt:

$$\boldsymbol{v}(t) = \frac{\boldsymbol{x}(t+\delta t) - \mathbf{x}(t-\delta t)}{2\delta t} \quad (2.8)$$

This implies that velocities are not available until the positions have been computed at the next step. One disadvantage of the Verlet algorithm is that two sets of positions,

GSK CONFIDENTIAL INFORMATION – DO NOT COPY

x(t) and *x(t-δt)*, have to be stored in memory. An alternative is the so-called *velocity Verlet algorithm*[82]:

$$\boldsymbol{x}(t+\delta t) = \boldsymbol{x}(t) + \boldsymbol{v}(t)\delta t + \frac{1}{2}\delta t^2 \boldsymbol{a}(t) \qquad (2.9)$$

$$\boldsymbol{v}(\mathrm{t}+\delta t) = \boldsymbol{v}(t) + \frac{1}{2}\delta t[\boldsymbol{a}(t) + \boldsymbol{a}(t+\delta t)] \qquad (2.10)$$

A general strategy for the above algorithm is the following:

1. Given x*(t)* and *v(t)*, first compute the forces on each atom using force field.
2. Update the positions *x* at *t+δt*.
3. Partial update of the velocity based on the current forces (gives *v* at half-time step).
4. Compute the new forces using the new position
5. Complete the velocity update
6. Go back to 1

The leap-frog algorithm[83] is equivalent to the Verlet algorithm but solves velocities at half time step intervals:

$$\boldsymbol{x}(t+\delta t) = \boldsymbol{x}(t) + \ \boldsymbol{v}\left(t+\frac{\delta t}{2}\right)\delta t \qquad (2.11)$$

GSK CONFIDENTIAL INFORMATION – DO NOT COPY

$$\boldsymbol{v}\left(t+\frac{\delta t}{2}\right)=\boldsymbol{v}\left(t-\frac{\delta t}{2}\right)+\boldsymbol{a}(t)\delta t \qquad (2.12)$$

The position and the velocities are not synchronized making it difficult to evaluate the total energy (kinetic + potential) at any point in time.

Each integration algorithm must preserve the physical properties of the equation of motion such that the total energy, which is the sum of kinetic and potential energy, can be kept constant. Without further modifications, MD simulations sample a statistical ensemble of microstates that are characterized by constant number of particles (N), constant volume (V), and constant energy (E) that is also known as microcanonical ensemble. It is also possible to control the system's temperature and pressure with the use of a thermostat or a barostat, and these can be applied to perform simulations in other thermodynamics ensembles such as the canonical ensemble (constant NVT) or isobaric-isothermal ensemble (constant NPT)[80]. In the canonical ensemble, the thermostat modifies the Newton's equations of motion, rescaling the velocities of the particles by ensuring that the average temperature of the system is the desired one; the energy is not conserved but fluctuates around a constant value. A variety of thermostat methods are available to add and remove energy from a system. Popular techniques to control temperature include velocity rescaling, the Nosé-Hoover thermostat[84] and the Berendsen thermostat[85]. The barostat in the isobaric-isothermal ensemble, rescales the positions of the particles causing the volume to fluctuate; also in this case, the energy is not conserved but oscillates around a constant value.

In order to have a more realistic model of a protein-ligand complex in solution, it is necessary to include a large number of solvent molecules along with the solute. This is achieved with the use of periodic boundaries condition (PBC) in which the particles being simulated are enclosed in a box which is then replicated in all three dimensions to give a periodic array. Coordinates and velocities are stored and propagated for the unit cell only, although, the evaluation of non-bonded terms has to be extended to every pair of atoms in the unit cell and periodic images. In doing so, the computational cost is significantly increased therefore the choice of an appropriate cut-off for the non-bonded interactions is necessary. Usually, for the calculation of the short-ranged van der Waals terms a spherical cut-off of 10 Å is used. The electrostatics contribution is instead calculated with the Ewald sum methods in which short-range interactions decay quickly and are negligible beyond some user-defined cutoff distance, and the long-range interaction decay more slowly. However, the Ewald summation is inefficient for large systems so to improve the efficiency of the electrostatic calculations the particle mesh Ewald (PME) method[86] can be used. The interactions are separated into short and long range. The direct interactions between two particles are replaced by two separate summations, a short range potential in real space of the simulation box and a long range potential in the Fourier space of the periodic boxes.

One of the main goals of MD simulations is to estimate/predict thermodynamic behaviour of real systems as observed in laboratory. Thermodynamic observables can be calculated from a MD simulation if the system is *ergodic* meaning that, given an infinite amount of time, the system will be able to explore all the possible configurations such that the *time average* and the *ensemble average* are identical. In general, the instantaneous value of an observable *O* of a system can be calculated by a *time average*:

$$O = \frac{1}{T}\int_{t=0}^{T} O(\boldsymbol{x}(t),\boldsymbol{p}(t))\,\mathrm{d}t \qquad (2.13)$$

where T is the duration of the experiment and ***x(t)***, ***p(t)*** represent the position and momenta of the state at time *t*. The true average value (denoted with <..>) is then given by the limit:

$$\langle \mathrm{O}\rangle_{time} = \lim_{T\to\infty}\frac{1}{T}\int_{t=0}^{T} O(\boldsymbol{x}(t),\boldsymbol{p}(t))\,\mathrm{d}t \qquad (2.14)$$

From an ensemble, a set of systems with the same types of degrees of freedom, by assuming that each state of the system has a statistical weight also referred to as probability distribution (*P*) on the phase space, it is possible to calculate the ensemble average of the observable *O*:

$$\langle \mathrm{O}\rangle_{ensemble} = \int \mathrm{O}(\boldsymbol{x},\boldsymbol{p})\mathrm{P}(\boldsymbol{x},\boldsymbol{p})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{p} \qquad (2.15)$$

The probability, *P(**x**, **p**)*, determines the extent to which the value *O(**x**, **p**)* contributes to the average. In principle the equivalence between *ensemble average* and *time average* offers a valid method to obtain thermodynamic properties. In fact, from MD simulation the thermodynamic averages are obtained as time averages using numerical integration of Eq.2.15

GSK CONFIDENTIAL INFORMATION – DO NOT COPY

$$\langle O \rangle = \frac{1}{M} \sum_{i=1}^{M} O\,(\boldsymbol{x}, \boldsymbol{p}) \qquad (2.16)$$

M is the number of steps. However, the computation of such quantities from an MD simulation remains far from trivial due to the limit of infinite sampling and to numerical errors.

2.3.1 MD limitations

Two of the main limitations of a MD simulation relate to force field issues and time.

2.3.1.1 Accuracy of the Force Fields (FFs)

The results of a simulation will be reliable only if the potential energy function mimics the forces experienced by the atoms as closely as possible to the “real” ones. At the same time, the potential energy has to be a simple function so that the calculation of the forces does not become too time consuming; an example of potential energy function or FF is reported in Section 2.1.

The accuracy of the FF depends on the form of the function and on the quality of the parametrization. This is usually done by combining available experimental data with the results of high level *ab initio* calculations on model systems that can be used as building blocks for macromolecules. Moreover, in the FF, the description of the same atom or functional group as found in different chemical settings is represented by different atom types. As a consequence, the transferability of the force field is highly restricted.

MD simulations with classical FFs cannot be used to treat phenomena in which quantum effects occur such as changes in chemical bonding, the presence of noncovalent intermediates and transfer of protons or electrons. Possible alternatives are for example the use of the bond-order depended FF called ReaxFF[87] (reactive force-field) that allows to describe bond breaking/formation or the use of QM/MM simulations[88]. In these simulations, which are especially useful in modelling enzymatic reactions, only a very limited portion of the system where the reaction of interest occurs is treated with high accuracy (QM) while the majority of the system is treated at a less accurate level of theory (MM). Another important challenge for FFs is accounting for electronic polarization. To overcome this limitation, polarizable FFs[19] have been developed which can mimic the electronic redistribution in response to an external electric field. Despite the advantage in the polarizable FFs they are still computationally demanding and less user-friendly than the fixed-charge ones.

2.3.1.2 Time limitations

The time limitation is a severe problem in MD simulations. At the moment, hundreds to millions of atoms can be studied for several nanoseconds to microseconds in a typical MD simulation[89]. Although, this is an impressive improvement over the first molecular dynamics simulations, such a time limit might still not be sufficient for studying certain quantities. For example, protein folding, ligand binding and unbinding processes are mostly at an inaccessible timescale for classical MD simulations. In particular, it is possible that when studying slow processes that require the overcoming of high energy barrier, biological systems get trapped in deep energy wells of their potential energy surface causing insufficient sampling and/or exploration of other conformations[90]. As will be discussed in 2.3.3, a strategy to resolve this issue is

GSK CONFIDENTIAL INFORMATION – DO NOT COPY

presented by enhanced sampling methods in which an external bias, such as an external force, is applied to the system to explore more efficiently the potential energy surface.

2.3.2 Free Energy

Considering a system of particles of coordinates *x* coupled to a thermostat bath at temperature T, which evolves under the influence of a potential *V(**x**)* following the canonical distribution, the probability of observing a given state of ***x*** is:

$$P\ (\boldsymbol{x}) = \frac{1}{Z}\mathrm{e}^{-\beta V(\boldsymbol{x})} \qquad (2.17)$$

where $\beta = 1/k_B T$ and $Z\ = \int exp(-\beta V(\boldsymbol{x})\ )\ d\boldsymbol{x}$, is the normalization factor such that the probability will sum to one and it is called the partition function of the system. The probability distribution $P(\boldsymbol{x})$ is called the Boltzmann probability distribution. For systems like proteins, which are characterized by a huge number of degrees of freedom, the space on which the probability distribution is defined is huge so that its computation for every possible state is impossible because of the incredibly large dimensionality. To allow a simpler description of the system, the probability distribution can be considered in terms of some reaction coordinate or collective variables (CV), $s(\boldsymbol{x})$. In this way, at equilibrium the probability to observe the system at a given point *x* on the CV is the sum of the probabilities of all configurations *s* which map to $s(\boldsymbol{x})$:

GSK CONFIDENTIAL INFORMATION – DO NOT COPY

$$P(\boldsymbol{s}) = \frac{1}{Z} \int \mathrm{e}^{-\beta V(x)} \delta\big(\boldsymbol{s} - \boldsymbol{s}(\boldsymbol{x})\big)\, d\boldsymbol{x} \qquad (2.18)$$

Where the Dirac δ function selects only the configurations corresponding to a specific value $s(\boldsymbol{x})$ of the collective variable.

For the canonical ensemble, the Helmholtz free energy, F, is defined as:

$$F = -k_B T \, ln\, Z \qquad (2.19)$$

where Z is the partition function defined above. The free energy, can then be expressed in terms of the probability distribution on the reduced coordinate, *s*:

$$F(\boldsymbol{s}) = -k_B T \, ln(P(\boldsymbol{s})) \qquad (2.20)$$

Eq. 2.20 allows the description of the equilibrium properties of the system as a function of relevant and carefully chosen set of variables.

A typical approach to compute free energy along a CV from a MD simulation is by using a histogram of the visited configurations:

$$F(\boldsymbol{s}) = -k_B T \, ln\, N(\boldsymbol{s}) \qquad (2.21)$$

where *N(**s**)* is the number of times the value *s* of the CVs has been explored. However, the use of this relationship in the biomolecular context is very often limited by the lack of enough statistics to construct a reliable histogram in the estimation of *P(**s**)*. This is especially true if the biologically relevant states of interest are separated by high energy barriers that are poorly sampled during the simulation. The free energy, together with entropy-related quantities (chemical potential and entropy itself), is the thermodynamic observable whose estimation suffers the most from sampling limitations. As functional biomolecular timescales tend to be in the μs-ms timescale and beyond, observing them in the timescale accessible through MD simulations is very challenging. Therefore, enhanced sampling techniques have been developed as a potential solution to tackle this problem. In the next section enhanced sampling methods are discussed with a focus on Metadynamics.

2.3.3 Metadynamics

Metadynamics[91] (MetaD) is an enhanced sampling technique aimed at accelerating the exploration of rare events and reconstructing the underlying free energy landscape as a function of a set of order parameters, referred to as Collective Variables (CVs). In this method, the sampling is accelerated by a history-dependent bias constructed as a sum of repulsive Gaussians in the CVs space.

A simple way to understand how MetaD works is to consider the one-dimensional free energy landscape in Figure 2-2. Once the CV has been identified, the system is simulated by conventional simulation for a small timestep (1-2 ps). Then, values of CVs are calculated and recorded as s_1. From this point, a bias potential in the form of a Gaussian hill centred in s_1 is added to the simulated system. The system evolves for another 1 or 2 ps, then another hill is added to s_1, and so on. These Gaussians are

GSK CONFIDENTIAL INFORMATION – DO NOT COPY

deposited along the system trajectory in the CVs space such that they encourage the system to explore configurations that have not been sampled before (Figure 2-2).

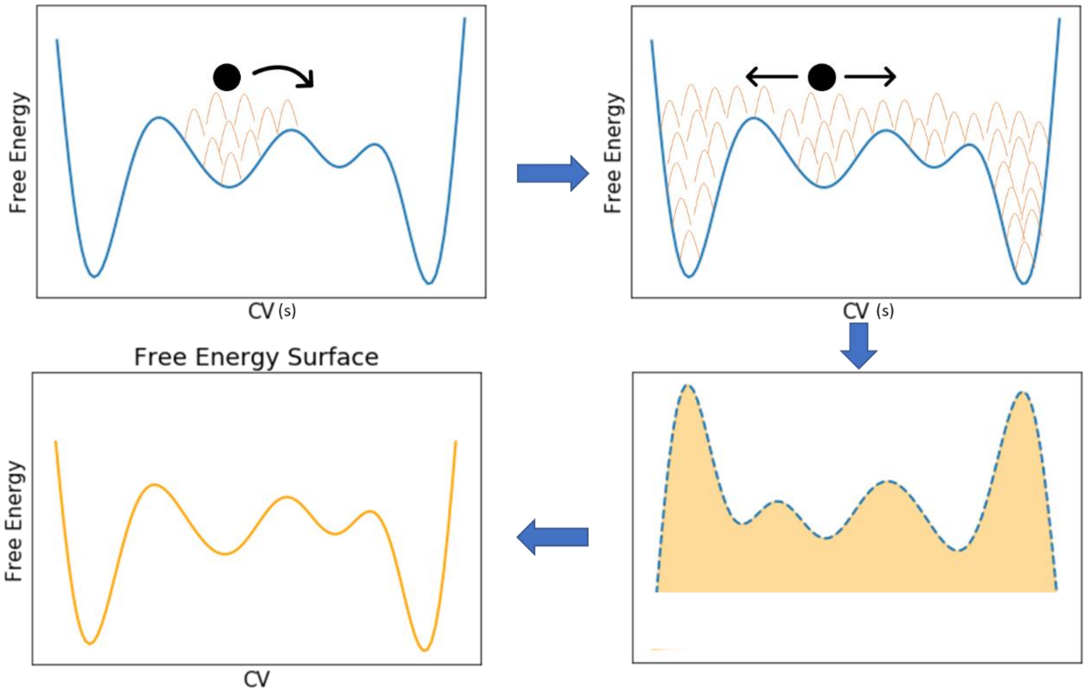


Figure 2-2 Schematic representation of the MetaD technique. Gaussians (orange) are added to the system (black dot) such that it will be possible to explore high-energy regions. The dynamics evolves in the modified free energy landscape until the sum of all the Gaussians will completely fill up all the basins. At that point, the system evolves in a flat landscape and the summation of all the bias deposited (gaussians) provides a negative estimate of the free-energy profile.

For a set of *d* CVs, ***s****i*(***x***), *i=1,2…d* where *d* is a small number and ***x*** is the set of microscopic coordinates of the system, the bias potential $V(\boldsymbol{s},t)$ added to the selected CV at the present time t and at all its past values t' < t, is defined as:

$$V(\boldsymbol{s},t) = \int_0^t dt'\omega\, exp\left(-\sum\nolimits_{i=1}^{d} \frac{\left(s_i(\boldsymbol{x}\,(t)) - s_i(\boldsymbol{x}\,(\mathrm{t}'))\right)^2}{2{\sigma_i}^2}\right) \qquad (2.22)$$

where *s_i(**x**(t'))* is the value taken by the CV at time t', ω is an energy rate and σ_i is the width of the Gaussian for the *i*-th CV. The CVs can be any function of ***x***, e.g. a distance or an angle, as follows:

$$\mathbf{s}(\mathbf{x}) \;=\; (\mathrm{s}_1(\mathbf{x}), \dots \mathrm{s}_\mathrm{d}(\mathbf{x})) \qquad (2.23)$$

The energy rate (ω) is constant and expressed as fraction between Gaussian height W and deposition stride τ_G:

$$\omega = \frac{W}{\tau_G} \qquad (2.24)$$

The bias is "history-dependent" because it is the sum of the Gaussians that have already been deposited in the CV space during time. The outcome of a MetaD simulations in the NVT ensemble is the Helmholtz free-energy as a function of the CV.

The Gaussian height *W*, width σ and deposition stride τ determine the accuracy and efficiency of the free energy reconstruction. If the Gaussians are large, the free energy will be explored at a fast pace, but the reconstructed profile will present large errors or even induce structural rearrangements that are artefacts due to the excess of energy flowing in the system. On the contrary, if Gaussians are too small or deposited infrequently, the reconstructed profile will be accurate, but it will take much longer to reach convergence. Typically, the width is chosen to be of the order of the standard deviation of the CV in a preliminary unbiased simulation in which the system explores a local minimum in the free energy surface. The bias potential is able to fill the minima

GSK CONFIDENTIAL INFORMATION – DO NOT COPY

in the free energy surface, allowing an efficient sampling of the space defined by the CVs.

The basic assumption in MetaD is that after a sufficiently long simulation time, $V(\boldsymbol{s}, t)$, as defined in Eq. 2.22, provides an estimate of the underlying free energy plus an irrelevant constant C:

$$V(\boldsymbol{s}, t \rightarrow \infty) = -\mathrm{F}(\boldsymbol{s}) + C \qquad (2.25)$$

In standard MetaD, constant Gaussians are added to the system in the whole course of the simulation. Therefore, two major limitations can be identified. First, it is difficult to decide when the simulation has converged and consequently when to stop it; secondly, the bias potential tends to overfill the underlying free energy and drives the system towards regions with higher energy, drifting the simulation toward nonrelevant configurations.

A solution to these problems is provided by well-tempered metadynamics[92], in which the bias potential $V(\boldsymbol{s}, t)$ has the same form as in standard MetaD (Eq. 2.22), but the Gaussian height (ω) decreases with increasing simulation time:

$$V\,(\boldsymbol{s}, t) = \sum_{t' = \tau, 2\tau, \ldots t' < t} W \exp\left(-\frac{V(s, t')}{k_B \Delta T}\right)$$
$$\exp\left(-\sum_{i=1}^{d}\left(\frac{(s_i(x(t) - s_i(x(t')))^2}{2{\sigma_i}^2}\right)\right) \qquad (2.26)$$

GSK CONFIDENTIAL INFORMATION – DO NOT COPY

where W is the initial Gaussian height, k_B is the Boltzmann constant, and ΔT is an input parameter with the dimension of a temperature. From equation 2.26, it can be understood that Gaussians of different height are added in the region of the CV space. In a well-tempered MetaD the bias potential $V(\boldsymbol{s}, t)$ can be related to the free energy $F(\boldsymbol{s}, t)$ of the system by:

$$V(\boldsymbol{s}, t \to \infty) = -\frac{\Delta T}{T + \Delta T} F(\boldsymbol{s}) + C \qquad (2.27)$$

The bias does not tend to become the negative of the free energy but instead a fraction of it ($\Delta T/(T + \Delta T)$). In the limit where ΔT→0, unbiased MD is recovered. For a finite ΔT, the system will sample a probability distribution proportional to $exp\left(-\frac{F(s)}{k_B(T+\Delta T)}\right)$ thus resulting in sampling the CVs at an effectively higher temperature $T + \Delta T$.

2.4 Docking

Docking is frequently used in SBDD; it consists of generating several conformations/orientations of the ligand also called ligand-binding poses within the protein binding site. Therefore, the availability of the 3D-structure of the protein is of fundamental importance. There are two basic components for each docking program 1) sampling of the different conformations of the ligand; 2) scoring of the generated ligand binding poses. The sampling process is done with a searching algorithm which should reproduce the experimental ligand binding mode. The scoring functions attempt to estimate binding affinity of the protein-ligand complex. Their main goal is to rank the correct docking solution highest among all the conformations generated by the sampling algorithm and in case of different ligands, to rank one ligand relative to another. Most of the docking programs can successfully predict the correct conformation of the ligand within the binding site but they are normally less accurate in reproducing the absolute interaction energy of the ligand-protein complex.

During the conformational stage, the structural parameters of the ligand (torsional, translational and rotational degrees of freedom) are incrementally modified into the binding site in order to identify the maximum of favourable intermolecular interactions. As described in Section 2.2, conformational search algorithms can be divided into *exhaustive* and *stochastic* search methods. In the exhaustive methods, the algorithm changes all the structural parameters until a minimum either local or global is found whereas in the stochastic search, the ligand is modified by random moves such that a broad coverage of the energy landscape is generated at a high computational cost[93].

GSK CONFIDENTIAL INFORMATION – DO NOT COPY

Once the poses are generated, molecular docking programs use scoring functions to estimate the binding energetics and finally rank the most promising results.

There are four types of scoring functions: 1) *force-field based*: the estimate of the binding energy is calculated by summing the contributions from bonded and non-bonded interaction from the force field equation, an example is reported in Eq.2.1; the major limitation is the difficulty in calculating the entropic contributions; 2) *empirical*: the functions are composed of multiple terms used to describe specific interactions or events associated with ligand-protein complexation such as hydrogen-bonding, ionic and apolar interactions, as well as desolvation and entropy terms. Those functions are fitted to experimental data using a series of known protein-ligand complexes as a training set. From the multiple linear regression analysis, weight constants are generated to be used as coefficients to adjust the terms of the main equation. In this way, they are strictly dependent on the quality of the test set but in general are faster than a force-field based ones; 3) *knowledge-based*: based only on frequency of atom pairs interactions observed in experimentally determined 3D structures of protein-ligand complexes; 4) *machine-learning-based*: a number of different machine-learning algorithms have been used such as, random forest and, support vector machines, to rescore poses generated by classical docking programs. The main difference to the other scoring functions is that the form of the function used to describe the ligand-receptor complex is not imposed on the algorithm and therefore allows more flexibility.

2.4.1.1 Grid-based ligand docking with energetics (GLIDE)

Glide[94, 95], a high-speed flexible docking software from Schrödinger, has been used as the docking program in this work. Glide uses a pre-computed van der Waals and electrostatic grid of the receptor and a highly efficient series of hierarchical filters for ligand conformational selection (Figure 2-3).

The first step in the docking experiment consists of the generation of a grid. Van der Waals and electrostatic potentials are evaluated at the vertices of a cubic grid using the OPLS3e force field. The origin and spatial extension of the grid is decided by the user usually as the centroid of a ligand with known binding orientation or as the centroid of the residues forming the binding site. Then, an extensive conformational search of the ligand with ConfGen[96] program is done after dividing the ligand into a core region and several rotamer groups. All the possible core conformations to which rotamer groups are attached in all possible permutations of conformations are generated and then passed on a set of hierarchical filters (Figure 2-3).

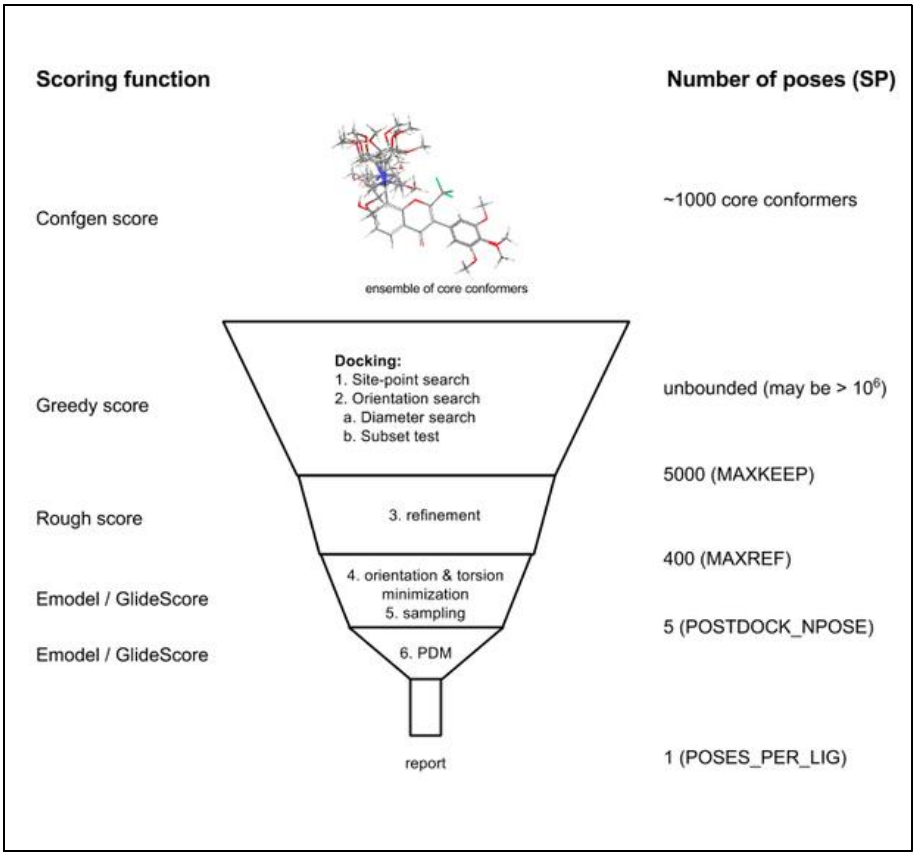


Figure 2-3 Glide-docking hierarchy. Reproduced from Glide User Manual in the Schrödinger 2018-04 Release.

The placement of the ligand in the binding site begins by choosing site-points on an equally spaced 2 Å grid on the active site region which serve as positions for the ligand centre. The ligand centre, defined as the midpoint of the line connecting the two most widely separated atoms in the core region (ligand diameter) is positioned in the grid and rotated about its diameter. Possible hydrogen bonds are scored and if the score is good enough all the rest of the interactions are also considered. Up to this point a discretized version of ChemScore[97] is used which is capable to recognize favourable hydrophobic, hydrogen bonding, metal-ligand interactions and penalizes steric clashes. This stage is called “greedy scoring” because the actual score for each atom depends on its position relative to the receptor but also on the best possible score it could get by moving it by ± 1 Å in all the directions. The final step is then to re-score the top greedy scoring poses via a "refinement" procedure (step 3, Figure 2-3) in which the ligand is moved as a rigid body by at most ±1 Å and the pose with the best interaction is passed on to the next filter. The resulting 100-400 poses are then subjected to force field minimization on the pre-computed electrostatic and van der Waals grids. Minimization begins with a pre-minimization step on smoothed grids and ends with a full-scale minimization on the original grids sampling translations, rotations and torsional motions of the ligand molecule. Finally, torsions are further sampled by a Monte Carlo procedure. The minimized poses are re-scored using the Schrödinger's proprietary *GlideScore* scoring function, an empirical scoring function that aims to maximize separation of ligand with strong affinity from those with little to no binding affinity. *GlideScore* is based on ChemScore, but includes a steric-clash term, adds other rewards and penalties such as buried polar terms (devised by Schrödinger to penalize electrostatic mismatches), amide twist penalties, hydrophobic enclosure terms, and excluded volume penalties, and has modifications to other terms:

$$GScore = 0.05 * vdW + 0.15 * Coul + Lipo + Hbond + Metal + Rews + RotB + Site \qquad (2.28)$$

where vdW refers to Van der Waals energy; $Coul$ to Coulomb energy; $Lipo$ to Lipophilic term; $HBond$ takes into account hydrogen bonds and it is separated in different weighted components that depends on whether the donor or acceptor are both neutral, one is neutral and the other charged or both charged; $Metal$ is the Metal-binding term in which the interactions with anionic or polar acceptor atoms are included. If the net metal charge in the apo protein is positive, the preference for anionic or polar ligands is included; if the net charge is zero, the preference is suppressed; $Rews$ includes rewards and penalties for various features, such as buried polar groups, hydrophobic enclosure, correlated hydrogen bonds, amide twists; $RotB$ represent the penalty for freezing rotatable bonds and $Site$ rewards situations in which a polar but not hydrogen bonding atom is found in a hydrophobic region.

The best-docked structure is given by the E_{model} score which combines the energy grid score and the binding affinity generated by GlideScore.

2.4.1.2 Induced Fit Docking Protocol with GLIDE

Schrodinger's Induced-Fit docking[98-100] protocol introduces protein flexibility into docking calculations. In this approach, side-chain degrees of freedom in the receptor are sampled while allowing minor backbone movements through minimization.

The extended sampling protocol uses the following steps:

1. *Initial Glide docking of each ligand using a softened potential and removal of side chains.* The residues for side chain removal are selected using properties such as solvent-accessible surface areas, B-factors (if present), and the presence of salt bridges. The potential is softened on a per-atom basis, using information from side-chain flexibility. Up to 80 poses are returned from several docking runs, some using a trimmed receptor, some using an untrimmed but softened receptor. The results of these runs are clustered to obtain representative poses, selected on the basis of the GlideScore and the descriptors used for clustering.

GSK CONFIDENTIAL INFORMATION – DO NOT COPY

2. *Prime side-chain prediction for each protein-ligand complex*. Performed on residues within a given distance of any ligand pose (default 5 Å), with optional inclusion or exclusion of other residues, followed by minimization of these residues and the ligand. Prime energy is calculated with a continuum solvation based molecular mechanics force field.

3. *Glide redocking of each protein-ligand complex structure*. The ligand is now rigorously docked, using default Glide SP settings, into the induced-fit receptor structure.

4. Scoring of each output pose with the scoring function called IDFScore. This scoring function is the sum between the GlideScore (empirical scoring function) from the redocking step and a contribution of the Prime energy (step 2).

GSK CONFIDENTIAL INFORMATION – DO NOT COPY

2.5 RISM and 3D-RISM

Reference Interaction Site Model (RISM) is a computational approach that calculates the distribution of solvent molecules around a solute; it has its roots in statistical mechanical integral equation theories (IET) of liquids. The IET of liquids consists of calculating the interactions between molecules in terms of density pair correlation functions (equivalent to determining a probabilistic structure of the solvent) by solving a set of equations based on the Ornstein-Zernike (OZ) equation and a closure relation. The OZ equation relates the total pair correlation function *h(r)* with the direct correlation function *c(r)*, for a liquid with density *ρ*; it is a convolution integral given by:

$$h(r) = c(r) + \rho \int c(|\mathbf{r} - \mathbf{r}'|) h(\mathbf{r}') d\mathbf{r}' \qquad (2.29)$$

The pair density distribution function *g(r)* is defined as:

$$h(r) = g(r) - 1 \qquad (2.30)$$

The OZ equation presents two unknown variables therefore a closure relation is needed to solve it and it is defined as follows:

$$h(r) = \exp(-\beta u(r) + h(r) - c(r) + B(r)) - 1 \qquad (2.31)$$

where $\beta = 1/k_B T$ and B is the bridge function.

GSK CONFIDENTIAL INFORMATION – DO NOT COPY

The OZ equation is only applicable to atomic liquids so in order to describe a non-spherical liquid the Molecular-OZ (MOZ) equation was provided in which both position (x, y, z) and orientations (ψ, θ, φ) of the particles are considered. The 6-Dimensional MOZ equation is not actively used to study molecular liquids because it is difficult (if not impossible) to solve, therefore a series of approximations that reduce the dimensionality of the MOZ equations have been developed starting from the work done by Chandler and Anderson in the seventies[101-106] which are now collectively referred to as RISM.

The 1D-RISM approach separates the solute and the solvent in a set of sites with spherical symmetry so that the correlation functions of the OZ equations depend only on the distance between those. It is extremely quick to calculate but it does not properly consider the spatial correlations of the solvent density around the solute due to the approximations in the method. In 1996 Berglov and Roux proposed a three dimensional extension of RISM, 3D-RISM[107]. 3D-RISM produces a solvent distribution around a rigid solute without the need for long MD or Monte Carlo simulations. The 3D solvent distribution can be calculated within minutes to hours from only solute structure, solvent structure and composition, and an intermolecular potential as input. After solving the 3D-RISM equations it is straightforward to calculate thermodynamics properties such as solvation free energy and partial molar volume. The three dimensional RISM (3D-RISM) is now one of the most used tools for investigating the liquid structure, for calculating physicochemical phenomenon but also recent applications are: calculating hydration thermodynamics of protein[108] and ions[109], computational drug design[110-112] and solvation effects.

For a solute-solvent system at infinite dilution, the 3D-RISM integral equation is written as:

$$\mathrm{h}_{\alpha}(\mathbf{r}) = \sum_{\xi=1}^{N_{\text{solvent}}} \int_{R^3} c_{\xi}(\boldsymbol{r}-\boldsymbol{r}')\chi_{\xi\alpha}(|\boldsymbol{r}'|)d\boldsymbol{r}' \qquad (2.32)$$

$$\alpha = 1, \dots, N_{solvent}$$

where subscripts α and ξ denote indexes of sites in solvent molecule (Figure 2-4), N is the total number of sites in a solvent molecule, h_α is the total correlation function, c_α is the direct correlation function and $\chi_{\xi\alpha}$ is the solvent susceptibility. It describes the correlations of the sites of solvent molecules in the bulk solvent and it is typically obtained by the solution of the 1D-RISM equations of the solvent only (without the solute present): $\chi_{\xi a}^{Solv}(r) = \omega_{\xi\alpha}^{Solv}(r) + \rho h_{\xi\alpha}^{Solv}(r)$.

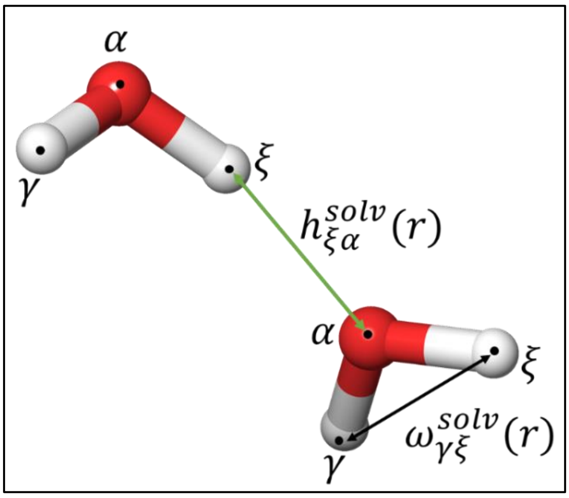


Figure 2-4 Solvent-solvent correlations in 1D-RISM and 3D-RISM methods. $\omega_{\gamma\xi}(r)$ represents the site-site intramolecular correlation functions and $h_{\alpha\xi}(r)$ the intermolecular correlation functions. Figure adapted from ref. [113]

The total pair correlation function is defined as:

$$h_{\alpha}(r) = g_{\alpha}(r) - 1 \qquad (2.33)$$

where g_{α} is the pair-distribution function, which gives the local density distribution of the solvent at grid points around the solute.

The 3D-RISM approximation has two unknown variables therefore, we need another equation that provides a connection between *h(r)* and *c(r)* or uniquely defines one of these functions. The general closure relation is:

$$\mathrm{h}_{\alpha}(\mathbf{r}) = \exp\left(-\beta u_{\alpha}(r) + h_{\alpha}(r) - c_{\alpha}(r) + B_{\alpha}(r)\right) - 1 \qquad (2.34)$$

$$\alpha = 1, \ldots, N_{solvent}$$

where $u_{\alpha}(r)$ is the 3D interaction potential between the solute molecule and α site of solvent, $\beta = 1/(k_B T)$, k_B is the Boltzmann constant, *T* is the temperature and *B(r)* is the so-called "bridge" functional. At this point, if *u(r)*, *T* and *B(r)* are known or estimated, the total and direct correlation function can be found by numerical solution of the previous equations.

The 3D interaction potential between the solute molecule and the α site of the solvent, $u_{\alpha}(r)$, is calculated as a site-site interaction potential between solute site and solvent site; it depends only on the absolute distance of the two sites and it is commonly composed of a long-range electrostatic term and a short-range Lennard-Jones term:

$$u_{s\alpha}^{ele} = \frac{q_s q_\alpha}{r} \qquad (2.35)$$

$$u_{s\alpha}^{LJ} = 4\varepsilon_{s\alpha}^{LJ}\left[\left(\frac{\sigma_{s\alpha}^{LJ}}{r}\right)^{12} - \left(\frac{\sigma_{s\alpha}^{LJ}}{r}\right)^{6}\right] \qquad (2.36)$$

where q_s, q_α are the partial electrostatic charges of the corresponding solute and solvent sites, $\varepsilon_{s\alpha}^{LJ}$and $\sigma_{s\alpha}^{LJ}$ are the LJ solute-solvent interaction parameters.

The exact bridge functionals, $B_\alpha(r)$, are practically incomputable and an approximate closure relation must be used. The Hyper Netted-Chain approximation (HNC) is the most straightforward model as it sets $B_\alpha(r)$ to zero. In this work, the Kovalenko-Hirata (KH) and partial series expansion of order-3 (PSE-3) closure relations have been used to perform 3D-RISM calculations. KH is a combination of HNC and the mean spherical approximation (MSA).

$$g_{s\alpha}(r) = \begin{cases} \exp\left(\Xi_{s\alpha}(r)\right) & \Xi_{s\alpha}(r) \le C \\ \Xi_{s\alpha}(r) + \exp(C) - C & \Xi_{s\alpha}(r) > C \end{cases} \qquad (2.37)$$

where $\Xi_{s\alpha}(r) = -\beta u_{s\alpha}(r) + h_{s\alpha} - c_{s\alpha} = -\beta u_{s\alpha}(r) + \gamma_{s\alpha}(r)$. The closure with C=0 was proposed by Hirata and Kovalenko and referred as the KH closure; when C goes to infinity we obtain the HNC closure.

PSE-*n* (in this work n=3) interpolates between KH and HNC:

GSK CONFIDENTIAL INFORMATION – DO NOT COPY

$$g_{s\alpha}(r) = \begin{cases} \exp\left(\Xi_{s\alpha}(r)\right) & if\ \Xi_{s\alpha}\ (r) \leq 0 \\ \dfrac{\sum_{i=0}^{n}\left(\Xi_{s\alpha}(r)\right)^{\mathrm{i}}}{i!} & if\ \Xi_{s\alpha}\ (r) > 0 \end{cases} \tag{2.38}$$

In the case of n = 1, the KH closure is obtained, while in the limit of n →∞ HNC is obtained.

2.5.1 Placevent Algorithm

Placevent[114] is a method that uses the 3D-RISM distribution function, *g(r),* to explicitly place solvent molecules. For all the methods detailed refer to the relevant paper[114]. In summary, the Placevent algorithm starts by reading the 3D-RISM distribution function and converting it to a population function using the following equation:

$$P(r) = \rho_{\mathrm{bulk}} V_{\mathrm{grid}} g(r) \tag{2.39}$$

where ρ_{bulk} is the density of the bulk solvent, V_{grid} is the volume of the grid and *g(r)* is the density distribution function. It then identifies the global maximum of the population distribution, where there is the highest population and probability of occupation, and it inserts an explicit atom (oxygen atom). At this point the population unit is reduced by one in the vicinity of the explicitly placed atom so that a new atom will be placed in the next highest population point. This process iterates until the water population reaches a user-defined multiple of the bulk density, which as default is set to 1.5. It is important to highlight that: a) one explicit atom is placed at each iteration b) atoms placed earlier with small index are more probable than atoms with high index that more closely resemble the bulk.

GSK CONFIDENTIAL INFORMATION – DO NOT COPY

2.6 Basic concept of WaterMap

WaterMap analysis is a protocol that performs post-MD trajectory analysis based on the inhomogeneous solvation theory (IST)[115, 116]. In this method, the free energy cost of moving a water molecule from bulk solvent into a protein hydration site is calculated.

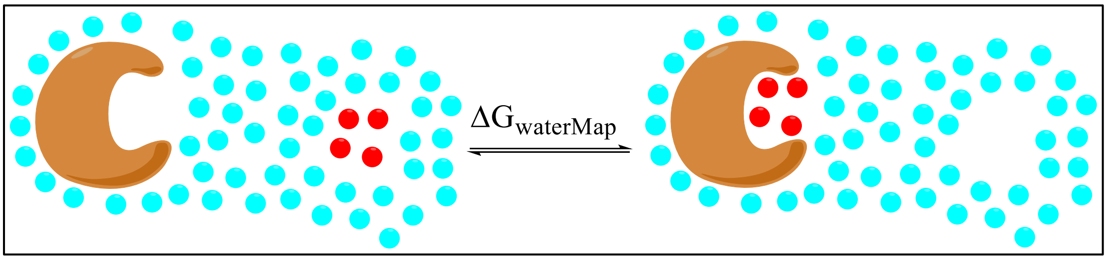


Figure 2-5Schematic representation of how the excess energy (enthalpy, entropy and free energy) in WaterMap are measured relative to bulk water.

To calculate entropy, an expansion of the orientational and spatial particle correlation functions is applied. In this model, the entropy of the bulk water is considered to be zero while local ordering of the waters due to the presence of the protein results in an unfavourable entropy. In the WaterMap protocol to estimate the excess entropy, the first order expansion and partial of the second order are calculated as follow:

$$S_e = -\frac{k_b \rho_\omega}{\Omega}\int g_{sw}(r,\omega)\ln g_{sw}(r,\omega)\,\mathrm{d}r\,\mathrm{d}\omega$$

$$-\frac{k_b \rho_w^2}{2\Omega^2}\int g_{sww}(r^2\omega^2)\ln \delta g_{Sww}(r^2,\omega^2)\,\mathrm{d}r^2\,\mathrm{d}\omega^2 \qquad (2.40)$$

where *r* is the cartesian coordinates and *ω* the Euler angle orientation of water, g_{sw} *(r,ω)* describe the single-body distribution of water at *r* and *ω*, g_{sww} *(r,ω)* gives the two body distribution, and ρ_w corresponds to the density of the bulk.

During the MD simulations, any water molecule whose oxygen atom is found to be in the binding site at given time is tagged and the positions and orientations of these water molecules are recorded. An algorithm then clusters the water positions with the aim to identify hydration-sites e.g. locations with the highest water-density. This clustering procedure is done until there are no more spatially distinct peaks with a density less than twice the bulk density of water.

The interaction energy between each water molecule and the system is calculated directly from the simulation, and the entropy of each water molecule (e.g. the entropic penalty of solvent ordering) is estimated with the inhomogeneous solvation theory (Eq.2.40) relative to the bulk water.

In general, hydration sites with positive enthalpic contribution make weaker interaction with the surrounding protein (e.g. near hydrophobic residues) than with the water molecules in solution whereas the ones with negative enthalpic contribution make stronger interactions with the surrounding proteins (e.g. charged groups) than with the water molecules in solution. The entropic contribution is always positive because any interaction between the protein and the hydration site will present some protein-water correlation entropy. Large values in entropy correspond to water molecules that have more significant entropic penalties and thus tend to be less stable in the binding site. In summary, displacing a water molecule which is thermodynamically unstable ($\Delta G > 0$) is expected to generate a gain in affinity and selectivity, while the displacement of a stable water molecule ($\Delta G < 0$) should be avoided[117].

GSK CONFIDENTIAL INFORMATION – DO NOT COPY

Chapter 3 Prediction of water molecules in BRD4-BD1

Water molecules make significant contribution in the ligand-protein binding process. Before utilising water molecules in binding studies an accurate knowledge of their location is needed. Therefore, being able to predict water location in a consistent, reliable and fast way is important in SBDD efforts. In fact, if a water molecule is neglected or misplaced the correct ligand conformation pose might be improperly scored resulting in an incorrect ligand binding pose. There are software solutions which are rigorous but time-consuming (e.g. WaterMap[40]) and prevent their usage on large numbers of protein-ligand complexes. At the same time, rapid scoring functions[118] [119] have also been used to reproduce the position of water molecules observed in crystal structures. However, an interesting alternative, that allows to accurately sample the position of water molecules in an efficient manner, is the RISM theory which calculates the distribution of the solvent around protein-ligand complexes.

At the bottom of the hydrophobic pocket of the BRDs there is a conserved network of water molecules which has been under investigation in the last years[63]. Given the importance of water molecules in the drug design, it was decided to study how this water network is predicted with 3D-RISM.

The aim of the work in is to primarily assess the predictive power of 3D-RISM followed by the Placevent algorithm in predicting the well-defined network of water molecules (W1 to W8, Figure 3-1) in the BRD4-BD1 protein. To address some of the limitation identified in the Placevent algorithm a new algorithm (GAsol) was developed by a former member of the GSK Computational Chemistry Group (Dr. Alvaro Cortes). In Section 3.3 the GAsol algorithm is evaluated and compared with the performance of Placevent.

3.1 Defining the water molecules to predict

The 184 crystal structures of the BRD4-BD1 protein (March 2017) were downloaded from the PDB to study the water molecules present in the ligand binding site. All the water molecules of each crystal structure that were less than 5 Å from the ligand were clustered according to agglomerative clustering procedure with average linkage and a modified Euclidean distance such that water belonging to the same protein could not be clustered together. This process allowed the identification of 8 clusters of highly conserved water molecules as reported in Figure 3-1 and Table 1. For each ligand-protein complex a separate file consisting of its water network was generated and used as reference for the water prediction. Correctly predicted water molecules which are represented as oxygen atoms are at a distance not greater than 2.0 Å, this cut-off distance has been selected because of the van der Waals diameter of the water (2.82 Å).

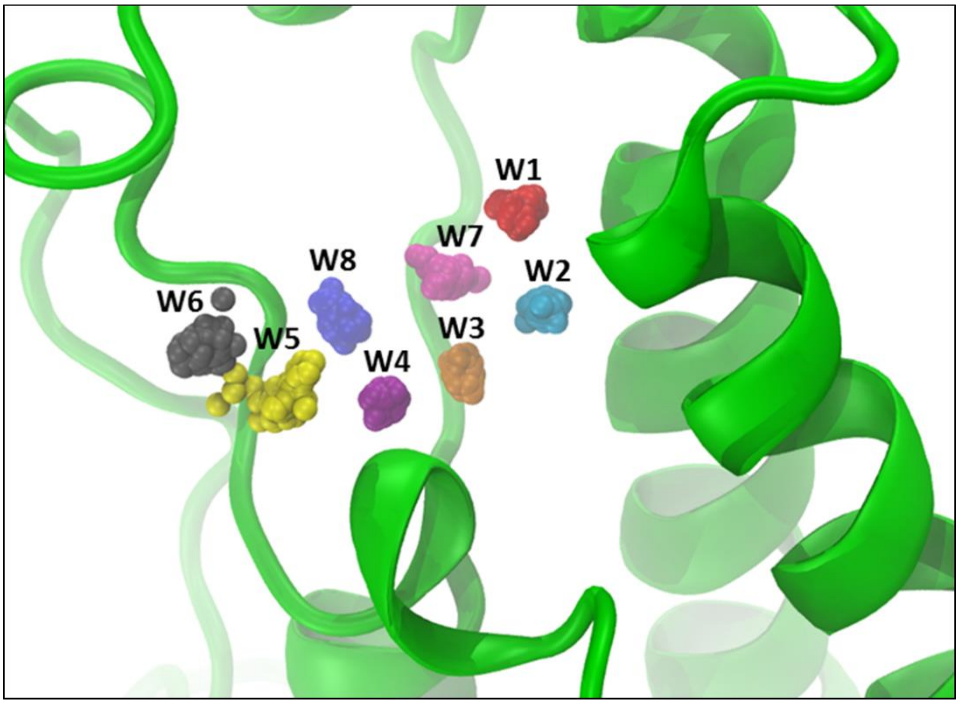


Figure 3-1 Overlay of the highly-conserved water network in the 184 Bromodomains BRD4-BD1 obtained after clustering method. The displayed backbone in green belongs to protein with PDB code 5I80.

Table 1 Presence or absence of the 8 conserved water molecules (W1 to W8) in the 184 BRD4-BD1 crystallographic structures at a distance of maximum 5 Å from the ligand.

Crystal structures (184)	Water molecule	W1	W2	W3	W4	W5	W6	W7	W8
	Present	174	174	171	173	155	153	160	168
	Not present	10	10	13	11	29	31	24	16

GSK CONFIDENTIAL INFORMATION – DO NOT COPY

3.2 Methods

3.2.1 Solvent Susceptibility Functions

The solvent susceptibility functions for 3D-RISM calculations were obtained by the 1D-RISM method present in AmberTools16[120]. The dielectrically consistent RISM (DRISM) was employed with the Kovalenko-Hirata (KH) closure. The grid size was set to 0.025 Å, which gave a total of 16384 grid points. SPC/E water at 55.5 M concentration was used as water model. Default parameters were employed for the convergence that is reached with modified direct inversion of the iterative subspace (MDIIS) method (20 MDIIS vectors, 0.3 as MDIIS step size and residual tolerance= 10^{-12}). The water dielectric constant was set to 78.497, at a temperature of 298K.

3.2.2 Input Structure and Parameters

Complexes were downloaded from the Protein Data Bank (PDB) and divided in their respective ligand-complex systems using the chain identifiers. PDB2PQR[121] was used to assign the protonation states for the protein residues. Ligands were manually protonated using the predicted interactions in the PDB as reference. Acpype[122] and Ambertools16 were used to build force field parameters for the ligands, while the Amber14SB[123] force field employed for the protein. Calculations were performed using the rism3d.snglpnt[124] program from the AmberTools16 package following the default parameters which included the usage of the KH as closure function.

3.3 Results

3.3.1 Placevent Algorithm

The 184 crystal structures belonging to the BRD4-BD1 protein were used as input to conduct 3D-RISM calculations as explained in the Section 3.2. The density distribution function was then converted into explicit oxygen atoms with Placevent algorithm (Section 2.5.1). The assessment of each prediction was done with a custom python script in which a prediction is considered correct if within 2.0 Å from the reference crystal structures. An example is reported in Figure 3-2.

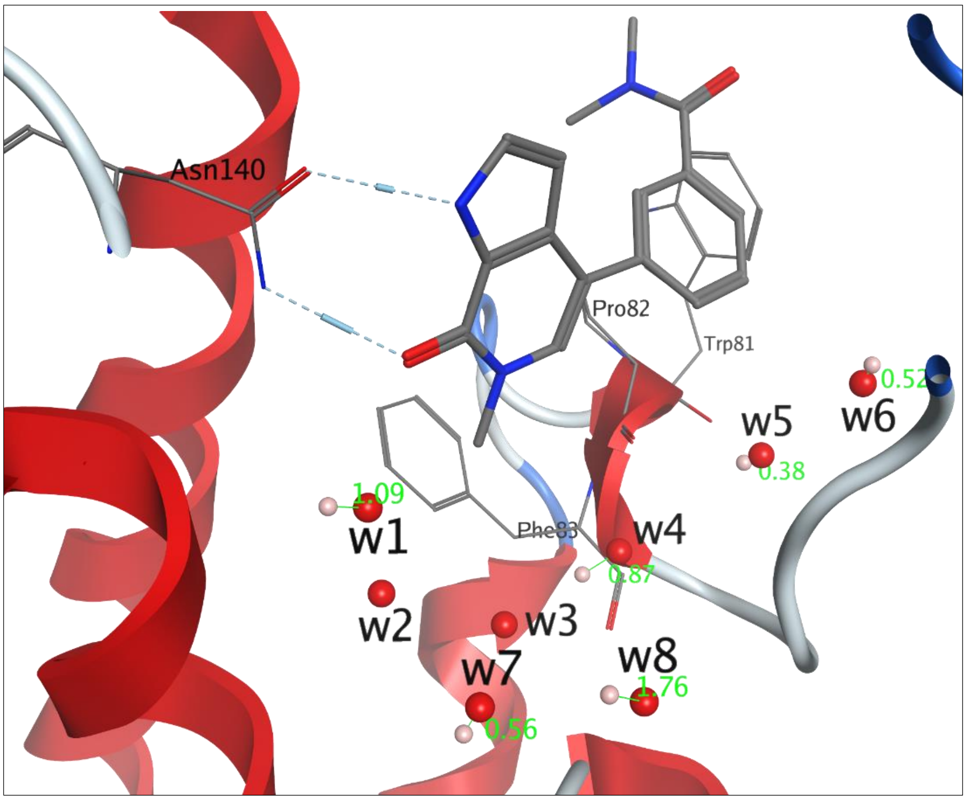


Figure 3-2 Example of water network (W1 to W8) of PDB 5I80 together with Placevent prediction in pink. The measurement between crystallographic water (in red) and prediction (in pink) is reported in Å and in green colour. The residue in the shelf (Trp81, Pro82, Phe83) and Asn140 are labelled and shown as stick.

In general, W1, W4, W5, W6, W7 and W8 are correctly placed for the majority of the cases (Table 2). On the other hand, W2 and W3 – although consistently present in most of the crystal structures, 174/184 (W2) and171/184 (W3) – are the least identified water molecules with the Placevent algorithm.

Table 2 Number of correct and incorrect predictions for the 8 conserved water molecules in the binding site of the 184 BRD4-BD1 crystal structures using the Placevent algorithm and a threshold of 2.0 Å.

		W1	**W2**	**W3**	**W4**	**W5**	**W6**	**W7**	**W8**
Placevent	# correct prediction	162	30	20	144	120	159	152	147
	# not predicted	22	154	164	40	64	25	32	37

A possible explanation for Placevent to be systematically missing W2 and W3 is due to its iteratively way of adding water molecules where it finds the highest population value [$P(r) = \rho * V * g\ (r)$] (Section 2.5.1) which in turn could result in a not globally optimized solution. An explicative example of this situation is reported in Figure 3-3: the density distribution from 3D-RIMS (orange mesh) which could be attributed to both W1 and W2 crystallographic water molecules, is converted into only one explicit oxygen atom by the Placevent algorithm.

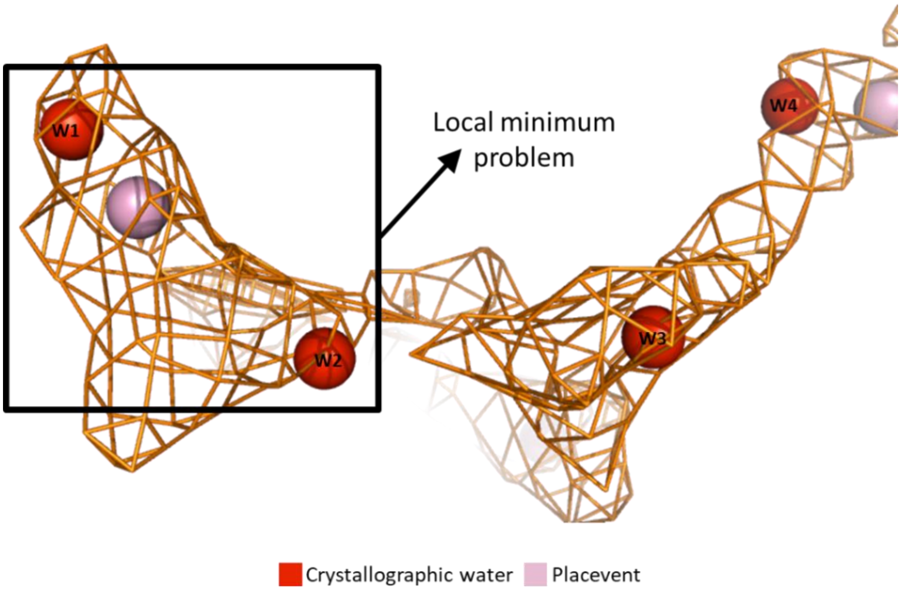


Figure 3-3 Example of local minima problem observed with the Placevent algorithm (pink) using PDB 5I80. 3D-RISM density distribution function is reported as mesh in orange. Crystallographic water molecules are reported in red.

3.3.2 GAsol

GAsol addresses the search for the optimal network of water molecules from a global point of view. It uses a genetic algorithm and a desirability function based on the 3D-RISM density that tries to avoid the local minima problem as observed with the Placevent algorithm (Figure 3-3).

3.3.2.1 Detecting the water sites and the genetic algorithm

At the start of the algorithm, it is crucial to correctly define the number of water sites to consider. This is done in the following way: the user specifies the region of interest by defining a spatial constraint in the form of a sphere with a user defined centre and radius, such that only grid points inside the defined region are chosen; then, the density distribution from 3D-RISM is converted into a population; the minimum value of *g(r)* to consider a grid point as a potential water site is by default set to a value of 5. After selecting the potential water sites, the algorithm initializes a population of individuals of potential solutions to the problem (chromosomes). Each chromosome is made of multiple genes, as many as water sites are available. Each gene is set to a value of 1, meaning the site is occupied by a water molecule or to 0, meaning that the site is empty (Figure 3-4). The initial population is then evolved during a maximum of 10,000 generations. In each generation, the population is subjected to selection, crossover and mutation. The selection procedure chooses individuals in the current generation with a tournament scheme. In this tournament, three individuals are selected randomly, allowing repetition, and only the best solution is allowed to reproduce. In the crossover phase, these individuals are mated by combining their chromosomes defining two random crossover points. Finally, in the mutation step, random gene flips are introduced in the offspring with a low probability.

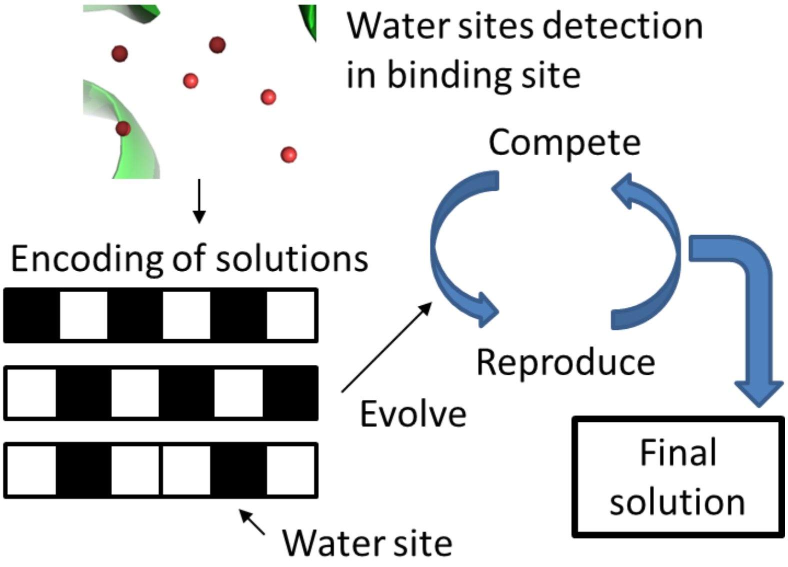


Figure 3-4 General overview of the GAsol algorithm.

3.3.2.2 Desirability function

Before the algorithm starts to generate solutions, the density distribution from the 3D-RISM calculation is transformed to a population function by using the equation $P(\boldsymbol{r}) = \rho_{bulk}V_{voxel}g(\boldsymbol{r})$ where ρ_{bulk} is the density of the bulk solvent, V_{voxel} is the volume of one voxel in the grid and $g(\boldsymbol{r})$ is the density function from 3D-RISM. Next to the conversion into population, there are two scoring phases for each individual water site and solution. For each water site detected in the first phase of the program, the minimum number of voxels required to account for one unit of the population is calculated. In this way, the minimum radius for a water molecule in the current grid point is effectively determined. Each water site is then scored by dividing the final population value (which should be around 1.0) by the radius of the sphere calculated. This scoring method guarantees that water sites with more compact populations, and therefore more likely, are selected preferentially. To score individual solutions (the final chromosome containing all the water molecules), a desirability function with two subcomponents and one penalty term is used. The function is defined as a weighted product of the subfunctions (3.1), corrected by the penalty term (3.2):

$$d = \left(\prod_{i}^{n=2} d_i^{w_i}\right)^{\frac{1}{\sum_i^{n=2} w_i}} \tag{3.1}$$

$$d_{final} = d - p \tag{3.2}$$

where d_i indicates the subcomponent d_1 and d_2, p is the penalty term and w is the weight value which is generally set to 1.

The first subcomponent (*d₁*) accounts for the amount of population considered for a particular solution by summing all the individual values for each occupied water site and normalizing by the sum of the values for all water sites in the solution space (occupied or not).

$$d_1(Y_i) = \frac{\sum_i^m Y_{i\_occupied}}{\sum_i^n Y_{i\_all}} \qquad (3.3)$$

$Y_{i\_occupied}$ is the score of occupied water site *i*, Y_i is the score of water site *i*, *m* is the total number of occupied water sites in the solution and *n* is the total number of water sites available.

The second subcomponent (d_2) tries to avoid double-counting the same part of the population multiple times in the case of proximal water sites:

$$d_2(y_i) = \left\{\begin{matrix} 0 & if\ y_i > 0 \\ 1 & if y_i = 0 \end{matrix}\right\} \qquad (3.4)$$

where y_i is the number of occupied sites found within the radius of another occupied water site. The function has a value of 1 by default except when two or more occupied water sites are at a distance less than the effective radius of any of them, which sets the value to 0. A penalty term has been introduced to improve the efficiency of the algorithm regarding the second subcomponent. As the desirability of the non-feasible solutions is always 0, the algorithm tends to waste several iterations at the start since the random solutions usually contain multiple incompatible occupied water sites. The penalty term is defined then as the weighted ratio of the number of incompatible water sites (S_i) and the total number of sites in the chromosome (S_T):

$$p(S_i) = \frac{S_i}{S_T} \qquad (3.5)$$

3.3.3 GAsol: Results and discussion

The density distribution functions obtained with 3D-RISM from the 184 BRD4-BD1 protein which were converted in explicit oxygen atoms with Placevent algorithm were also evaluated with the GAsol algorithm. As for Placevent, a prediction was considered correct when a water molecule (oxygen atom) was placed at a distance not greater than 2.0 Å from the corresponding crystallographic water molecule. A different algorithm able to give a globally optimized solution was needed because in the case of the water network at the bottom of the BRD4-BD1 protein, W2 and W3 which are close to each other, were systematically missed due to local minima problem (Figure 3-3 and Figure 3-5). GAsol, the newly developed genetic algorithm by Dr. Alvaro Cortes which can be downloaded at https://github.com/accsc/GASOL, has the advantage of placing water molecules in a user defined region of the protein by finding the solution that best fit the density distribution function.

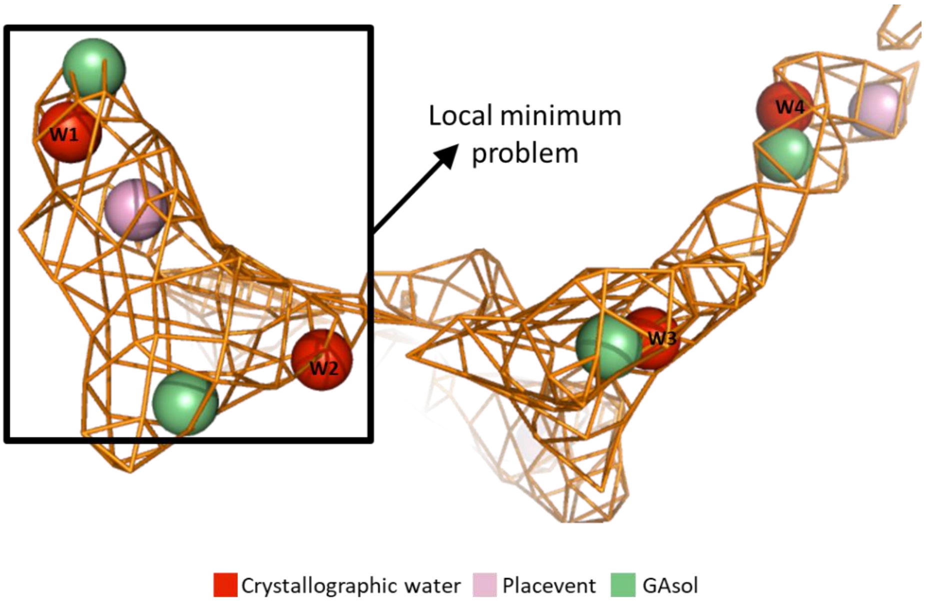


Figure 3-5 Example of conversion of 3D-RISM density distribution function (orange) into explicit water molecules from Placevent (pink) and GAsol (green). The crystallographyc water are reported in red.

GAsol predicted the correct 8-water molecules network for the majority of the crystal structures and it produced improved results as compared to Placevent in 90% of the BRD4-BD1 complexes. More specifically, the number of times in which W3 is correctly placed is 141/171 cases as opposed to Placevent in which a correct prediction is observed in only 20 cases. In Placevent algorithm, W2 is correctly predicted in 20 cases whereas in GAsol it is always successfully predicted (174/174).

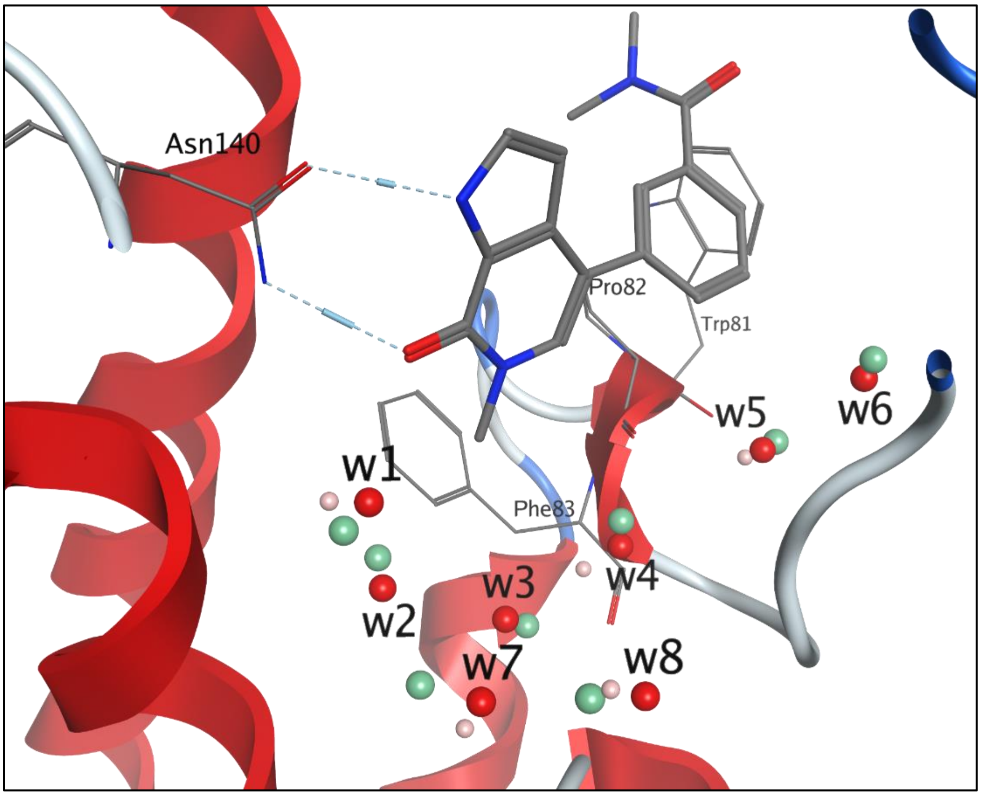


Figure 3-6 Example of water network (W1 to W8) of PDB 5I80 together with Placevent prediction in pink and Gasol prediction in green. Placevent fails to predict W2 and W3 which are instead predicted by GAsol. The residue in the shelf (Trp81, Pro82, Phe83) and Asn140 are labelled and reported as stick.

Table 3 Number of correct and not present prediction by GAsol algorithm.

GAsol	**# correct prediction**	**169**	**174**	**141**	**154**	**144**	**145**	**160**	**166**
	# not predicted	15	10	43	30	40	39	24	18

Given the metric used to successfully predict a water molecule by checking if a prediction is matching a crystallographic water at a distance not greater than 2.0 Å, it was key to test the correlation between the threshold distance from the crystallographic water molecules and the predictive power of both Placevent and GAsol algorithm. As reported in Figure 3-7, it is clear that with Placevent the percentage of the correct prediction increases constantly with increasing distance; if a less restrictive distance of 3.0 Å from the reference crystallographic water is used, the percentage of correct prediction increases up to 85%. In the case of GAsol, the rapid increase in the percentage of the correct prediction up to a distance of 2.0 Å is followed by a plateau in which the number of the correct prediction is constant. Placevent performance even at 3.0 Å is lower than the GAsol one.

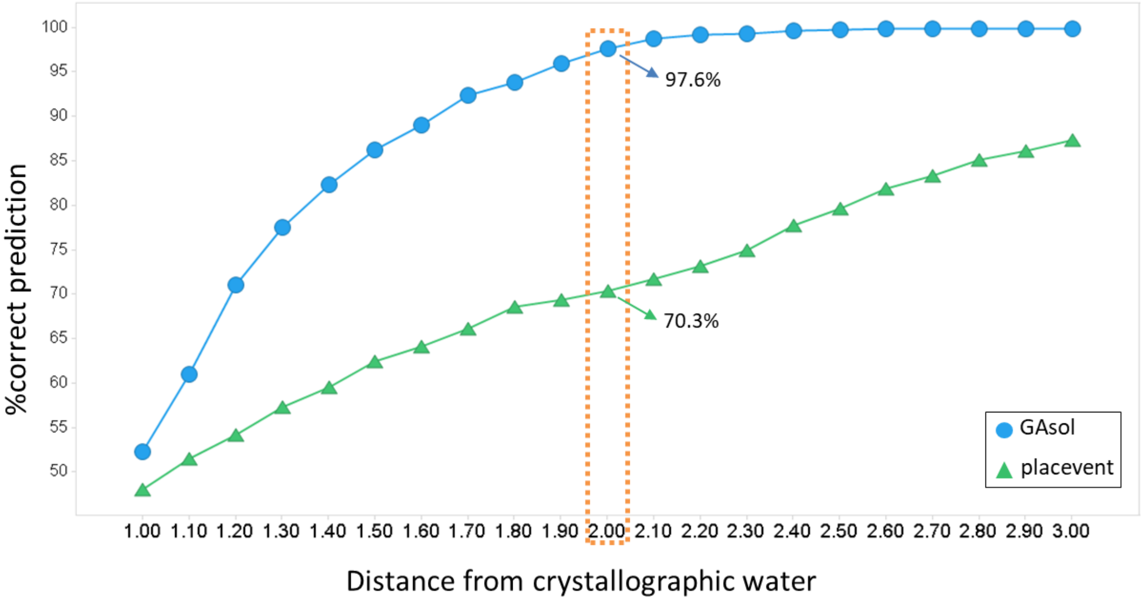


Figure 3-7 Percentage of correct prediction depending on the threshold used to match predicted and crystallographic water molecules. The dotted line indicates the distance threshold which was used during the comparison of the two algorithms.

Lastly, it was verified that the improved performance observed with the GAsol algorithm was not dependent on simple addition of more water molecules to match the crystallographic ones. To study this, it was decided to evaluate the number of additional water or false predictions that were found in the results from Placevent and GAsol. Given that the network of 8 water molecules is correctly predicted, a false prediction is defined as any additional water molecule which is found to be at a distance up to 2.0 Å from a predicted water molecule and such prediction is not associated with any other reference crystallographic molecules. From the overall results, it is evident that the accuracy of GAsol is higher than that of Placevent of about 95% and 71% respectively, see Figure 3-8. Most importantly, it is observed that the increased performance of GAsol does not come with a higher number of false positive but it is similar to the Placevent percentage (1.1% vs 0.7%).

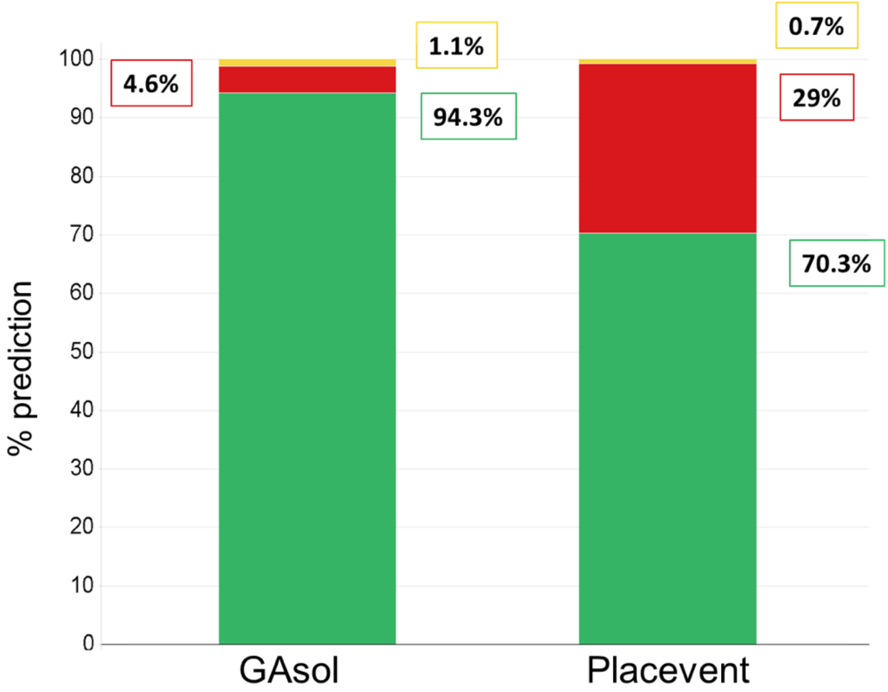


Figure 3-8 Percentage of correct (green), missing (red) and incorrect (false positive, yellow) prediction by GAsol and Placevent algorithm.

GAsol is outperforming Placevent algorithm in placing water molecules in the cavity of BRD4-BD1 protein. Nevertheless, it should be highlighted that it was not developed to replace *in toto* Placevent but rather to have a reliable alternative in situations where a precise information or globally optimized solution of the water network is needed, such as in the protein-ligand cavity.

3.4 Conclusions

The conserved water network in the BRD4-BD1 protein was initially predicted with 3D-RISM. The continuous density distribution functions were then converted into explicit water molecules by Placevent algorithm. It was noted that, despite the conserved presence of two particular water molecules, here called W2 and W3, their correct prediction was rarely obtained. To address this issue, a new algorithm called GAsol was developed by Dr. Alvaro Cortes and tested in this work with the intention to specifically solve the local minima problem observed with Placevent and provide alternatives to the conversion of the density distribution functions obtained from 3D-RISM.

By analysis of the results, GAsol is able to correctly predict the conserved water network in the 184 BRD4-BD1 protein under investigation with a 90% improvement with respect to Placevent. The increased performance of GAsol is not attributed to simply adding more false predictions. In fact, the number of incorrect water molecules which are not matching any crystallographic one is only 0.4% higher than with Placevent. Overall, GAsol is able to correctly predict water molecules in confined regions such as the ligand binding site and it has shown to be a valuable alternative when a globally optimized solution is needed.

GSK CONFIDENTIAL INFORMATION – DO NOT COPY

Chapter 4 Exploring Ligand stability with Binding Pose Metadynamics

4.1 Overview

A crucial aspect in SBDD is obtaining the crystal structure of the biological macromolecule with the ligand of interest. It is equally important if not even more critical, to choose the correct geometry of the protein-ligand complex that will then constitute the starting point for computational experiments such as virtual screening for hit identification, molecular docking for pose identification or more rigorous binding free energy methods for predicting the ligand binding affinity. Therefore, it is clear that the quality of the protein-ligand complexes ultimately determines the success of the applied computational methods.

The ligand reported as bound to the protein needs to be supported by the primary evidence of the X-ray diffraction experiment which is the Electron Density (ED) and also by the primary expectations which are the known distributions of stereochemical descriptors such as bond lengths, bond angles, and general stereochemistry. As reported in several papers[125-129], there are still cases deposited in the PDB in which the presence and/or location of the ligand is not fully supported by the underlying ED. The ED of a protein-ligand complex can be downloaded directly from the PDBe (Protein Data Bank in Europe: http://pdbe.org/); in the absence of a precomputed ED maps, structure factors can be downloaded from the PDB and then converted by ED maps using different software.

The primary goal of the following work is evaluating the BPMD protocol to assess the quality of the ligand binding pose in the starting crystal structure and inform on cases in which the ligand binding mode needs to be revised before undertaking SBDD campaigns.

4.2 Methods

4.2.2 Datasets

The first three test cases were identified by searching the primary literature: Epothilone[130-133] bound to tubulin alpha-1β chain protein, Ampicillin[134] bound to penicillin-binding protein and a ligand bound to IκB kinase β[135]. Extra cases were identified from the RCSB PDB in a semi-automated fashion. The Real Space

Correlation Coefficient[136, 137] (RSCC) is a local measure of how well the calculated density of a ligand matches the observed electron density and is defined as,

$$RSCC = \frac{cov(\rho_{obs}, \rho_{calc})}{\left[var(\rho_{obs})var(\ \rho_{calc)})\right]^{\frac{1}{2}}}$$

where ρ's are the electron density values at grid points that cover the residue in question, *obs* and *calc* refer to experimental and model electron density and *cov* and *var* are the sample covariance and variance. RSCC is a statistical measurement which is publicly available for deposited PDB structures through the Protein Data Bank in Europe (PDBe; http://pdbe.org/). In the *world-wide Protein Data Bank* (wwPDB) X-ray validation report, it is stated that a RSCC value above 0.95 normally indicates a good fit whereas a poor fit results in a RSCC value around or below 0.8. All the ligands present in the PDB with RSCC annotation were taken from the *Twilight*[138] database and were merged with their Uniprot (https://uniprot.org/) entry name as found in the primary structure section of the PDB file (DBREF section). In this way, all the proteins in which at least one structure was present with RSCC < 0.8 (ligand not fully supported by the underlying electron density) and one with RSCC > 0.9 (ligand with good electron density fit) were selected; the set of two structures with different ligands but same Uniprot entry will be referred to as a "pair". Furthermore, crystal structures with resolution worse than 2.5 Å were discarded as well as all the proteins in which the biological assembly was not monomeric. A total of about 11000 structures were retrieved, ligands that are part of the crystallization buffer or solvent were removed (7538 total structures). The number of structures was further reduced by grouping them with their Uniprot entry name and RSCC value. A representative member of each protein was visually inspected; all the structures of that protein were discarded if the protein was challenging to model e.g. it contained unresolved portions or was difficult to parameterise. This procedure provided a set of 63 structures (61 unique ligands) as a reference set to validate the BPMD tool. Since it was difficult to find sufficient pairs for the dataset, the 2.5 Å resolution criterion was relaxed to allow five extra structures (PDB: 2ITY, 3W16, 3QCQ, 4QE8 and 5HIB) to be selected.

The ($2m\mathrm{F}_o$-$D\mathrm{F}_c$) maps and ($m\mathrm{F}_o$-$D\mathrm{F}_c$) maps for the reference dataset were downloaded from the PDBe database (http://www.ebi.ac.uk/pdbe/) contoured at +1 σ and ±3 σ respectively and visually inspected with Coot[139]. *Fo* and *Fc* are the experimentally measured and model-based amplitudes, respectively, *m* is the figure of merit, *D* is the σ_A weighting factor[140]. In general, protein-ligand models with RSCC < 0.8 have poor ligand ED-fit. This can range from complete absence of ED for significant portions of the ligand to broken ED throughout the entirety of the ligand[128]. A low RSCC score can also be the result of a combination of good ED-fit for the portions of the ligand interacting with the protein target and the remaining portion(s) having very poor ED-fit due to high ligand moiety mobility. It is clear that the interpretation of the final ED-fit is a subjective process and there is likely to be a wider range of individual fitting interpretations when the ED is poorly defined.

In this work, we want to address the capability of BPMD to discriminate between high and low probability ligand binding modes, therefore it is important to separate cases in which the ED is almost absent, indicating that the ligand presence is not supported by experimental evidence, from cases where the ED quality does not allow a complete determination of the ligand pose due to partial disorder, indicating that the ligand presence is partially supported by ED. In order to better define the dataset, omit maps were calculated with σ_A style maximum likelihood-weighted $m\mathrm{F_o}$-$D\mathrm{F_c}$ and $2m\mathrm{F_o}$-$D\mathrm{F_c}$ map coefficients for all the structures by removing the ligand in question followed by maximum likelihood refinement in *REFMAC*[141]. The omit map here refers to the fact that the ligand is omitted from the model refinement to reduce model bias in the electron density map. If the ligand molecule is present, the shape of the resulting difference electron density will provide corresponding evidence. After the omit map was created and visually inspected with the original coordinates overlaid, the dataset was more accurately divided in three distinct categories (Table 4): 1) Green: ligands with RSCC > 0.9 that show very good fit with the ED; 2) Amber: ligands with RSCC < 0.8 that are only partially supported by the ED, i.e., the ligand presence showing partially occupied and/or disordered portions and a fractional positive difference density is observed in the ($m\mathrm{F_o}$-$D\mathrm{F_c}$) omit maps; 3) Red: ligands with RSCC < 0.8 that have no moieties supported by the ED, i.e., in the ($2m\mathrm{F_o}$-$D\mathrm{F_c}$) map there is no ED that could explain the ligand, no interpretable positive difference ED in the ($m\mathrm{F_o}$-$D\mathrm{F_c}$) omit maps and/or presence of negative difference density.

Table 4. Overall classification of the dataset used for validating BPMD tool. The number in parenthesis represents the overall number of ligands identified from both manual literature search and Twilight database filtering.

Category	Number of structures
Green: Ligand supported by ED	29 (30)
Amber: Ligand partially supported by ED	18
Red: Ligand with ambiguous density	16 (21)

4.2.3 System Preparation

Complexes were downloaded from the PDB and prepared with the Protein Preparation Wizard[142] in Maestro[143] v.2018.04. Hydrogens atoms and missing residues were added to the initial coordinates; bond orders were assigned to the ligand in the crystal structures. The protein termini were capped with ACE and NMA residues. Epik was used to find the most likely protonation state of the ligand and the energy penalties associated with alternate protonation states. The protein's hydrogen bond network was optimized using the ProtAssign algorithm in the H-Bond Refine Tab of the Protein Preparation Wizard[142] (Maestro v.2018.04) by correcting both potentially transposed heavy atoms in asparagine, glutamine and histidine side chains and also the protonation states of histidine residues. Finally, a restrained minimization using the Impref module of Impact with the OPLS3e[144] force field was used such that hydrogen atoms were freely minimized while allowing for sufficient heavy-atom movement to relax strained bonds, angles and clashes. The Force Field Builder (FFBuilder) tool was used to automatically generate accurate force field torsional parameters derived from quantum mechanics for all ligands containing substructures not fully covered by the standard OPLS3e parameters.

4.2.4 Binding Pose Metadynamics (BPMD)

BPMD as implemented in Maestro[143] v.2018.4 is a variation of MetaD simulation in which 10 independent MetaD simulations of 10 ns are performed using as CV the measure of the RMSD of the ligand heavy atoms relative to their starting position. The hill height and width were set to 0.05 kcal/mol (about 1/10 of the characteristic thermal energy of the system, k_BT) and 0.02 Å respectively. Before the actual MetaD run, the system was solvated in a box of SPC/E water molecules followed by several minimization and restrained MD steps that allow the system to slowly reach the desired temperature of 300 K as well as releasing any bad contacts and/or strain in the initial starting structure. The final snapshot of the short unbiased MD simulation of 0.5 ns is then used as reference for the following MetaD production phase.

The key concept in BPMD is that under the same biasing force, ligands that are not stably bound to the receptor will experience a higher fluctuation in their RMSD as compared to stably bound ones. Three scores are provided by BPMD that are related to the stability of the ligand during the course of the MetaD simulations: 1) *PoseScore* indicative of the average RMSD from the starting pose. Rapid increase in the *PoseScore* is indicative of ligands that are not in a well-defined energy minimum and hence might not have been accurately modelled. 2) PersistenceScore (*PersScore*) is a measure of the hydrogen bond persistence calculated as the fraction of the frames in the last 2ns of the simulation that have the same hydrogen bonds as the input structure, averaged over all the 10 repeat simulations. Low *PersScore* is found in structures in which their contact network is weakened by the BPMD bias. It ranges between 0, indicating that either the ligand at the start did not have any interaction with the protein or that the interactions were lost in due course and 1, indicating that the interactions between the starting ligand binding mode and the last 2 ns of the simulations are fully kept. 3) CompositeScore (*CompScore*) is a linear combination of the *PoseScore* and *PersScore* obtained from fitting the results on 42 different systems from the primary paper[145] and is calculated as follows: $CompScore = PoseScore - 5 * PersScore$.

4.2.4.1 Performance evaluation

For the evaluation of the *PoseScore* and *PersScore* performance, several statistical metrics were generated such as sensitivity, specificity, and k-score. Those metrics were generated by using the number of ligands in the green category that were predicted to have a *PoseScore* < 2 as True Positives (TP) or PoseScore > 2 as False Negative (FN) and the number of ligands in the red category predicted with *PoseScore* < 2 as False Positive (FP) and *PoseScore* > 2 as True Negative (TN).

The sensitivity was calculated as the proportion of positives that are correctly predicted:

$$Sensitivity = \frac{TP}{TP + FN} \tag{4.1}$$

The specificity was calculated as the proportion of negatives that are correctly predicted:

$$Specificity = \frac{TN}{TN + FP} \tag{4.2}$$

The accuracy was calculated as the probability to correctly classify the compounds:

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \qquad (4.3)$$

The Cohen's kappa, symbolised by κ, is a statistic useful for intrarater reliability testing. It can range from -1 to +1, where 0 represents the amount of agreement that can be expected from random chance, and 1 represents perfect agreement between the raters.

It takes the following form:

$$k = \frac{P_o - P_c}{1 - P_c} \qquad (4.4)$$

Where P_o is the proportion of observed agreement (accuracy) and P_c is the proportion of agreements expected by chance. An example of the kappa statistic may be found in Table 5.

Table 5 Data for κ calculation.

		Rater1		Total
		Yes	No	
Rater2	Yes	**A**	**B**	*g1*= **A+B**
	No	**C**	**D**	*g2*= **C+D**
	Total	*f1*= **A+C**	*f2*= **B+D**	*n* = **A+B+C+D**

Summing the observations in the main diagonal of Table 5 (**A** and **D**) and dividing it by *n*, gives the proportion of observed agreement (P_o). The proportion of expected agreement is calculated as follow:

$$P_c = \frac{\left(\frac{f_1 g_1}{n}\right) + \left(\frac{f_2 g_2}{n}\right)}{n} \tag{4.5}$$

4.2.5 Bootstrapping

The prepared complexes with the PDB code 4XYA and 4ZW1 were submitted to 100 MetaD simulations as explained in Section 4.2.4 to generate the total bootstrap population. From the total of 100 simulations, bootstrap samples of 5, 10, 20 and 30 simulations were selected. For each combination, the *PoseScore*, *PersScore* and *CompScore* were calculated and this process was repeated 100 times. Finally, statistics relative to the 100 scores from the bootstrap samples were calculated and then analysed by looking at their distribution in box plots.

4.3 Results and Discussion

Each complex was run using Desmond[146] on a single node with 4 GPU cards, taking 2-3h for a typical system (1 complex = 1 x 10 metadynamics run). The results were assessed for pose stability based on the *PoseScore*, that is the RMSD of the ligand with respect to the initial ligand heavy atoms coordinates. A *PoseScore* ≤ 2Å was considered stable (this value of RMSD is often used as a threshold defining success in prospective docking simulations). In addition, the results were also analysed looking at the *PersScore* which is an indication of the strength of the hydrogen bonds formed between the ligand and the protein residues. If 60 % of the total hydrogen bonds were kept during the simulations (e.g. *PersScore* ≥ 0.6), it was considered as a sign of good persistence. Finally, the linear combination of the two scores, *CompScore*, was also investigated but not used as the primary metric to assess pose stability as reported in the Results from Twilight Database (4.4.1).

4.3.1 Initial test cases from literature searching

The evaluation of BPMD started from three well characterised literature cases which are representative of problematic situations: 1) ligand with incorrect ligand binding mode (Epothilone A in Tubulin alpha-1β chain); 2) ligand placed into ED belonging to another entity (Ampicillin in Penicillin-binding protein 4) 3) ligand modelled with incorrect geometry (inhibitor in IκB kinase β). Each case will be discussed in detail in the following sections.

4.3.1.1 Tubulin alpha-1β chain

Epothilones are natural compounds belonging to the microtubule stabilizing antimycotic agent class that bind to the common binding site in β-tubulin. The first atomic model of Epothilone A (EpoA, Figure 4-1, B) bound to α,β-tubulin was solved by Nettles et al.[132] with a combined approach of electron crystallography, NMR and molecular modelling in 2004 (PDB: 1TVK). Serious doubts have been raised for the proposed bound conformation of EpoA in this model due to inconsistencies with NMR information. In fact, in 2013, Prota et al.[133] deposited a 2.3 Å X-ray crystal structure of EpoA in complex with α,β-tubulin, the stathmin-like protein RB3, and tubulin tyrosine ligase (PDB: 4I50) which showed a different binding mode for the ligand as compared to the 1TVK model (Figure 4-1A). For these reasons, the structure of

Tubulin alpha-1β chain in complex with EpoA was employed to examine the ability of BPMD to differentiate between correct and incorrect ligand binding mode. During the metadynamics calculation on the initial structure (1TVK), EpoA shows an average RMSD over the 10 runs that increased from the beginning to the end of the simulation time and the hydrogen bonds were present for a limited time of the simulation run (Figure 4-1, C). The overall *PoseScore* and *PersScore* are 4.0 and 0.1. Both these observations are indicative of instability of the binding mode and are consistent with reports questioning the original structure. On the other hand, EpoA in 4I50 shows a different behaviour, where the averaged RMSD reaches a steady *PoseScore* of 0.9 that is kept constant until the end of the simulations (Figure 4-1, C). At the same time, the *PersScore* indicates that the hydrogen bonds identified at the start of the metadynamics run are kept for 70% of the averaged time. In this case the results suggest a stable binding mode. There is a clear difference between the original structure (1TVK) and the more recent one (4I50) both in the *PoseScore* and the overall RMSD profile during the metadynamics run, showing that in this case BPMD can differentiate between a stable and unstable binding geometry. Another potential sign of instability for the ligand in 1TVK could also be seen in the high RMSD of up to 3.6 Å between the ten structures used as reference for the metadynamics run; in 4I50 the ligand RMS deviation reaches 0.75 Å at maximum. Finally, in the case of 4I50, EpoA under the BPMD bias experiences no drastic rearrangement as compared to the initial pose, hence the ligand conformation in 4I50 is in a stable conformation as opposed to the ligand modelled in 1TVK.

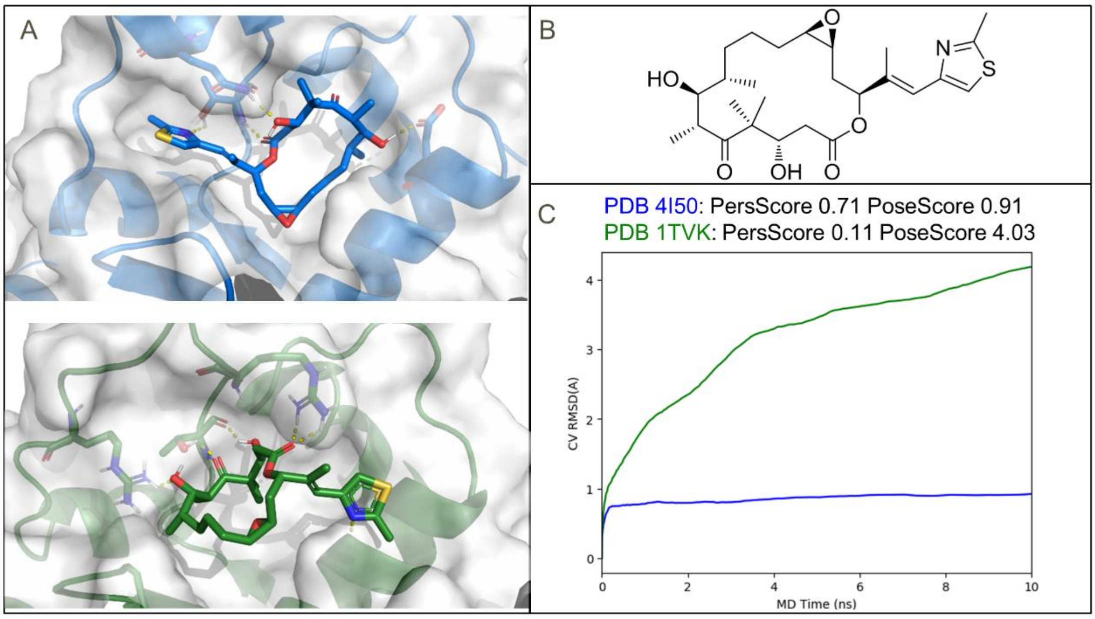


Figure 4-1 Summary of EpoA results. A) Binding site and surface of Tubulin alpha-1β chain complexed with EpoA. (PDB 4I50 blue colour, upper left corner and PDB 1TVK green colour, lower left corner). The protein structures were aligned using the backbone; the ligand RMSD between the two structures is 9.35 Å. Hydrogen bonds between EpoA and protein are shown with yellow dashed lines. B) Structure of EpoA. C) Average RMSD of EpoA during the 10 x 10 metadynamics run in PDB 4I50 (blue line) and 1TVK (green line).

4.3.1.2 Penicillin-binding protein 4

Penicillin-binding proteins are membrane-associated proteins that catalyse the final step of murein biosynthesis in bacteria. Penicillins bind irreversibly to the active site of those enzymes disrupting the cell wall synthesis. The crystal structure of the penicillin binding protein 4 from staphylococcus aureus in complex with ampicillin (PDB 3HUN[134], resolution 2Å) was deposited in 2010 showing two separate chains in the asymmetric unit although the preferred biological assembly of the complex is monomeric. In both the deposited chains, A and B, the ampicillin is modelled with a different binding mode and the RSCC is low, suggesting that there is poor fit between observed and modelled ED: RSCC $_{\text{Chain A}}$ = 0.52 and RSCC $_{\text{Chain B}}$= 0.73 (Supp. Info).

Moreover, as stated by Weichenberger et al[147], the phenyl moiety of the ampicillin (ZZ7-501, chain B) is placed in a density that could better fit a sulphate ion. Therefore, each chain containing the ligand ampicillin was submitted to the BPMD protocol. The *PoseScore* of Ampicillin in chain A and B are 4.6 and 5.1 respectively (Figure 4-2). The high RMSD deviation from the reference conformation of the ligand and the weakened hydrogen bond network during the simulations are consistent with the fact that the ED does not support the ligand presence in either of the chains.

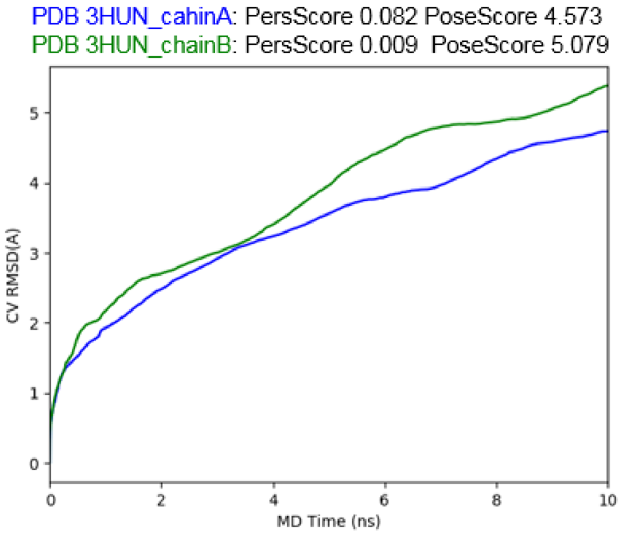


Figure 4-2 Plot of RMSD estimate averaged over all 10 trials vs simulation time for the Ampicillin from PDB 3HUN chain A in green and chain B in blue.

4.3.1.3 IκB kinase β

The structure of IκB kinase β bound to the ligand as depicted in Figure 4-3, B was solved at a resolution of 3.6 Å and deposited as PDB code 3QAD[135]. This has been the focus of several papers concerning protein-ligand crystal structures with poorly refined ligand geometries[127, 128]. In this crystal structure, the aminopyrimidine ring of the bound inhibitor had a pyramidal carbon in the pyrimidine ring instead of the expected planar one and the piperazine moiety is in an unfavourable boat conformation (Figure 4-3A). The authors released a second structure (PDB:3RZF) after re-refinement of the erroneous one. However, even in the newly released structure the ligand is highly strained[148]. Moreover, by analysing the deposited ED, as explained in the Section 4.2.2, the (2mFo-DFc) map doesn't support the presence of the ligand. The BPMD protocol was applied to the ligand as modelled in both crystal structures. The averaged *PoseScore* and *PersScore* are 6.7 and 0 for the ligand in the obsolete structure (PDB: 3QAD) and 4.9 and 0.009 for the ligand in the refined structure (PDB: 3RZF). In both cases the ligand RMSD increases significantly from the starting conformation suggesting that the ligand binding pose is highly unstable under the BPMD bias supporting the hypothesis that the ligand is not correctly modelled in the binding site in either crystal structure (Figure 4-3, C).

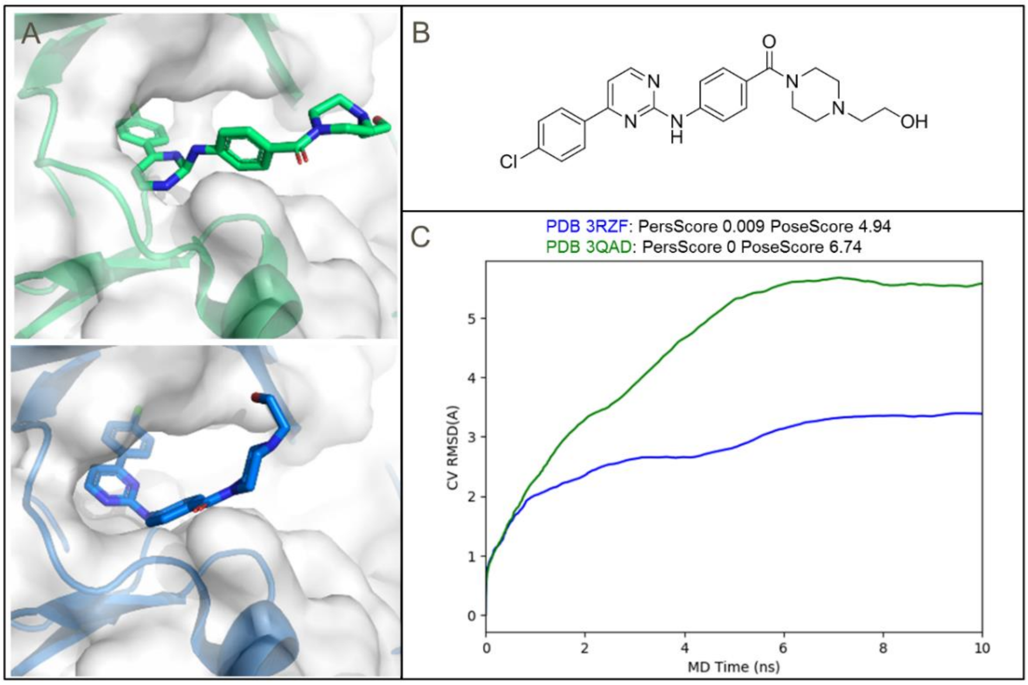


Figure 4-3 A) Binding site of Iκβ kinase in complex with inhibitor XNM. On the left: upper corner, originally deposited crystal structure (PDB 3QAD) in which the ligand geometry is highly strained with a pyramidal aromatic carbon in the amino-pyrimidine moiety and the piperazine ring in a boat conformation; lower corner, the re-deposited but still highly strained structure is shown (PDB 3RZF). B) Structure of ligand. C) Average RMSD of ligand during the 10 x 10 metadynamics run in PDB 3QAD (green line) and 3RZF (blue line).

4.4 Basic features of ligand data set and their protein target

To further validate the methodology, we identified and analysed a dataset of 64 unique ligands bound to 32 different proteins, resulting in a total of 69 complexes, including the structures from manual literature search. As reported in Figure 4-4, the majority of the structures have a crystallographic resolution below 2.5 Å, except for the cases identified from manual literature search (PDB:1TVK, 3QAD, 3RZF) and the five extra cases (see 4.2). Transferases are the most well represented structures (about 30% of the whole dataset); the remainder of the dataset is distributed across 11 protein families including hydrolase, isomerase, DNA-binding protein, signalling protein and ligase. The ligands appear to be widely distributed in druglike physicochemical space implying the generality of the data set used (Figure 4-5).

GSK CONFIDENTIAL INFORMATION – DO NOT COPY

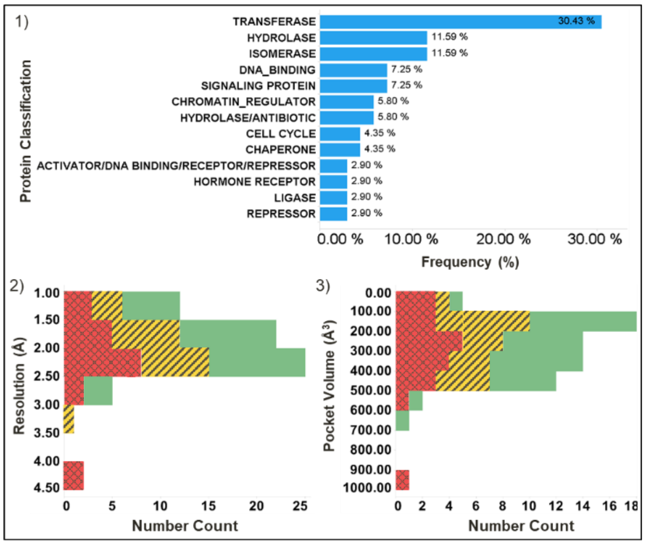


Figure 4-4 Characteristics of the protein, structures and binding pockets.1) Distribution of proteins, as displayed in the PDB and according to the Uniprot access code. 2) Distribution of crystallographic resolution values. 3) Distribution of ligand-binding pocket volumes using SiteMap. In 2 and 3, the distributions are colour coded accordingly to the category definition: ligand supported by ED (green), partially supported by ED (amber with tick diagonal stripes) and not supported by ED (red with diamond grid).

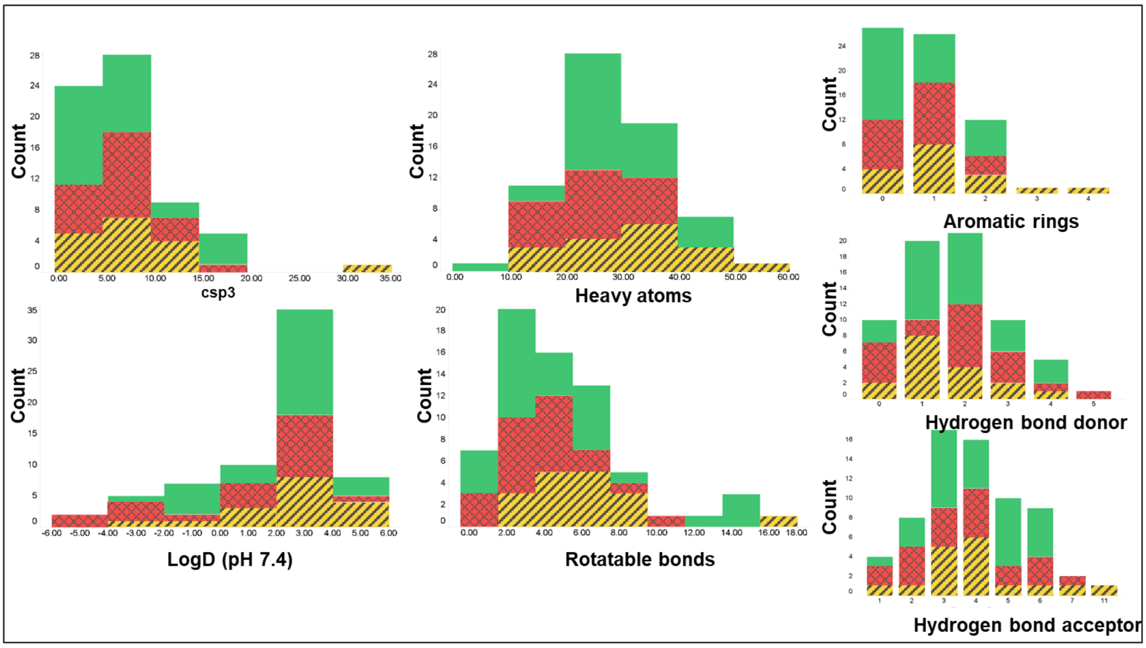


Figure 4-5 Selected physicochemical properties of the data set. From the upper left: number of sp3 carbons, heavy atom, and aromatic rings, logD at pH=7.4, number of rotatable bonds, hydrogen bond donor and acceptor. The histograms are colour coded by the ligand category definition: green, ligand supported by ED, red with diamond grid, ligand not supported by ED and amber with tick diagonal stripes, ligand with ambiguous density.

4.4.1 Results from Twilight Database

The results of the complexes from categories Green (ligand supported by ED) and Red (ligand not supported by ED) are firstly discussed. A total of 51 crystal structures, 30 with RSCC > 0.9 and 21 with RSCC < 0.8 that have been confirmed by inspection of the ED's map were subjected to the BPMD protocol to assess ligand stability. Initially, the Twilight database was searched for pairs of compounds crystallized in the same protein, in which one was well supported by the electron density and the other was not. By analysing the results for pairs of structures only, where each pair contains one structure with RSCC > 0.9 and one with RSCC < 0.8 as defined in Section 4.2, it is observed that in 7/11 pairs a cutoff of *PoseScore* = 2 clearly distinguishes between structures supported by ED and those that are not (Figure 6). In 2/11 pairs (EPHA3 and PPARG), the structures cannot be distinguished by *PoseScore*, while in 2/11 pairs the structures can be distinguished by the *PoseScore*, but both fall below (BACE1) or above (ANDR) the threshold of 2. All the test cases with RSCC > 0.9 have *PoseScore* < 2, except for the androgen receptor (Figure 6).

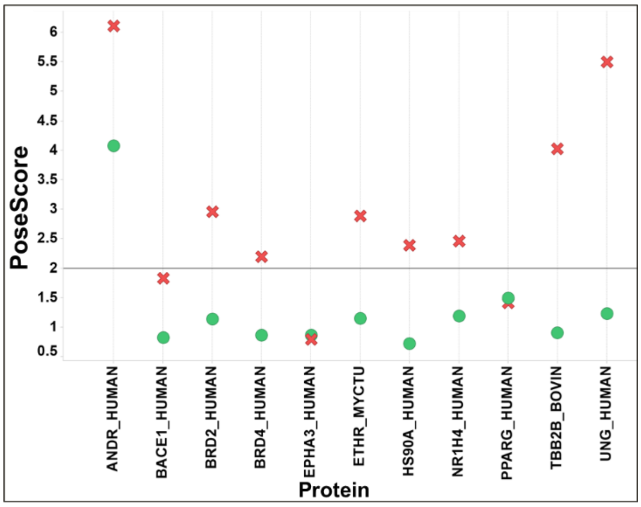


Figure 4-6 *PoseScore* overview of complexes shown by protein in pairs. Test cases are colour coded by electron density (red and cross shape= underlying density is too poor to model the ligand, green and circle shape= density is present). On the y-axis a cutoff at a *PoseScore* equal to 2 is reported to identify ligands that are stable during course of the BPMD runs.

To get a more meaningful picture of the performance of the method a larger dataset was needed. Since it proved difficult to find pairs of compounds from the same protein where one of the ligands was clearly not supported by the density, this requirement was removed, and the protocol was tested on an expanded set as shown in Figure 4-7.

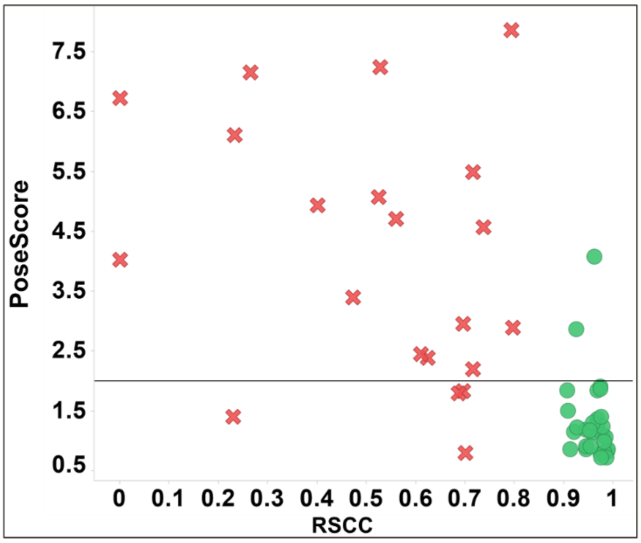


Figure 4-7 *PoseScore* of all the test cases supported by ED, in green and circle, and not supported by ED in red and cross.

Overall, if the results of category Green and Red are analysed ignoring the pair definition, 28/30 crystal structures supported by the ED have a *PoseScore* below the cutoff threshold of 2. The only two outliers are the surface exposed ligand of the androgen receptor, PDB 2PIT, and the fragment bound to the PHIP protein with 6 heavy atoms, PDB: 3MB3 (see Section 4.5). Conversely, 17/21 crystal structures where the ligand is not supported by the ED have a *PoseScore* > 2. In general, a *PoseScore* of 2 (which is indicative of the average RMSD deviation from the starting pose) has been identified as a practical threshold for distinguishing between stable and unstable ligands as proposed previously by Clark and co-workers[145]. From these results, by using the number of structures with RSCC > 0.9 and *PoseScore* < 2 as true positives (TP= 28) or *PoseScore* > 2 as false negative (FN= 2) and with RSCC < 0.8 and *PoseScore* > 2 as true negatives (TN= 17) or *PoseScore* < 2 as false positive (FP= 4), combining structures found both in the literature with manual search and in the

Twilight database (Table 4), the confusion matrix of the *PoseScore* (Table 6) gave a sensitivity of 0.94, specificity of 0.84 and a kappa value of 0.78 which confirmed the ability of BPMD to correctly separate crystal structures where the ligand has a satisfactory ED from crystal structures where the ED doesn't explain the ligand placement. The general trend previously observed for the initial test cases identified by manual search of the literature has proven to be generalizable. This separation of the two classes of ligands also gives confidence that the force field is performing well. The stability of the green ligands under BPMD bias, suggests that the instability of the red ligands is not a consequence of issues with the force field. Indeed the OPLS3e[27] force field is well validated for drug-like ligands and has been show to perform well in comparison to other ligand force fields such as those in CHARMM[149], AMBER[150] and older versions of OPLS[151].

Table 6 Confusion matrix of the structures from the Green and Red category which are supported and not by ED using *PoseScore* as metric.

	PoseScore < 2	*PoseScore* > 2	Total
Green: RSCC > 0.9	28	2	30
Red: RSCC< 0.80	4	17	21

Table 7 Summary statistics for the Green and Red category which are supported or not by ED using *PoseScore* as metric. Statistics were generated by using the green scored with *PoseScore* <2 as true positive (28) or *PoseScore* > 2 as false negative (2), the red scored with *PoseScore* < 2 as false positive (4) or *PoseScore* > 2 as true negative (17).

Sensitivity	0.94
Specificity	0.84
Precision	0.88
Accuracy	0.88
Negative Predictive Value	0.81
Kappa	0.78

The overall results were also analysed by *PersScore* (Figure 4-8). The correct threshold to separate Green and Red categories is more difficult to identify unambiguously for *PersScore* as compared to *PoseScore*. However, if a threshold of 0.6 is adopted, which corresponds to 60% of the interactions being maintained on average across all ten simulations, 23 out of 30 of the ligands supported by ED are correctly identified, while in the remaining 7 cases the interaction networks were kept between 40% and 57% of the total averaged simulation time. The fragment in PDB 3MB3, scored as unstable by the *PoseScore* metric, is the only example with RSCC > 0.9 that had no hydrogen bonds at the start of the simulation. The absolute numbers of ligands in the Red category that fell below and above the threshold of 0.6 were equal to the number of complexes identified by the *PoseScore* threshold: 17 Red protein-ligand complexes have an interaction network that is significantly altered by the BPMD bias and in 4 Red cases the network is preserved. The sensitivity, specificity and kappa value if the *PersScore* is used to evaluate ligand stability are 0.81, 0.84 and

0.58, respectively. The *PoseScore* appears to be a better metric to separate cases where the ligand is correctly modelled in the ED from those which are not. Finally, the *CompScore*, which is a combination of the *PersScore* and *PoseScore*, gives results that are comparable to *PoseScore*.

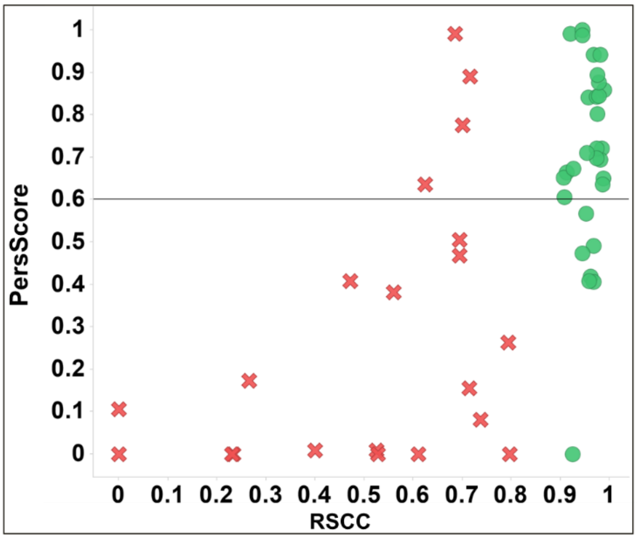
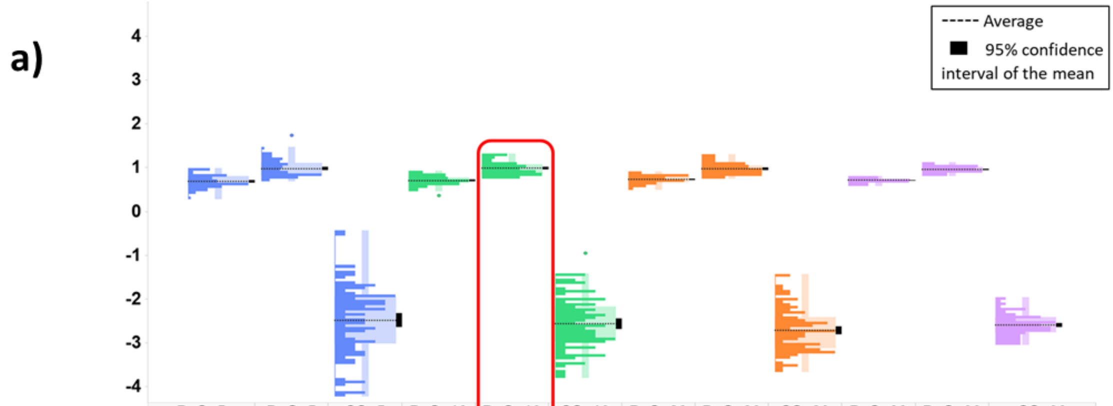


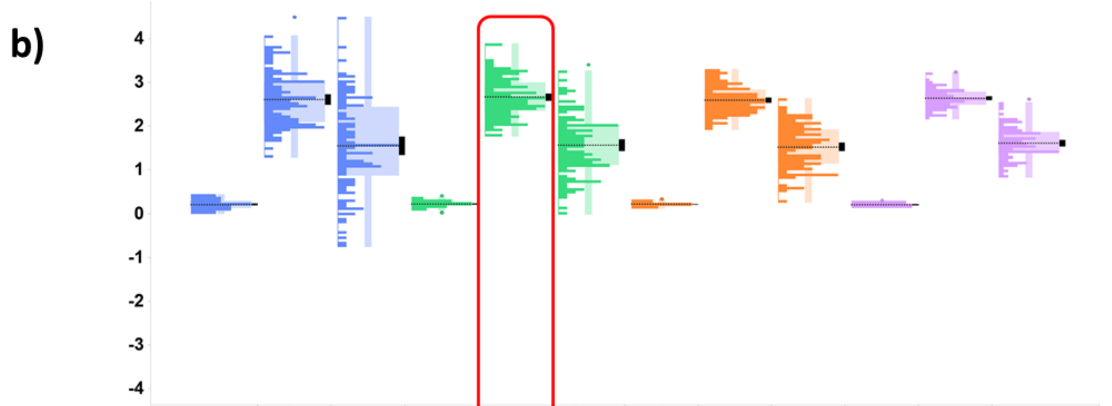
Figure 4-8 *PersScore* of all the test cases supported by ED, in green and circle, and not supported by ED in red and cross.

As a further comparison of the scores, a bootstrapping study was carried out as explained in Section 4.2.5 to estimate the standard errors associated with each metric. Two crystal structures were selected from the Green (PDB 4XYA) and Red (4ZW1). The ligand supported by the ED from Green category is in a stable conformation, the BPMD scores are: *CompScore*= -0.51, *PersScore*= 0.34 and *PoseScore*= 1.22. The ligand not supported by the ED from the Red category is unstable during the MetaD runs and the *PoseScore* is above the identified threshold of 2. The scores are *CompScore* = 1.43, *PersScore* = 0.15 and *PoseScore* = 2.2.

In general, the results of this study combined with the bootstrapping analysis (Figure 4-9) showed that *PoseScore* is the preferred metric because it can be estimated with greater precision than *CompScore* and is easier to interpret than the *PersScore*. Moreover, the addition of more replicas can decrease the variability of the scores but it does not seem to justify the increased time required for a quick and quantitative method able to distinguish between stable and not ligand poses.



	PeS_5	PoS_5	CS_5	PeS_10	PoS_10	CS_10	PeS_20	PoS_20	CS_20	PeS_30	PoS_30	CS_30
Count	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Outliers	0.00	1.00	0.00	1.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Avg	0.69	0.98	-2.48	0.71	0.99	-2.56	0.74	0.97	-2.71	0.71	0.96	-2.59
StdDev	0.15	0.19	0.79	0.11	0.15	0.58	0.09	0.13	0.49	0.05	0.07	0.25
StdErr	0.01	0.02	0.08	0.01	0.02	0.06	0.01	0.01	0.05	0.00	0.01	0.03
Var	0.02	0.04	0.62	0.01	0.02	0.34	0.01	0.02	0.24	0.00	0.01	0.06
MostCommon	0.62	0.76	-2.52	0.71	0.81	-1.63	0.82	0.78	-2.90	0.72	0.89	-2.99



	PeS_5	PoS_5	CS_5	PeS_10	PoS_10	CS_10	PeS_20	PoS_20	CS_20	PeS_30	PoS_30	CS_30
Count	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Outliers	0.00	1.00	0.00	3.00	0.00	1.00	3.00	0.00	0.00	1.00	2.00	1.00
Avg	0.21	2.61	1.55	0.22	2.66	1.56	0.21	2.60	1.52	0.20	2.64	1.62
StdDev	0.12	0.63	1.06	0.07	0.45	0.71	0.05	0.32	0.50	0.04	0.23	0.38
StdErr	0.01	0.06	0.11	0.01	0.05	0.07	0.01	0.03	0.05	0.00	0.02	0.04
Var	0.01	0.39	1.12	0.01	0.21	0.50	0.00	0.10	0.25	0.00	0.05	0.14
MostCommon	0.00	1.70	-0.77	0.21	2.46	2.13	0.18	2.27	1.22	0.19	2.50	1.39

Figure 4-9 Statistics generated from bootstrapping experiments on the crystal structure with PDB 4XYA (a) and 4ZW1 (b). The red boxes indicate the results from the current protocol with 10 replicas. The scores are abbreviated as PeS= *PersScore*, PoS= *PoseScore* and CS= *CompScore*.

4.4.1.1 Analysis of structures in the "Twilight" dataset

Inspection of the ED maps of the ligands belonging to the Amber category revealed that the regions of the ligand that were not supported by ED were often outside the binding site and solvent-exposed while the regions inside the protein binding site were mainly supported by the ED (Figure 4-10). For some of these cases, the ligands may be stable even though they have a low RSCC. Hence, a BPMD *PoseScore* below the threshold of 2 might be expected.

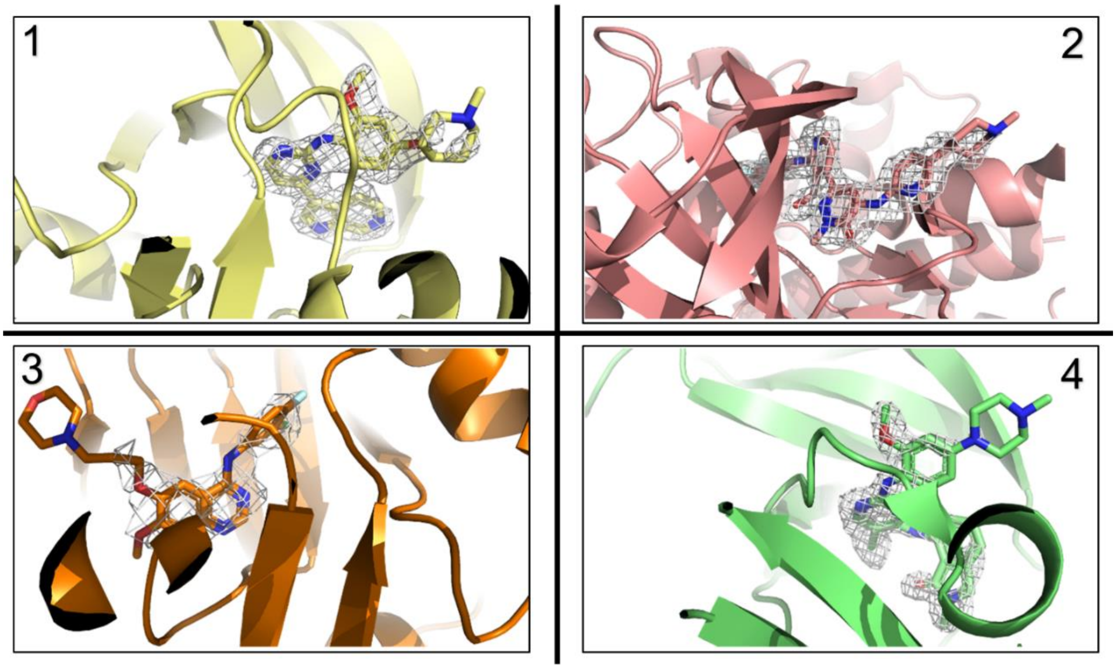


Figure 4-10 Examples of ligands belonging to Amber category in which the portion not described by the ED is solvent exposed and not interacting with the protein. The (2mFo-DFc) map contoured at 1 σ is shown as wireframe and the protein is in cartoon representaion. 1) PDB 4D2R[152], ligand DYK, RSCC= 0.68 2) PDB 5P9H[153], ligand 7G7, RSCC= 0.76 3) PDB 2ITY[154], ligand IRE, RSCC= 0.76 4) PDB 2JKO[155], ligand BJI with RSCC= 0.84.

The percentage of ligand atoms that are uncovered by the ED in the Amber category varies from about 6% to 38%, signifying that most of the ligand is supported by the density with some portion that is not supported by experimental evidence. It is observed that the ligand *PoseScore* increases with increasing disorder in the ligand structure with a r^2 of 0.54 (Figure 4-11). The ligand BJI in PDB 2JKO is about 26% uncovered by the ED and given the trend of the other structures in the Amber category, it could be thought to be unstable. Its predicted BPMD *PoseScore* is 1.1 suggesting high ligand stability. After inspecting the original publication[155], this is not surprising, because the reasoning behind its design was not to add new protein ligand interactions but to improve the ligand solubility without affecting potency by including a piperazine ring[155] that is solvent exposed (Figure 4-10, panel 4). Therefore, the high ligand stability obtained from BPMD confirms the hypothesis that the piperazine ring would not be detrimental to the stability of the ligand's binding mode, and the structural result can be trusted regardless of the missing electron density.

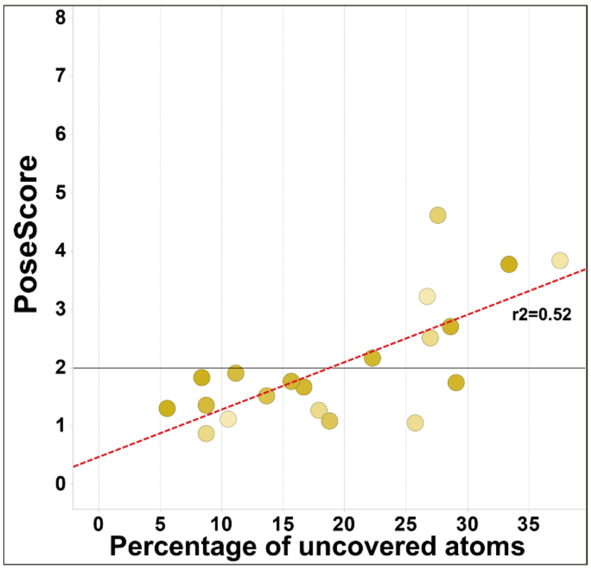


Figure 4-11 *PoseScore* of the structures belonging to Amber category (ligand partially supported by ED) with respect to the percentage of atoms that are uncovered by ED and colour coded by RSCC from low (light-yellow) to up to 0.8 (dark-yellow).

4.5 Discussion

Overall, BPMD showed good ability to identify ligands whose modelled pose is not supported by their electron density (Figure 4-6 and Figure 4-7), even for the cases in which the ligand is only partially supported by the ED (Figure 4-11). Investigating the results in more detail reveals some insights into situations that can be challenging for the method and caveats users should be aware of, which are worthy of further discussion. Before being submitted to the final metadynamics simulation, every complex underwent an equilibration procedure as explained in the Method Section (4.2). The last step of the last short unbiased MD simulation (0.5 ns) is used as the reference structure for the *PoseScore* calculation. Consequently, the reference structure for BPMD is not the input structure but the equilibrated one which can, in some cases, differ significantly. Given this observation, the *PoseScore* was compared to the RMSD between the initial crystal structure and the reference equilibrated structure, which is here called MD-RMSD, to understand if unstable poses could be flagged in advance from the MD equilibration procedure. The *PoseScore* correlates with the MD-RMSD, that is the average displacement among the reference structures with a r^2 of 0.75 (Figure 4-12): the higher the MD-RMSD from the originally deposited structure, the more likely the ligand will be unstable.

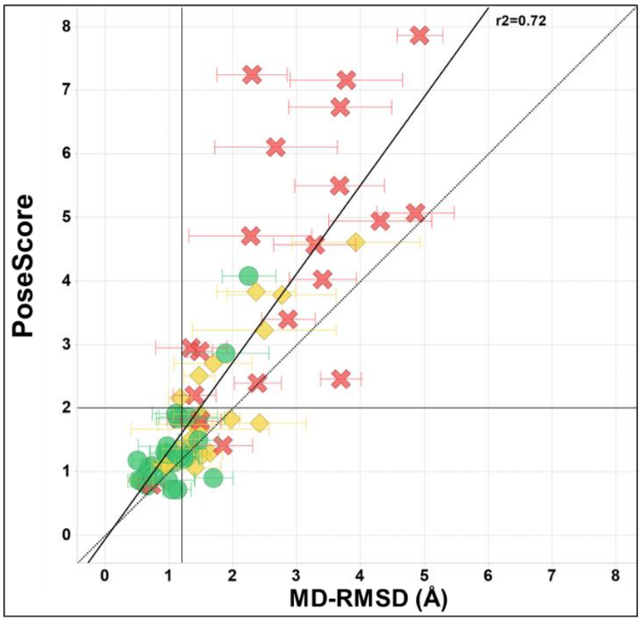


Figure 4-12 *PoseScore* correlation with the ligand RMSD of the reference structures obtained after the equilibration procedure, MD-RMSD. The structures are colour coded by category: Green (ligand supported by ED, circle shape), Amber (ligand partially supported by ED, diamond shape) and Red (ligand not supported by ED, cross shape). The error bar calculated as the standard deviation of the RMSD from each of the 10 metadynamics run is reported on top of each point. The black line is the best fit line between *PoseScore* and MD-RMSD whereas the dashed black line indicates the line y=x.

It should be noted that the RMSD between the ten reference structures can be quite significant, especially in the cases of the Red category (Figure 4-12) and in the region of MD-RMSD > 3 Å. Here, the conformation of the ligand reference structures is not a very good representation of the initial crystal structures. The high correlation between *PoseScore* and MD-RMSD (Figure 4-12) suggests that MD-RMSD could be used as a preliminary indication that the initial structure is not correctly modelled. When a shorter equilibration step of 10 ps instead of 500 ps is used for all the cases in which at least one reference structure shows a RMSD ≥ 3 Å, a systematic decrease of MD-RMSD is observed, while the *PoseScore* does not change significantly confirming that the ligand in those structures are highly unstable (Figure 4-13).

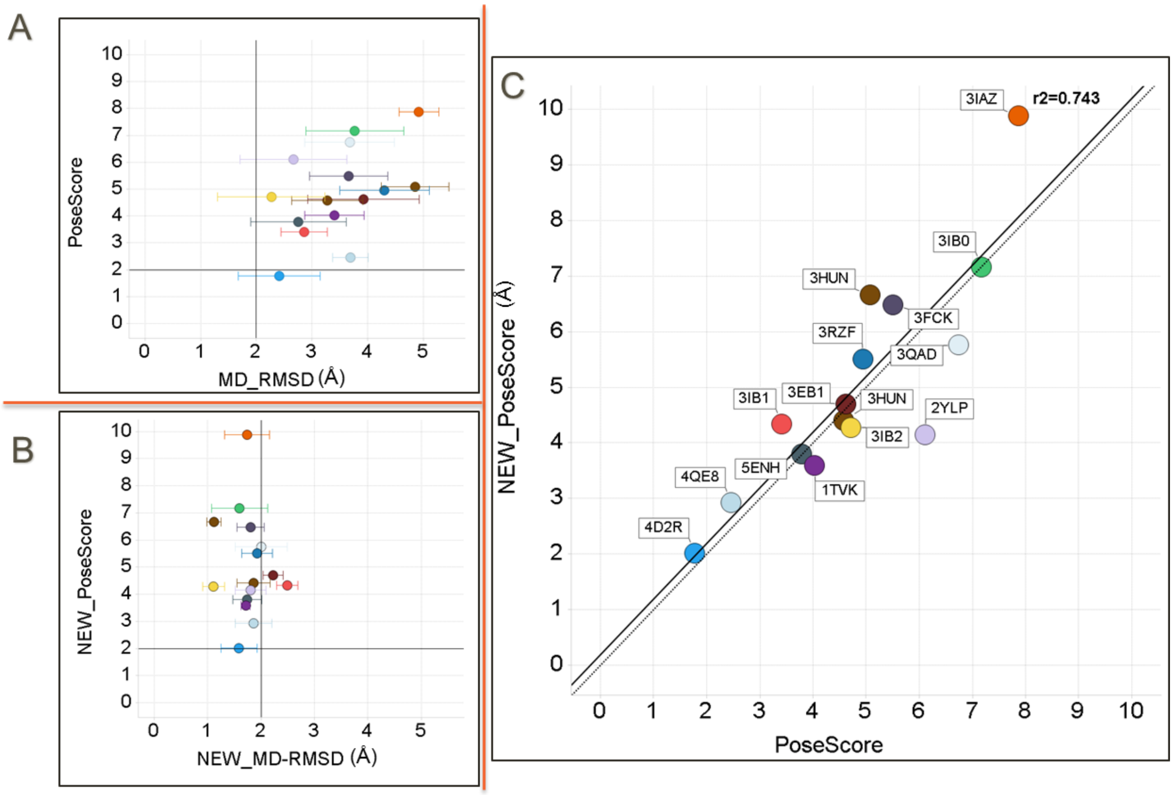


Figure 4-13. Left panel: A) *PoseScore* of the structures in which at least one of the reference structures has MD-RMSD > 3 Å. B) *PoseScore* of the same structures, run after a shorter unbiased MD simulation of 10 ps instead of 500 ps (called NEW_*PoseScore*). The MD-RMSD is below 2.5 Å for all cases. C) *PoseScore* obtained with shorter equilibration procedure (NEW_*PoseScore*) and with default protocol. Best fit line is displayed in black line and y=x in black dotted line. In general, the overall results did not change. Each structure is coloured identically in A, B and C.

In the region of MD-RMSD up to 2 Å (Figure 4-14), the average reference structure can be considered close enough to the initial starting structure as opposed to the ones in the region with MD-RMSD > 2.5 Å. The ligands in the Green category have lower structural variability (low MD-RMSD), in agreement with the *PoseScore* below the threshold of 2. The presence of outliers in the region of the MD-RMSD between 1.3 Å and 2 Å (Figure 4-14), shows the advantage of using the metadynamics production phase. In the case of 5XHK, 5MYM and 4WZ1, the benefit of BPMD is clear; in fact, despite the MD-RMSD being below 1.5 Å, which could be indicative of ligand stability, the BPMD bias was needed to correctly classify them as unstable (*PoseScore* >2) in agreement with the missing ED from the experiment. A different situation is observed for 2HWQ and 5Q17 in which *PoseScore* is found below the threshold of 2 and also the MD-RMSD is lower than 2 Å. The results obtained for this data set suggest that BPMD can successfully discriminate between crystal structures that are correctly modelled and those that are not, but also that the MD-RMSD obtained from the short equilibration step might be informative of dubious structures. A thorough investigation of MD as compared to BPMD in identifying questionable crystal structures, is required to identify the optimum protocol to balance accuracy against computational efficiency.

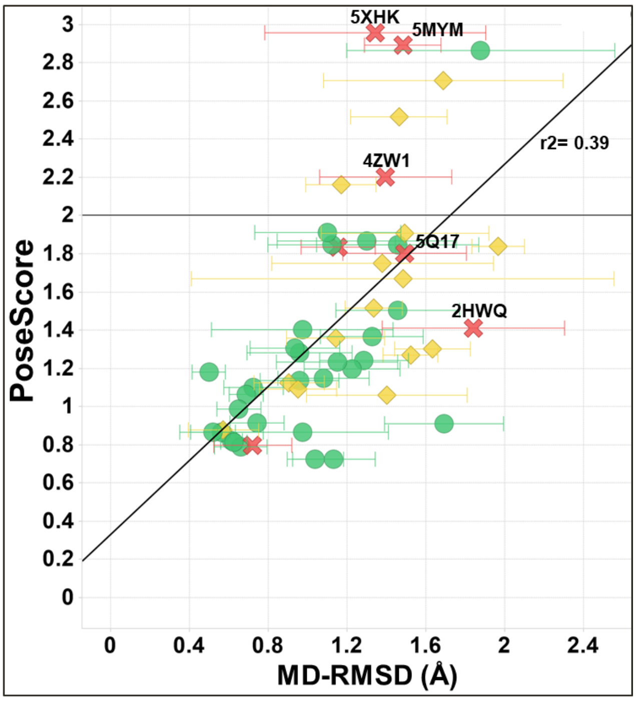


Figure 4-14 *PoseScore* in the range of MD-RMSD up to 2 Å where the reference structures can be considered similar to starting structure. The structures identified with their PDB name represent the advantages (5XHK, 5MYM, 4ZW1) and potential limitation (5Q17 and 2HWQ) of using the BPMD protocol. All the structures in the Green category show low MD-RMSD.

Several of the structures that were considered to be misclassified by BPMD, that is there was disagreement between the ED classification and the *PoseScore*, deserve further discussion. The ligand 4HY found in PDB 2PIT[156] binds to the surface-exposed allosteric pocket called Binding Function 3 of the androgen receptor. It has a RSCC= 0.93 but resulted in a high *PoseScore* of 4.1. In the original paper[156] it is claimed that it binds weakly (IC50 ~ 50 μM) to the surface weakening the contacts between the androgen receptor and coactivator proteins. In this case of low binding affinity in an open solvent exposed site, the ligand can be readily displaced under the BPMD bias, despite being supported by the electron density. A similar exposed pocket was observed in the case of three ligands PDB 5MRD, 2XPA, 2XP6 where there is agreement between ED and *PoseScore*, so this is not a consistently observed problem. Another interesting case is that of the MB3 ligand crystallized in the second bromodomain of PHIP (PDB 3MB3[157]), which has a high RSCC value of 0.93 but showed a *PoseScore* higher than 2. Interestingly, the ligand is a very small fragment possessing only 7 heavy atoms, therefore, very different from others investigated here and in previous BPMD studies. It is possible that the metadynamics parameters may not be well optimised for such a small fragment.

Among the incorrectly predicted crystal structures in the Red category, it is worth mentioning PDB 2HWQ[158]. Despite a poor ligand RSCC score of 0.23 and ambiguous fragmented density in the omit maps it has a *PoseScore* of 1.21. The binding pocket is narrow, and the ligand binds deep in it. The same authors have solved structures of related ligands bound to the same protein (PPARG) such as PDB 2F4B[159] (RSCC=

0.73) and 2HWR[158] (RSCC= 0.71). Therefore, the ligand may have been modelled by using prior knowledge from these studies in addition to the electron density. A crystal structure with comparable ligand binding mode and good RSCC value (PDB: 5GTN[160]) was also solved a few years later corroborating the deposited ligand binding mode. Therefore, in this instance, although the electron density doesn't support the binding mode, the ligand may have been correctly modelled using other information, explaining why it is miscategorised in this work.

The cases discussed above suggest that particularly small ligands or open binding sites may be challenging for BPMD. To investigate whether these factors are a key influence on the BPMD score, heavy atom count, protein-ligand H-bond count and binding pocket volume have been plotted against *PoseScore* (Figure 4-15). No correlations were seen, suggesting these factors are not driving the prediction from *PoseScore*.

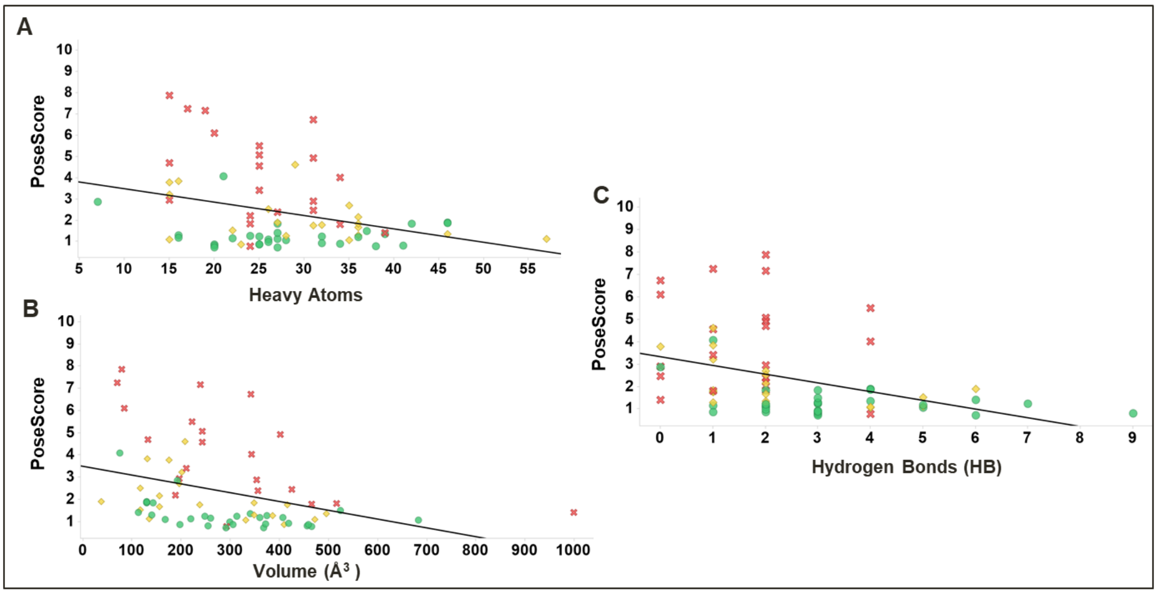


Figure 4-15 Relationship between *PoseScore* and number of heavy atoms (A), volume of binding site (B) and hydrogen bonds (C) of the ligands in the data set. Structures belonging to category Green, ligand supported by ED are coloured in green with circle shape; from category Amber, ligand partially supported by ED are in yellow with diamond shape and from category Red, ligand not supported by underlying ED, are in red and with cross shape.

Out of the 69 total structures, 34 had binding affinity data as retrieved from the primary literature citation where they were firstly discussed, in the form of IC50. The binding data were converted into pIC_{50} and a poor correlation was observed with *PoseScore* (r^2 = 0.33, see Figure 4-16) as opposed to what was found by Clark et al[145] in which they use BPMD as a tool to prioritise Induce-Fit Docking ideas. In this work, despite the binding affinity of the ligand under investigation, if the starting position is correctly modelled it will result in a *PoseScore* below 2; therefore, BPMD should not be regarded in this way as a tool to prioritize ideas based on the *PoseScore*.

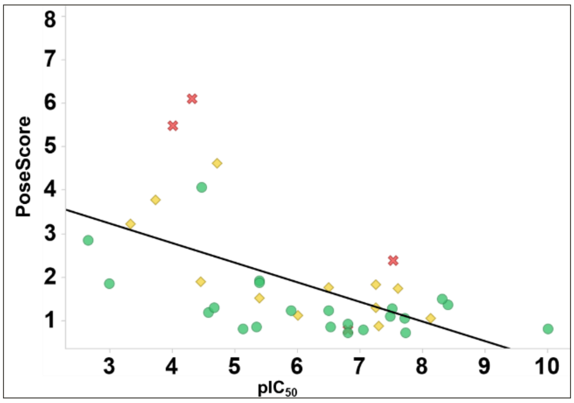


Figure 4-16 *PoseScore* of all the crystal structures with pIC50 data colour coded by category (green and circle, complexes supported by ED, amber and diamond, complexes partially supported by ED and red and cross, complexes not supported by ED).

4.6 Conclusions

The capabilities of the BPMD tool to correctly identify stable ligands were investigated in detail using a diverse data set of crystal structures from published literature and collected from the *Twilight* database. Primarily, this study has been used as a validation that BPMD is a useful predictor of binding mode stability, in agreement with previous studies looking at docking and scoring. From the conducted validation on 69 complexes with different physical chemical properties, BPMD has shown good performance categorising ligand binding modes supported by ED from those not supported. In particular, it was identified that ligands supported by ED show a *PoseScore* below 2 as opposed to the unsupported ones which have a *PoseScore* above 2. BPMD scores do not seem to correlate with binding affinity, suggesting that the method is useful for assessing the stability of individual poses or the relative stability of different poses of the same ligand, but not for scoring different ligands.

From the more challenging cases in which the ligand is partially supported by the underlying ED, it was observed that if the disorder comes from parts of the ligand that do not participate in any protein interaction, then the overall stability of the ligand, given that the rest of it is correctly modelled, won't be affected. Thus, information from BPMD simulations could help medicinal and computational chemists in designing more potent compounds maximising the ligand-protein interactions where the most stable portion of the ligand lies and incorporating flexibility and solubilising groups in the non-interacting portion without compromising the overall binding stability.

Aside from the validation of the methodology, the data presented here suggest BPMD may be a useful tool for the crystallographer. Solving crystal structures is an expert job incorporating information beyond that contained in the electron density but this makes it open to a degree of subjectivity. These results suggest that BPMD could be useful in assessing preliminary poses generated by crystallography and highlighting those that would need further investigation or confirmation before undertaking more time consuming and expensive strategies. Specific cases have been identified which can be challenging for BPMD, including very small fragments, surface exposed binding pockets and interactions with no hydrogen bonds. There are a number of parameters within BPMD that could be investigated to see whether they give an improvement in these cases, such as gaussian width and height, choice of collective variable and run time. However, they are not investigated further here because, overall, the default parameters offer good discrimination. Finally, the average RMSD of the equilibrated structures prior to the metadynamics simulations has appeared to be informative of inherently unstable ligands. Interestingly, it is observed that the reference structure used in the BPMD protocol might have substantially changed from the input structure especially for the cases in which the ligand is not supported by ED. A brief investigation of using a shorter equilibration procedure on the cases with MD-RMSD > 3 Å showed overall unchanged results with a *PoseScore* that does not change significantly from the one obtained with the default protocol. It would be interesting to know in a more systematic way which protocol of unbiased molecular dynamics is needed to obtain comparable BPMD results and this will be the subject of further studies.

GSK CONFIDENTIAL INFORMATION – DO NOT COPY

Chapter 5 Structure-based drug design of a selective probe for BRD4-BD1 domain

5.1 Introduction

In 2010, two potent and selective triazolodiazepine-based inhibitors of the BET proteins were first publicly disclosed, (+)-JQ1[161] and I-BET762[162], which were shown to have in vivo on-target activity in models of NUT midline carcinoma and inflammation respectively, (Figure 5-1). Since then, the development of BET bromodomain small molecule inhibitors as a potential strategy to treat several human diseases such as cancer, immunological diseases, heart diseases, metabolic diseases, has been the focus of both industry and academia[163, 164].

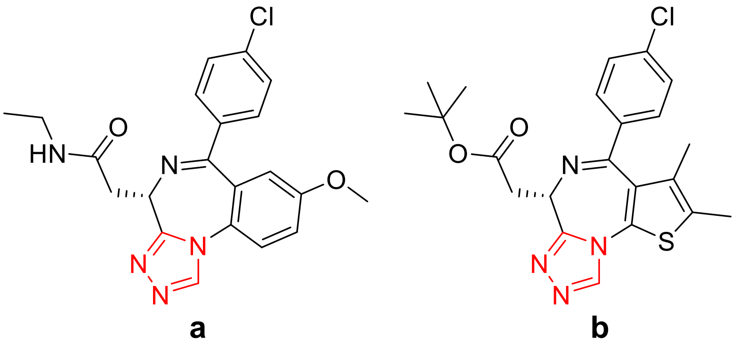


Figure 5-1 The first BET BRDs inhibitors, I-BET762 (**a**) and (+)-JQ1 (**b**).

Selective inhibition of the BET subfamily over the other BRDs members has been achieved with several compounds[165-167] that do not discriminate between the eight (4 BD1 and 4 BD2) BET family member (see Figure 1-6), from here the name of pan-BET inhibitors. This has led to a better understanding of the bromodomain mode of action in gene regulation and their role in diseases.

X-ray crystal structures for both **a** (PDB 3P5O[162]) and **b** (PDB 3MXF[161]) in BRD4-BD1 are available (Figure 5-2). The ligands show similar interactions given the high structural similarities (same triazolo-diazepine core). The key interaction with the evolutionarily conserved asparagine, Asn140 in BD1 (Asn429 in BD2), is established by the nitrogen of the triazole ring (marked in red in Figure 5-1). The adjacent nitrogen makes the water-bridged interaction with the tyrosine residue while the methyl substituent occupies a hydrophobic pocket. The region between the helices (ZA-channel) is occupied by the aromatic rings of each compound. Finally, the chlorophenyl substituents are interacting with the lipophilic shelf formed by the tryptophan (Trp), proline (Pro) and phenylalanine (Phe) residues, therefore also known as the WPF lipophilic shelf. A known strategy to increase binding affinity for BET and selectivity over non-BET bromodomains is designing substituents that occupy and interact with the WPF shelf[168] (discussed later).

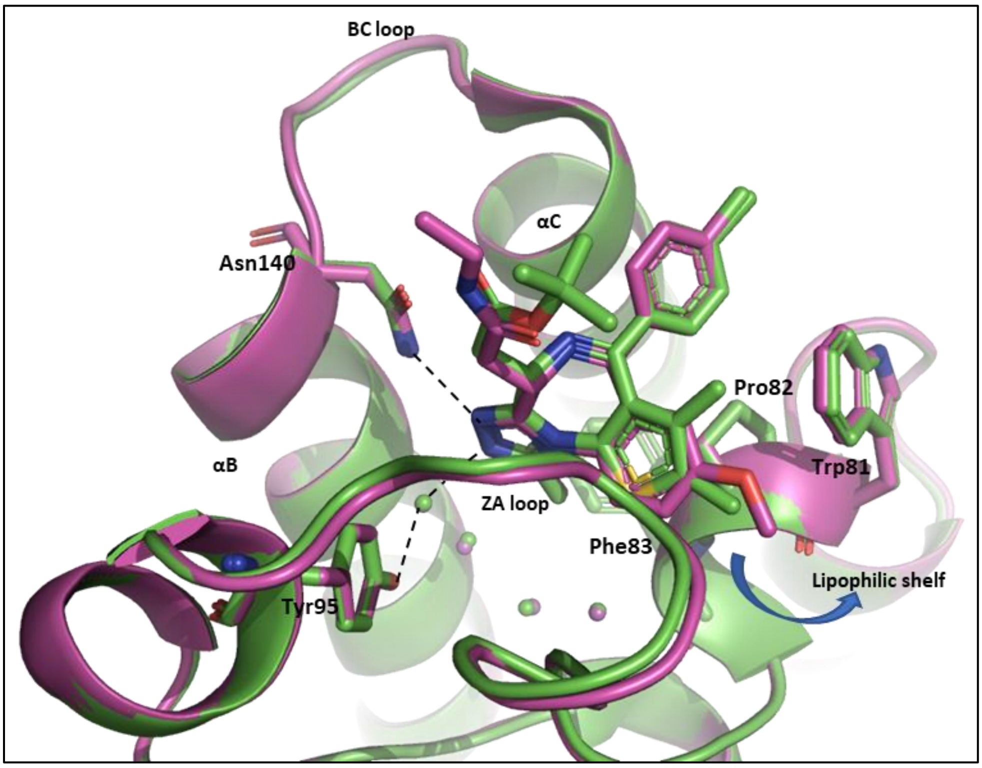


Figure 5-2 Overlay of BRD4-BD1 protein in complex with JQ1 (magenta, PDB: 3MXF) and I-BET762 (green, PDB: 3P5O). The key residues are represented in stick, the lipophilic shelf is highlighted with a blue arrow.

Following the discovery of **a** and **b**, the number of reported BET inhibitors has increased significantly leading to the discovery of different scaffolds (**c-i,** Figure 5-3) able to interact with the KAc binding site and particularly with the conserved asparagine residue[165] (portions highlighted in red). Compounds **a-c** show little or no selectivity between members of the BET family.

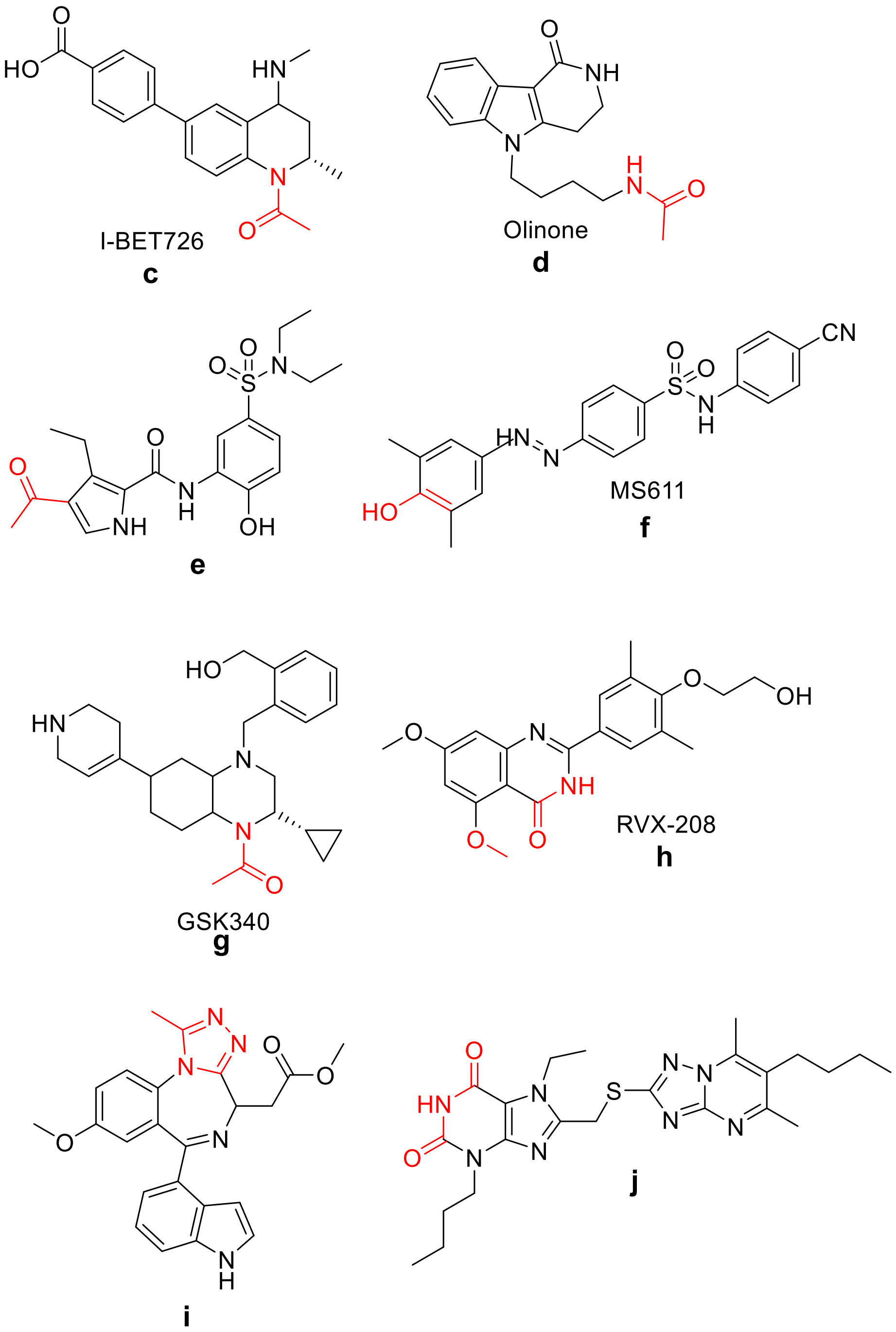


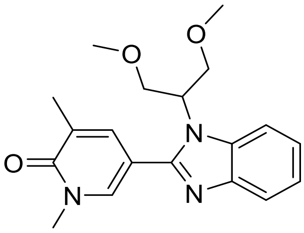
Figure 5-3 Examples of BET inhibitors. The KAc mimetics are highlithed in red.

Despite the high structural similarity of the BET proteins, the function of each member is likely to be different[169, 170]. The identification of compounds that allow intra-BET selectivity and associate with high confidence the biological function to a particular BET protein has been a major challenge.

Therapeutically, the effects of the pan-BET inhibitors on different transcriptional pathways have raised concerns about the safety and tolerability of pan-BET inhibitors in humans[170]. Therefore, an alternative strategy to improve the understanding of the BET BRDs and decrease the toxic profile of the pan-BET inhibitors is to selectively inhibit one of the two tandem domains, BD1 or BD2, in the BET BRDs.

Olinone[171] (**d**) showed 3.4 μM affinity for BRD4-BD1 with no activity for BD2; the acylpyrrole, **e**[172], showed 0.24 μM affinity for BRD4-BD1 and 10-fold selectivity over other members of the BET family. The diazobenzene MS611[173] (**f**) with the phenol acting as KAc mimetic, was reported to be moderately active for BRD4-BD1 (Kd= 0.41 μM) with 100-fold selectivity over BRD4-BD2. GSK340[174] (**g**) showed nanomolar BRD4-BD2 potency and 50-fold selectivity over BRD4-BD1. RVX-208 (**h**) was developed at Resverlogix[175] and was found to have ~20 fold selectivity for BRD3-BD2 over BRD3-BD1 domain; it is currently in phase III clinical trials for cardiovascular diseases. The indole, **i**[176] had 20-fold selectivity for BRD2-BD2 over BRD2-BD1 (reduced to ~10 fold for BRD4). Lastly, the xanthine derivative (**j**[177]) exhibited 5μM affinity for BRD4-BD2 with 10-fold bias over other BD1 domains and undetectable BD2 binding. Overall the compounds reported in Figure 5-3 are not stronger binders and have limited selectivity, showing that finding a domain-selective compound is far from trivial.

In GSK one of the potential cores for obtaining pan-BET inhibitors was the pyridone benzimidazole. A representative example of the series is compound **k (**Figure 5-4), for which X-ray structures in both BD1 and BD2 domains were solved (see infra).



k

Figure 5-4 Benzimidazole scaffold identified by GSK and the subject of this study. The substituent group (R1-R5) were explored in more details.

Many compounds belonging to the benzimidazole series were synthesized and tested (~ 920). The benzimidazole series was attractive because it showed an intrinsic preference for the first domain (BD1) of the BRD4 protein. Given the high structural conservation of BET bromodomains and the many transcriptional pathways and tissue specific functions that BET members have, finding inhibitors of the BET protein member which are domain-specific could improve not only the biological understanding of the specific domain but also of the risk associated with their safety and tolerability in humans.

5.2 Domain selectivity

The sequence identity between the four dual-BRDs BET proteins is high; however, several differences exist between the BD1 and BD2 domains. For BRD4 (Figure 5-5), the "gatekeeper" residue which restricts the access to the WPF region, switches between Ile146 in BD1 and Val439 in BD2; in addition a Gln85/Lys378 mutation is observed at the base of the ZA channel. The greatest difference is in the BC loop, with Asp144 in BD1 exchanged for His437 in BD2, aspartic to glutamic acid mutation of the adjacent residue and lastly a lysine to proline mutation close to the conserved Asn140/433.

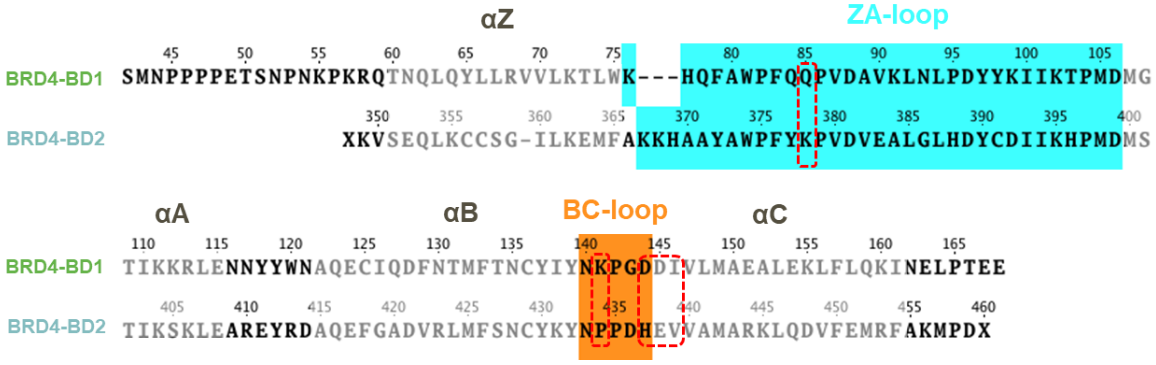

αZ
ZA-loop
BRD4-BD1
SMNPPPPETSNPNKPKRQTNQLQYLLRVVLKTLWK---HQFAWPFQQPVDAVKLNLPDYYKIIKTPMDMG
BRD4-BD2
XKVSEQLKCCSG-ILKEMFAKKHAAYAWPFYKPVDVEALGLHDYCDIIKHPMDMS
αA
αB
BC-loop
αC
BRD4-BD1
TIKKRLENNYYWNAQECIQDFNTMFTNCYIYNKPGDDIVLMAEALEKLFLQKINELPTEE
BRD4-BD2
TIKSKLEAREYRDAQEFGADVRLMFSNCYKYNPPDHEVVAMARKLQDVFEMRFAKMPDX

Figure 5-5 Sequence alignment of BRD4-BD1 and BRD4-BD2; the key differences are highlighted in a box with dashed red line: Q85/K378, K141/P434, D144/H437, D145/E438, I146/V439.

In relation to the benzimidazole series, the crystal structures of **k** in BRD4-BD1 and BD2, showed that the pyridone ring, also called the “warhead”, is mimicking the interaction with Asn140 of the KAc. The benzimidazole extends into the ZA channel, situated between the αZ and αA helices of the bromodomain and lastly, the diethyl-ether group is pointing towards the WPF shelf region (Figure 5-6).

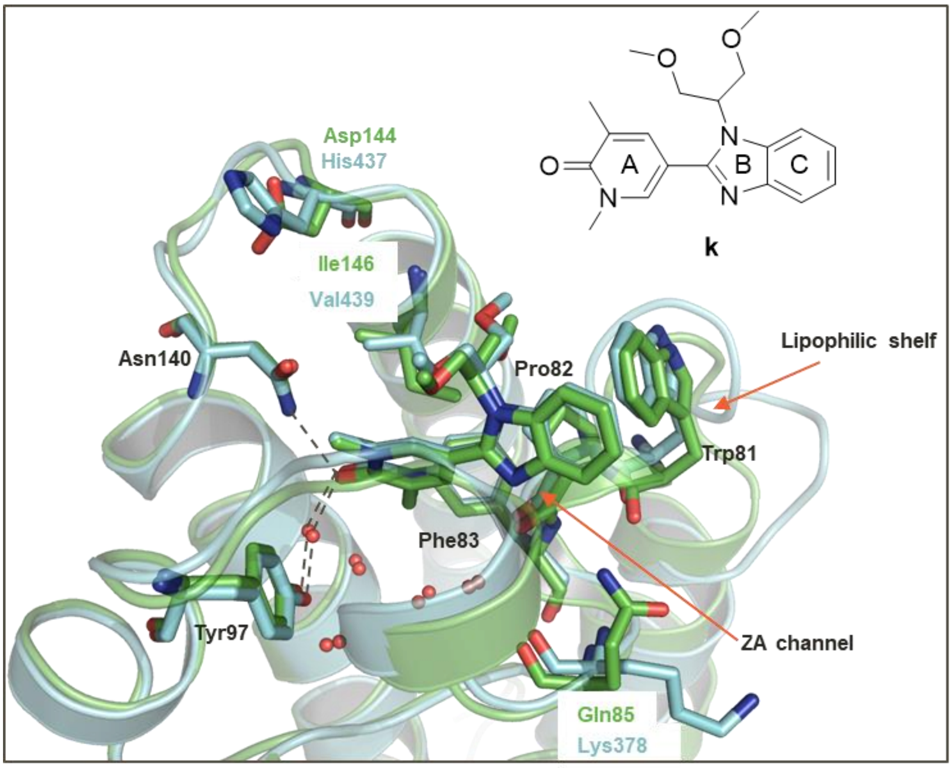


Figure 5-6 X-ray crystal structures overlay of BRD4-BD1 (green) and BD2 (cyan) with compound **k** bound. The lipophilic shelf (Trp81, Pro82, Phe83) and the ZA channel are indicated by the orange arrows.

5.3 Structure–Activity Relationship (SAR) of benzimidazole series.

The benzimidazole scaffold with the pyridone “warhead” allows functionalization at up to five positions, R1-R5 which are extending in different regions of the BRD4 binding pocket (Figure 5-7, core B and C).

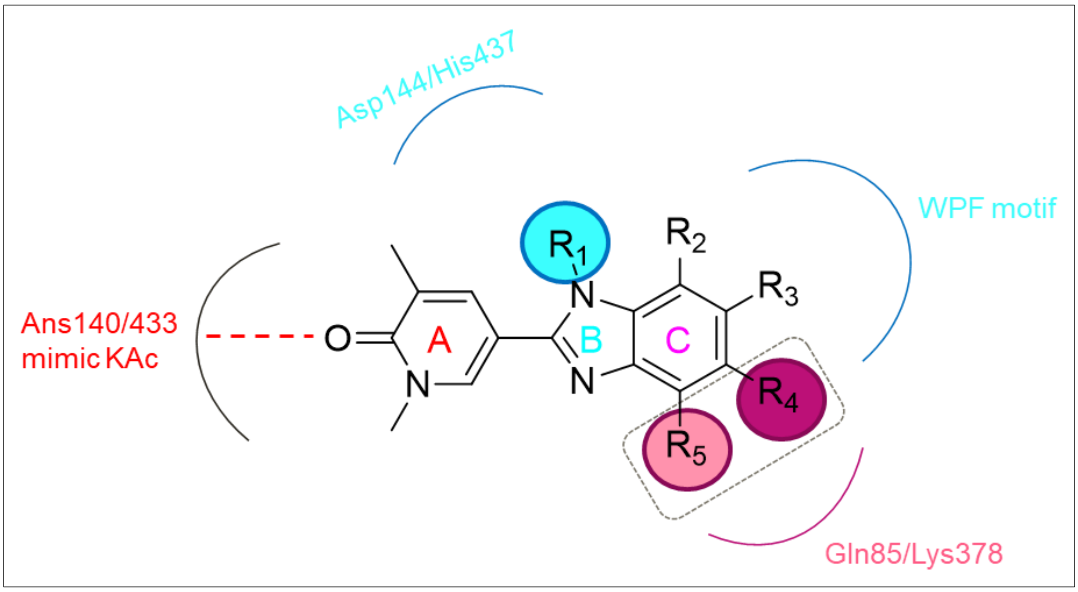


Figure 5-7 Schematic representation of the benzimidazole scaffold with the functional groups (R1-R5) in relation to the residues in ligand-binding site.

In order to maximise the understanding of the already synthesised compounds, Structure-Activity Relationship (SAR) analysis was generated by studying the effect of the R substituents on compounds with a common scaffold, also called matched molecular pairs. The information obtained from a successful SAR analysis can be used to make rational structural modifications to optimize some properties, in this case potency and selectivity.

The SAR analysis was conducted by studying the R1 substituents which are likely to extend towards the WPF shelf or in the region where the Asp144/His437 single point mutation is present and then, the R4 and R5 position which are inserting in the Gln85/Lys378 amino acid single point mutation.

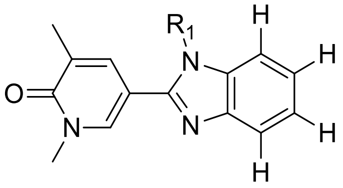
5.3.1 R1 substituents

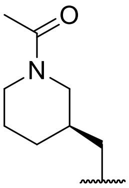
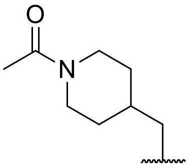
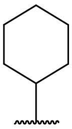
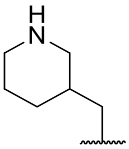
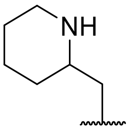
The effect of the R1 group was studied using the core as reported in Table 8. Compound **9** exhibits modest potency for both domains, pIC_{50} values of 5.7 and 4.8, respectively, and 0.9 log unit selectivity towards BD1. In Table 8 the effect of varying potency and selectivity by changing R1 is shown. The comparisons in the following discussion are with compound **9**, unless otherwise stated.

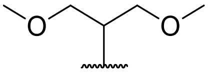
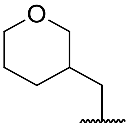
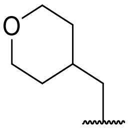
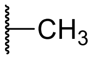
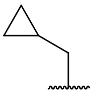
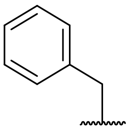
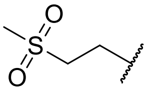
It was observed that by extending the methyl substituent with a six-member aliphatic ring containing nitrogen (**4**,**5**) or replacing it with cyclohexane (**3**) the BD1 potency increases followed by an increase in selectivity toward BD1 up to 1.4 log unit. This is in line with the hypothesis of gaining potency and selectivity by accessing the WPF shelf. At the same time, in the case of cyclopropane (**10**) or tetrahydropyran (**7**,**8**) an increase in potency for both domains is observed with no improvement in selectivity. The sulfone substituent is not potent (**12**, $pIC_{50} < 6$) and has no selectivity. The aromatic ring in **11** shows improved potency (pIC_{50} BD1= 7.6 and BD2= 7.2) but for both domains therefore lacking in the required selectivity. Given the higher BD2 potency of **11**, it could be hypothesized that the aromatic ring is likely to interact with both the WPF shelf and His437, which is only present in BD2, therefore such aromatic substituents are not appropriate for the required objective. The highest selectivity of up to 1.8 log units, is reached by the acetyl capped piperidine substituent (**1**) when the nitrogen is inserted in the meta position while when the nitrogen is in para position (**2**) a significant drop in BD1 potency is observed.

Given the potency and selectivity of compound **1**, it was used as the starting compound for further SAR exploration and was the subject of computational studies.

Table 8 R1 analysis of the benzimidazole scaffold. Compounds are shown from high to low selectivity calculated as $pIC_{50(BD1)} - pIC_{50(BD2)}$.



R1	*pIC_{50} BD1/ pIC_{50} BD2* *(selectivity)*	*Cmpd*
	8/6.2 (1.8) (starting compound)	**1**
	6.8/5.10 (1.7)	**2**
	7.4/6.0 (1.4)	**3**
	6.2/5 (1.2)	**4**
	6.8/5.7 (1.1)	**5**

	7.3/6.2 (1.1)	**6 (k)**
	7.6/6.6 (1)	**7**
	7.2/6.4 (0.8)	**8**
	5.7/4.8 (0.9)	**9**
	6.7/6 (0.7)	**10**
	7.6/7.2 (0.4)	**11**
	5.7/5.4 (0.3)	**12**

5.3.2 R4 and R5 substituents

The effect of the R4 substituents was analysed using matched molecular pairs with three different R1 substituents (Figure 5-8): piperidine capped with acetyl (Core 1), diethyl-ether (Core 2), oxane (core 3) while the rest of the substituents (R2, R3 and R5) were kept as hydrogens.

The effect of the R4 seems independent from the R1 group, as can been seen from the parallel lines in Figure 5-8. In addition, despite the wide range of R4 substituents, the change in selectivity is minimal, hence, R4 substituents are not important for selectivity.

Compared to the starting compound **1** with hydrogen at position R4, the piperazine motif appears to be the best one for Core 1 with an increase of 0.3 log units in selectivity and a total selectivity of 2.1 log units, the methoxy-piperidine motif is best for Core 2 although it is not contributing to add selectivity and the morpholine motif is best for Core 3 adding 0.6 log units of selectivity.

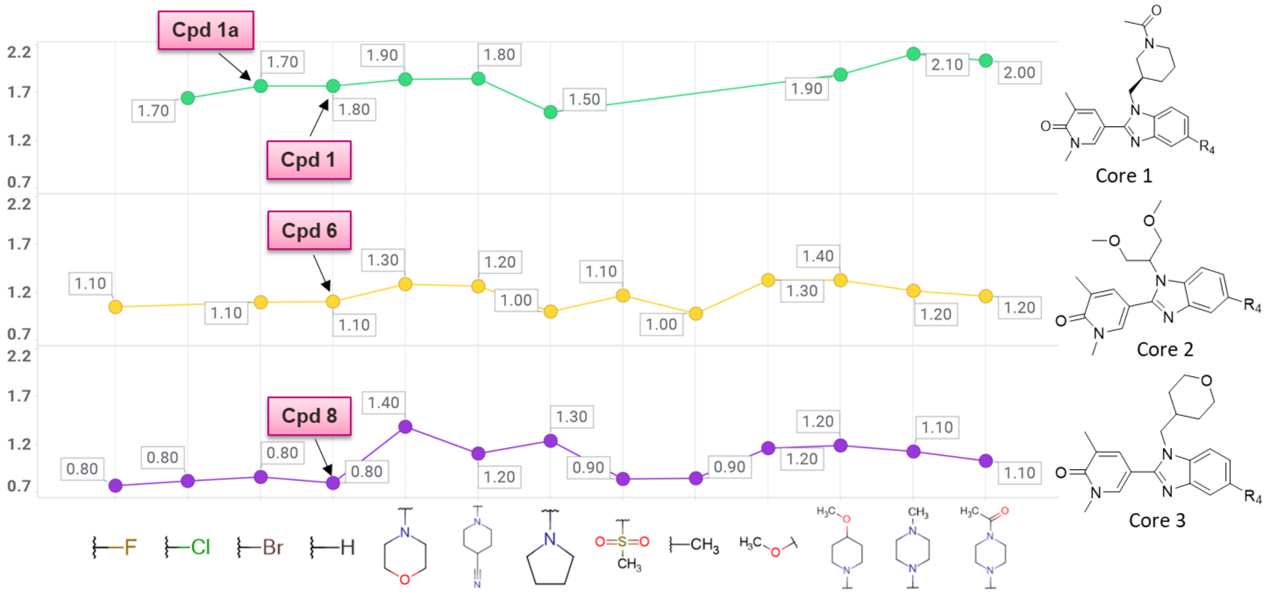


Figure 5-8 Match-molecular pairs analysis of R4 substituents with respect to selectivity. Three different cores were used as reference Core 1 in green, Core 2 in yellow and Core 3 in purple. Size and labels are accordingly to selectivity. The compound **1a** for which a crystal structure was present is highlighted as well as compounds present in Table 8.

Finally substituents on the R5 position were analysed (Figure 5-9). The capped acetyl piperidine (**1**) has a selectivity of 1.8 log unit when all substituents are hydrogens while when R5 is either methoxy or chlorine the selectivity decrease to 1.4 log unit (Figure

5-9, blue). It is important to highlight that the data retrieved for the capped acetyl piperidine (except of **1**) refers to the racemic mixture. In this case, the R5 substituent is not significantly contributing to the selectivity. For the R1 substituents equals to tetrahydropyran (Figure 5-9, magenta), a marginal improvement in selectivity is reported. In general, it is observed that the R5 group is detrimental for BD2 potency leaving BD1 potency approximately unchanged.

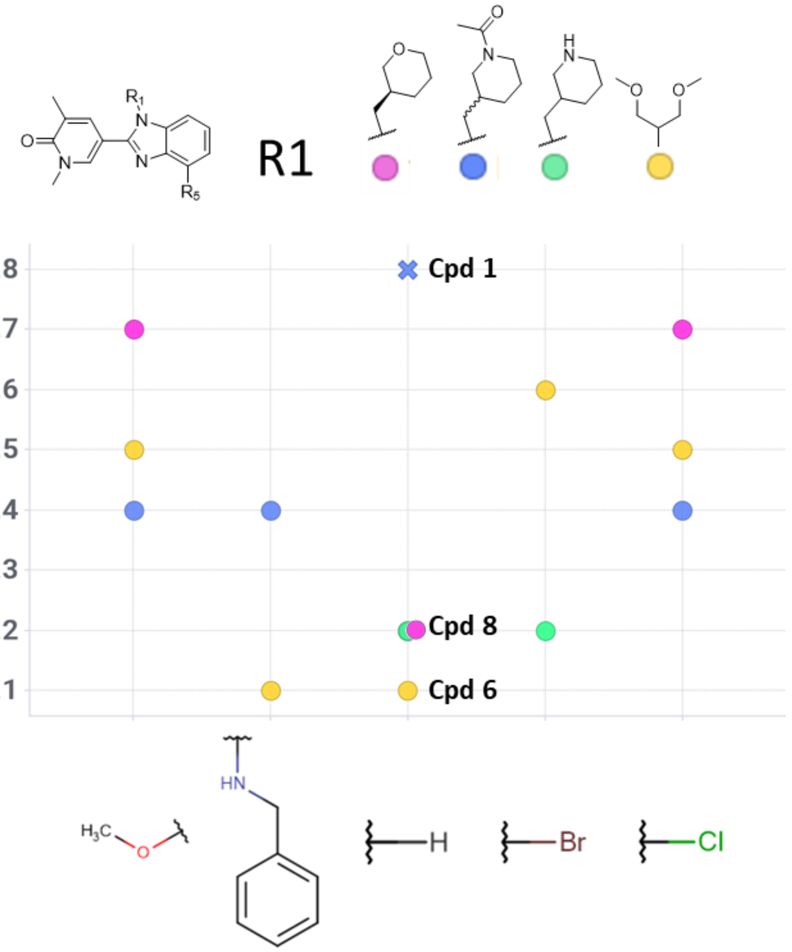


Figure 5-9 Analysis of R5 position with varying R1 substituents. Cpd **1** in cross shape refers to the enantiomer and has a

In summary from the SAR analysis some conclusions could be drawn. Firstly, despite the possibility to decorate the benzimidazole core with many substituents, the extension of the R1 position seems key to achieve BD1 selectivity. Additionally, aromatic rings at the R1 position should be avoided because of the potentially additive hydrophobic interaction with His437. The acetyl capped piperidine ring is providing both potency and selectivity towards BD1 hence, the design was structured around this starting compound. The R4 and R5 positions were not adding any significant boost in potency and/or selectivity despite being in the region where the Gln/Lys mutation occurs.

5.4 Aims

The aim of this project is to guide with computational tools the design of a selective (> 1000 fold) chemical probe for the BD1 BET family by using the BRD4 protein. The initial SAR analysis was helpful to identify the R1 position as the key region for the design of new molecules. In the protein such region corresponds to the Asp144/His437 mutation. Computational studies are carried out to understand how such structural difference can be explored to obtain selective and potent ligands for the BET BD1 domain.

5.5 Methods

Through this work the OPLS-3e force field was used for protein preparation, ligand preparation, Glide Grid generation, docking, MD and MetaD simulations. The BRD2-BD2 protein was used as surrogate for the BRD4-BD2 protein because it was difficult to crystallize.

5.5.1 Complex preparation

The apo form of BRD4-BD1 (PDB: 2OSS) and of BRD4-BD2 (PDB: 2OUO), the complex of **1a** in BRD2-BD2, and the complexes with GSK code 5urkn and 8bltk were prepared with the Protein Preparation Wizard[142] tool in Maestro v.2018.04. Bond orders were assigned to the ligands, hydrogens were added. The hydrogen bonding network was optimized with the ProtAssign algorithm at neutral pH such that the amide group of Asn, and Gln residues, thiol and hydroxyl groups and the imidazole rings of His residues were adjusted. In particular, the His437 protonation state was selected to be HIE given the interaction with the NH of the backbone of the gatekeeper (Val439) when in the closed conformation. Restrained minimization using the Impref module of Impact and the OPLS-3e force field was performed for refinement of the structures. This minimization continued until an average RMSD of the non-hydrogen atoms reached the specified limit of 0.3 Å.

5.5.2 Molecular docking

Prior to docking calculations, the ligands **13** to **20** were drawn in the Maestro[143] (v.11.8.012, release 2018.04) panel interface using the 2D-Sketcher panel. Then the ligands were prepared with the LigPrep[178] module to determine the 3D structure and ionization state at pH 7.0 ± 2.0. The compounds **13** to **18** were modelled in their protonated form.

The Glide v8.1 molecular docking tool[94, 179] (Schrodinger, 2018.04) in Standard Precision (SP) mode was used to generate binding mode for compound **13** to **20** in the BRD4-BD1 and BRD2-BD2 bromodomains.

The x-ray crystal structures with gsk code 5urkn (BRD4-BD1) and 8bltk (BRD2-BD2) were initially processed with the Protein Preparation Wizard in Maestro[143] as described in 5.5.1 and used to generate the pre-computed grid. The bound ligands were selected to define the binding site. The four water molecules at the bottom of the pocket were retained during the preparation step. Default van der Waals scaling (1.0 for the receptor and 0.8 for the ligand) were used. Under the Advanced Settings it was specified to use enhanced sampling for conformer generations and expanded sampling for the selection of initial poses. It was selected to keep at least 4 poses per ligand.

The docked complexes of **13** and **16** were then visually inspected and the poses that were satisfying the H-bond interaction with Asn140/433 and inserting the R1 group in the Asp144/His437 cavity were selected and used for WaterMap and MD simulations.

An x-ray crystal structure of **13** in BRD2-BD2 was solved later. The overlay between docking and x-ray crystal structure is reported below.

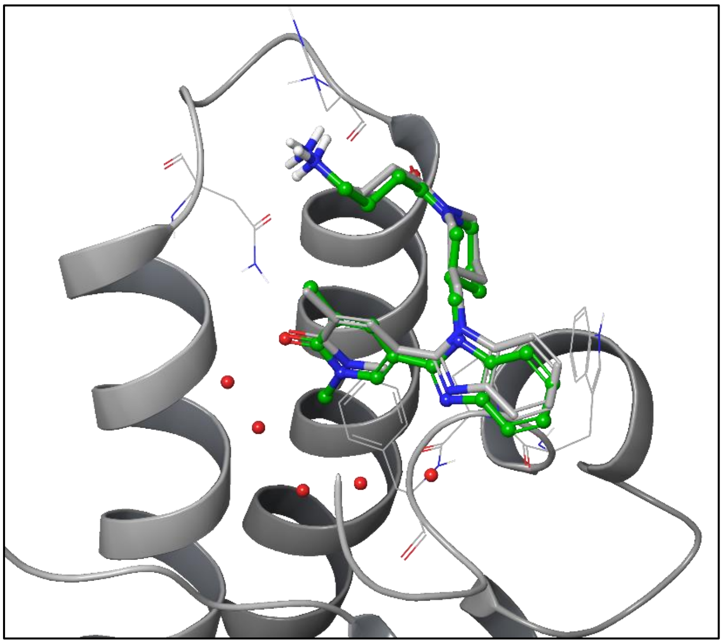


Figure 5-10 Overlay of x-ray structure of **13** in grey with docking solution in green. The most important residue in the ligand binding site are reported in lines. The ligand heavy atoms RMSD between the x-ray structure and docking solution was 0.48 Å. The ZA loop is omitted for clarity.

5.5.3 Molecular Dynamics

The docked solutions of compound **13** and **16** in BRD4-BD1 and BRD2-BD2 appropriately prepared as described in Section 5.5.1 were used as starting structure for MD experiments. MD were carried out using Desmond[146] v.5.6 and the OPLS3e force field. By using the System Builder tool, the prepared complexes were solvated in a box of TIP3P water with a size of 10x10x10 Å, ions were added first to neutralize the system and then to reach a concentration of 0.15 M. At this stage, the system was relaxed and submitted to the MD protocol which consist of several stages. The system was relaxed at 10 K by 100 ps Brownian dynamics NVT simulation with restraints on the solute heavy atoms (50 kcal/mol/Å^2). The solvent box was then equilibrated at 10 K by 12 ps of NVT simulation and 12 ps of NPT simulation with restraints on the solute heavy atoms (50 kcal/mol/Å^2). The system was then heated to 300 K and full equilibration was performed in the NPT ensemble for 12 ps with restraints on the solute heavy atoms (50 kcal/mol/Å^2). The last relaxation procedure was a 24 ps NPT dynamics run at 300 K without restraints. Finally, the production simulation was carried out in the NPT ensemble for a length of 1µs. A constant temperature of 300K was maintained using the Nosé-Hoover thermostat algorithm and Martyna-Tobias-Klein barostat algorithm to maintain the pressure of 1 atm. The short-range coulombic interactions were evaluated using the short-range method with a cut-off value of 9 Å. The long-range electrostatic interactions were calculated using the particle mesh Ewald method. The nonbonded forces were calculated using a RESPA integrator where the short-range forces were updated every 2fs and the long-range forces were updated every 6 fs. The trajectories were saved every 500 ps for analysis, for a total of 2000 frames for each replica.

Trajectories were additionally processed in VMD 1.9.2[180], *i.e.*, the protein was centered in the water box by using the pbc tool and the trajectories were aligned on the protein backbone heavy atoms using the RMSD Trajectory tool. The trajectories were loaded in a custom Jupyter notebook and further analysed with MDtraj and Pytraj packages.

5.5.4 Metadynamics

The prepared complexes of the BRD4-BD2 apo form and the BRD2-BD2 model bound to compound **1** were solvated with TIP3P water molecules in a box of 10x10x10 Å. The model of **1** was generated by removing the R4 group of **1a** crystal structure (gsk8bltk). Ions were added to neutralize the system and to reach a concentration of 0.15 M.

The prepared complexes were then submitted to the MetaD simulations with Desmond[146] v5.6 using as CVs the χ_1 and χ_2 torsional angle of His437. Gaussians were deposited every 0.09 ps with a hill's height of 0.03 kcal/mol. The width of the Gaussian was set to 5°. The MetaD simulation lasted less than 100 ns and was replicated three times changing the seed to assign starting velocities.

The evaluation of the convergence was based on three criteria.

1) The FES was compared through an interval of 10 ns, and to consider a simulation converged, the free-energy profile should not change significantly. This was achieved by comparing the FES at different simulation time.
2) The evolution of each CV was monitored so that they were not trapped in specific energy minima.
3) Finally, it was checked that the FES profiles of different replicas were superimposable.

In the case of the three replicas of the apo BRD4-BD2 form the FES were calculated by using 65 ns, 218 ns and 171 ns. The block average was calculated every 5ns, 10 ns and 20 ns respectively to have a total of 12 FES to compare.

5.5.5 WaterMap

The WaterMap calculations were run with the default parameters using BRD4 BD1 and BD2 apo protein (PDB:2OSS and 2OUO) and the docked solutions of compound **13** and **16** in BRD4-BD1 and BRD2-BD2 appropriately prepared as described in Section 5.5.1.

From WaterMap the location, occupancy, enthalpy, entropy and free energy of the water molecules are calculated by doing a combination of molecular dynamics, solvent clustering and statistical thermodynamics analysis.

For the definition of the binding site, the ligand was selected and retained for the 2.0 ns long MD simulation with OPLS3e force field; in the case of the apo protein the residues Asn140/433, Asp144/His437 and Ile146/Val439 were selected as reference for the binding site. Water sites around Asp144/His437 were analysed. The existing water molecules at the bottom of the pocket were retained and treated as solvent in the calculations. The coordinates of the protein are restrained with a 5.0 kcal/mol/Å^2 harmonic potential on the heavy atoms to ensure convergence of the water sampling around the protein conformation. The frames from the MD simulation are then spatially clustered to form localized hydration sites meaning that in those positions there is a high water occupancy, and their thermodynamic properties are calculated relative to bulk solvent using IST. In particular, enthalpy is calculated as the average non-bonded molecular mechanics interaction energy of the water in the hydration site with the rest of the system; entropy is calculated numerically by integrating a local expansion of spatial and orientational correlation functions (Section 2.6) and described in ref. [115, 116]

5.6 Results

It should be noted that because of the difficulties in crystallizing the compounds in the BRD4-BD2 protein, the BRD2-BD2 protein which is instead possible to crystallize was used as a surrogate. This is a common practice used in GSK given the high residue conservation among the two isoforms.

From the SAR analysis, compound **1** (pIC$_{50}$= 8 BRD4 BD1 and BD2= 6.2) was optimized further to obtain higher BD1 over BD2 selectivity. At the outset of this work, the X-ray of a close analogue of compound **1**, compound **1a** (Figure 5-11, pIC$_{50}$= 8 BRD4 BD1 and BD2= 6.3) bound to BRD2-BD2 was available. The overlay of the X-ray structure of the apo form of BRD4-BD2 (PDB: 2OUO) and **1a** (gsk8bltk), revealed that the R1 substituent, acetyl capped piperidine, is orienting towards the His437 and not the WPF shelf as observed for compound **k** (Figure 5-6). Moreover, the His437 is found in a different conformation from the apo form, pointing outward from the ligand-binding site.

The higher selectivity of **1a** of 1.7 log unit as opposed to other synthesised compounds could be attributed to the R1 substituents extending in the region of the BC loop where the Asp144/His437 mutation occurs. Therefore, investigation into R1 substituents that extend toward the Asp144/His437 region was proposed as primary strategy to achieve selectivity for BD1 over BD2.

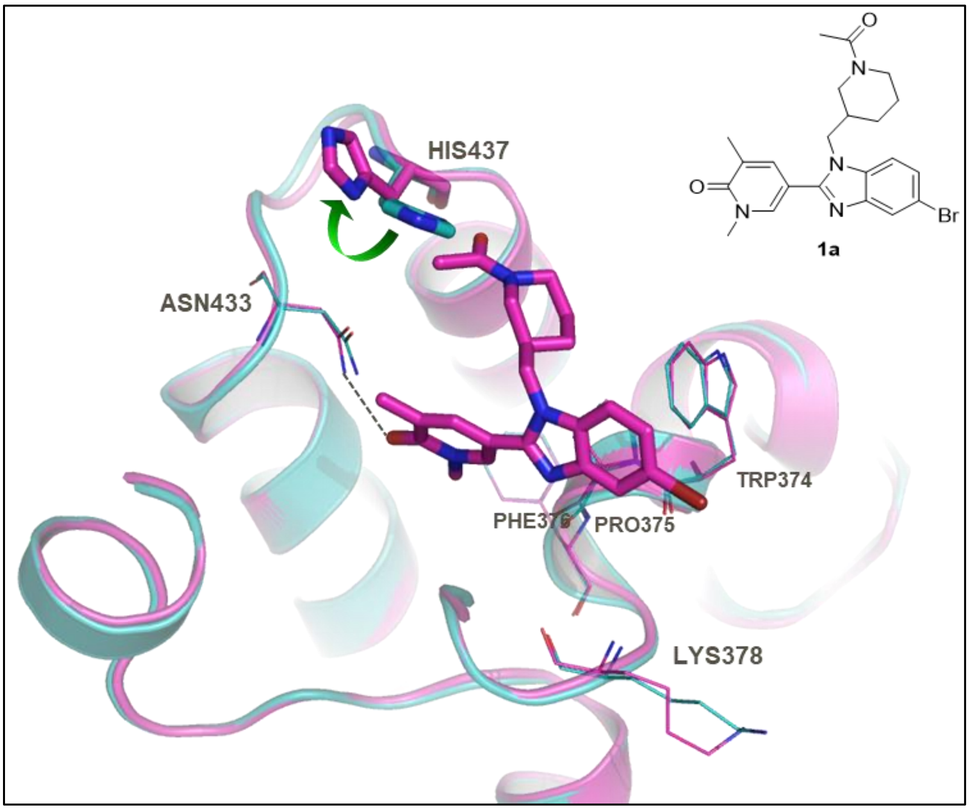


Figure 5-11 Crystal structure of ligand **1a** in BRD2-BD2 in magenta overlaid to the BRD4-BD2 APO crystal structure in cyan. The hydrogen bond interaction with Asn433 is reported in black. The most important residues for ligand binding are reported in lines. The rearrangement of His437 (stick) is represented by a green arrow.

5.6.1 Analysis of the conformational states of His437: crystal structures

To study in more detail the role of the His437 in ligand binding, all crystal structures of the BRD2-BD2 protein, a close analogue of the BRD4-BD2 protein, available in the PDB and GSK repository were firstly analysed. From the His437 χ1 (N-CA-CB-CG) and χ2 (CA-CB-CG-ND1) torsional it was observed that four distinct orientations can be identified, as reported in Table 9.

Table 9 Classification of His437 with respect to the $\chi1$ and $\chi2$ torsional angles of the 628 BRD2-BD2 complexes under investigations.

Name	χ_1 (range)	χ_2 (range)
Closed (1)	-179.6, -159.7	78.7, 146.6
Closed (2)	-177.3, -153.5	-157.3, -46.5
Open (1)	-89.7, -72.9	17.1, 74.0
Open (2)	-109.7, -73.0	-162.7, -137.8

By visual inspection, the 4 categories can be simplified into two if only the $\chi1$ orientation is used: “closed” if the $\chi_1$varies between -179.6 and -153.4 and “open” if χ_1 varies between -109.7 and -72.9 (Figure 5-12).

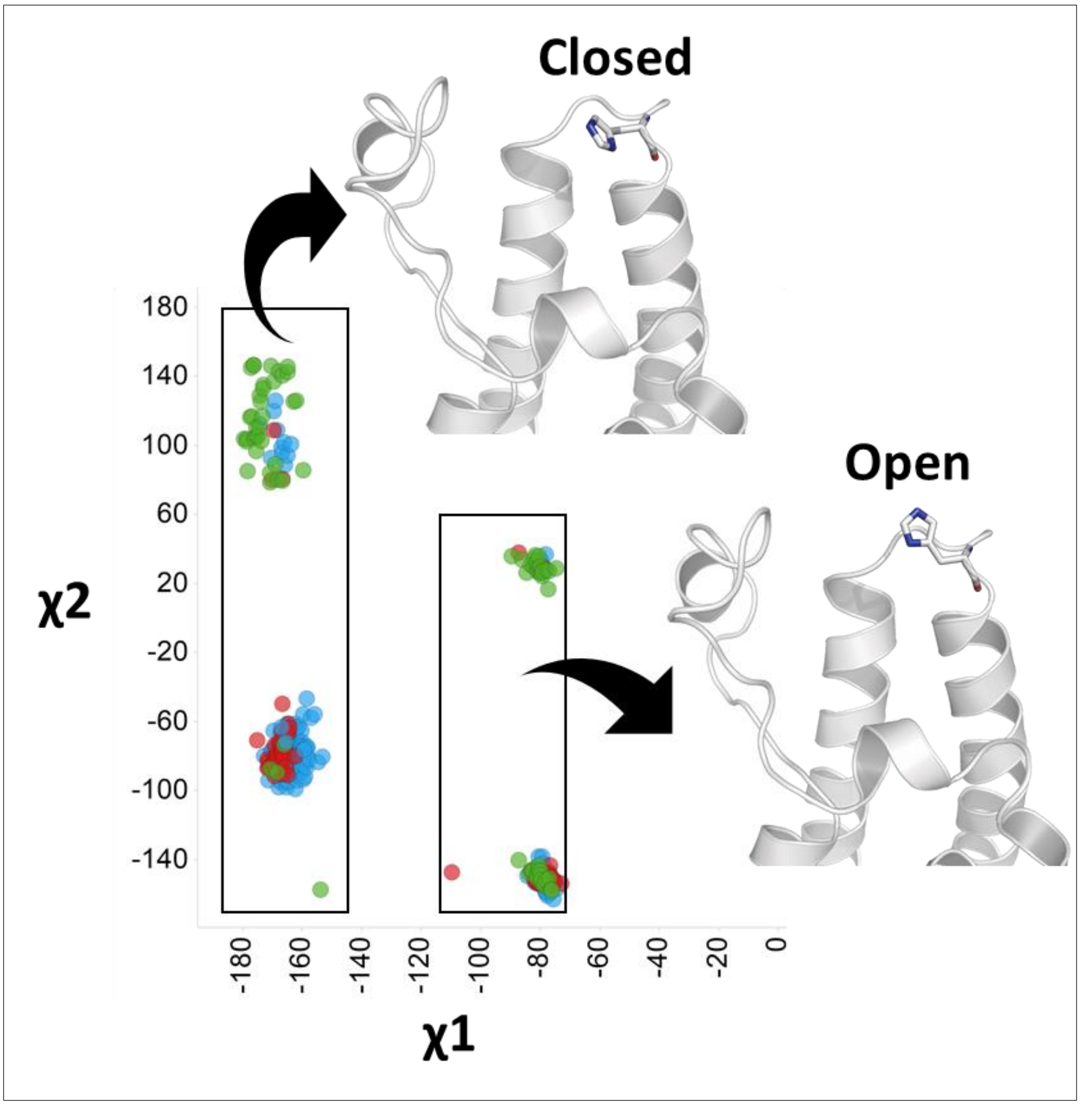


Figure 5-12 χ1 χ2 orientation of the deposited crystal structures of BRD2-BD2 protein. The complexes in blu refer to structures with only 1 conformation, the structures in red and green refers to dual His437 conformation.

From the 628 BRD2 structures analysed, the majority, 330, are clustered in the space of the closed conformation and the remaining 214 are in the open conformation. Interestingly, in 84 complexes, the His437 is reported to have a dual orientation (A/B letter in the PDB file) meaning that the residue is mobile and thought to occupy both open and closed conformations.

Out of the 207 complexes with only the open conformation (Figure 5-13, upper left), 75 (36%) ligands show from 0.5 up to 2 log units of selectivity towards BD1, about half of them (115 ligands) bind equally to both BD1 and BD2 domains and only 17 ligands (8%) have BD2 preference. On the contrary, the majority of ligands with only the closed conformation (Figure 5-13, lower left) bind with higher preference towards BD2; more precisely out of 314, 211 (68%) have from 0.5 to 3.5 log units selectivity towards BD2 and only 16 (5%) have from 0.5 to 1.5 log units selectivity towards BD1. In the 84 cases with dual conformations (Figure 5-13, right), there is not a clear trend with selectivity: 19 (23%) ligands show preference towards BD1, 23 (28%) towards BD2 and 39 (48%) are not selective.

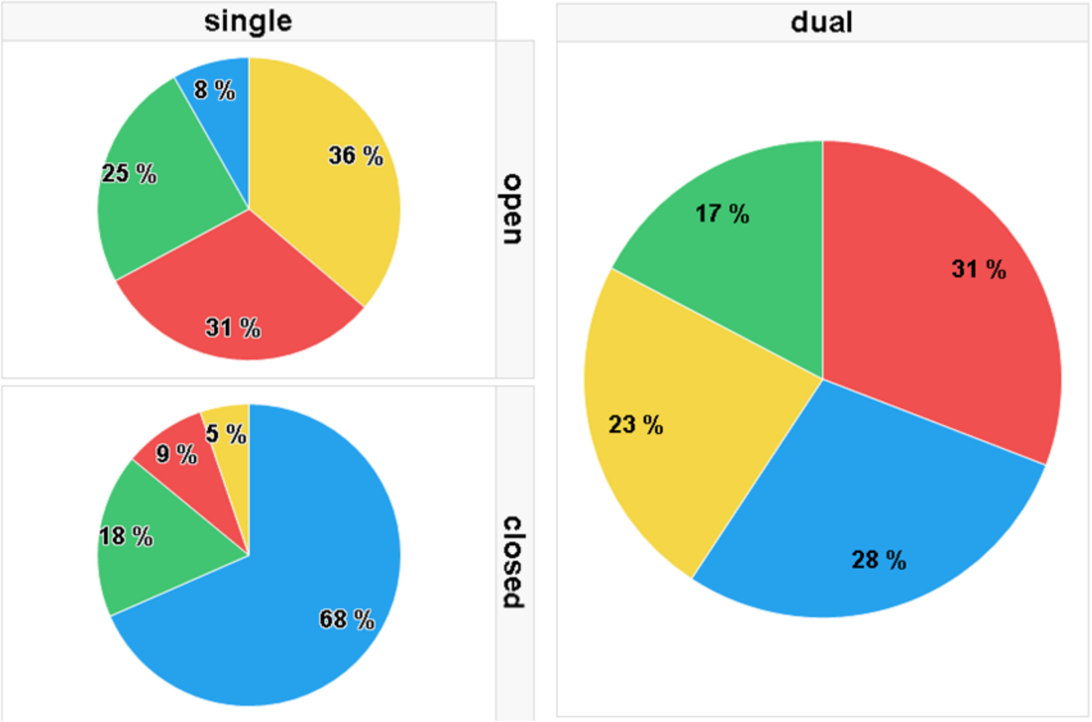


Figure 5-13 Pie chart representing the selectivity profile of the ligand in the BRD2-BD2 protein crystal structures with respect to the His437 orientation being deposited as single open/closed on the left or dual orientation on the right. The selectivity (s) was calculated by subtracting to the BD1 potency from the BD2 ones [pIC$_{50\ (BD1)}$) – pIC$_{50\ (BD2)}$] and four categories were identified: 1) compounds BD1 selective in yellow, s > 0.5 ; 2) compounds BD2 selective in blue, (s ≥ -0.5); 3) compounds with slight preference towards BD1 in green, (0 < s ≤ 0.5); 4) compounds with slight preference towards BD2 in red, (-0.5 < s ≤ 0).

In summary, from the analysis of the $\chi1$ and $\chi2$ of His437 in the available crystal structures it is observed that molecules that bind to BD2 when His437 is in: i) *open* conformation show intrinsic preference for BD1 domains (Figure 5-13, upper left); ii) *closed* conformation show intrinsic preference for BD2 (Figure 5-13, lower left); iii) *dual* conformation do not have a clear domain preference (Figure 5-13, right).

5.6.2 Analysis of the conformational states of His437: MetaD simulations

The observed correlation between His437 being in an open conformation and BD1 selectivity could be attributed for example to an unfavourable energetic cost due to the rearrangement of the His437, creation of better ligand and protein interactions (enthalpic gain), water rearrangement (entropy loss) or a combination of all.

To understand the energetic contribution of the His437 conformations, open and closed, and whether biasing it in a specific conformation could be used as a strategy to achieve selectivity, as observed with **1**, metadynamics simulations of the protein in its apo form (PDB: 2OUO) and in complex with **1** were run as described in 5.5.4.

In the apo form of BRD4-BD2 protein it is assumed that the protonation state of the histidine is HIE (Figure 5-14) because of the hydrogen bond interaction between the N-δ and the NH-backbone of the gatekeeper, Val439.

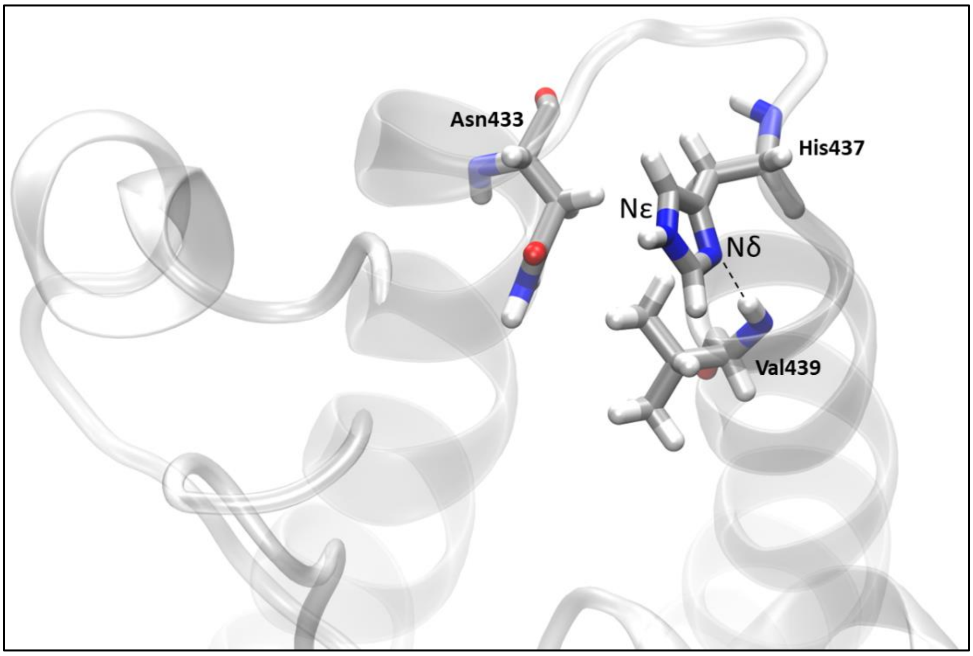


Figure 5-14 BRD4-BD2 apo form with the His437 in the HIE protonation state forming hydrogen bond with the NH-backbone of the Val439 residue in black dotted line. The Nδ and Nε of the imidazole side chain of the histidine are indicated. The ZA loop is hidden for clarity.

In those calculations, the χ_1 and χ_2 torsional were used as the CV to allow a full exploration of all the possible orientations of His437. The resulting Free Energy Surface (FES) showed that in the apo simulation, His437 has three distinct energy minima (Figure 5-15) in the same χ_1/χ_2 region that correspond to the open and closed conformation identified during the analysis of the deposited crystal structures (Figure 5-12). Minima *A* and *B*, which correspond to closed and open conformations, respectively, have a comparable free energy and are expected to be similarly populated in solution. Minimum *C* corresponds to an open conformation of the His437 in which the Nδ of the imidazole ring is pointing out from the binding site. According to the MetaD simulation, minimum *C* is at a 1 kcal/mol ± 0.6 kcal/mol higher than both minima *A* and *B*.

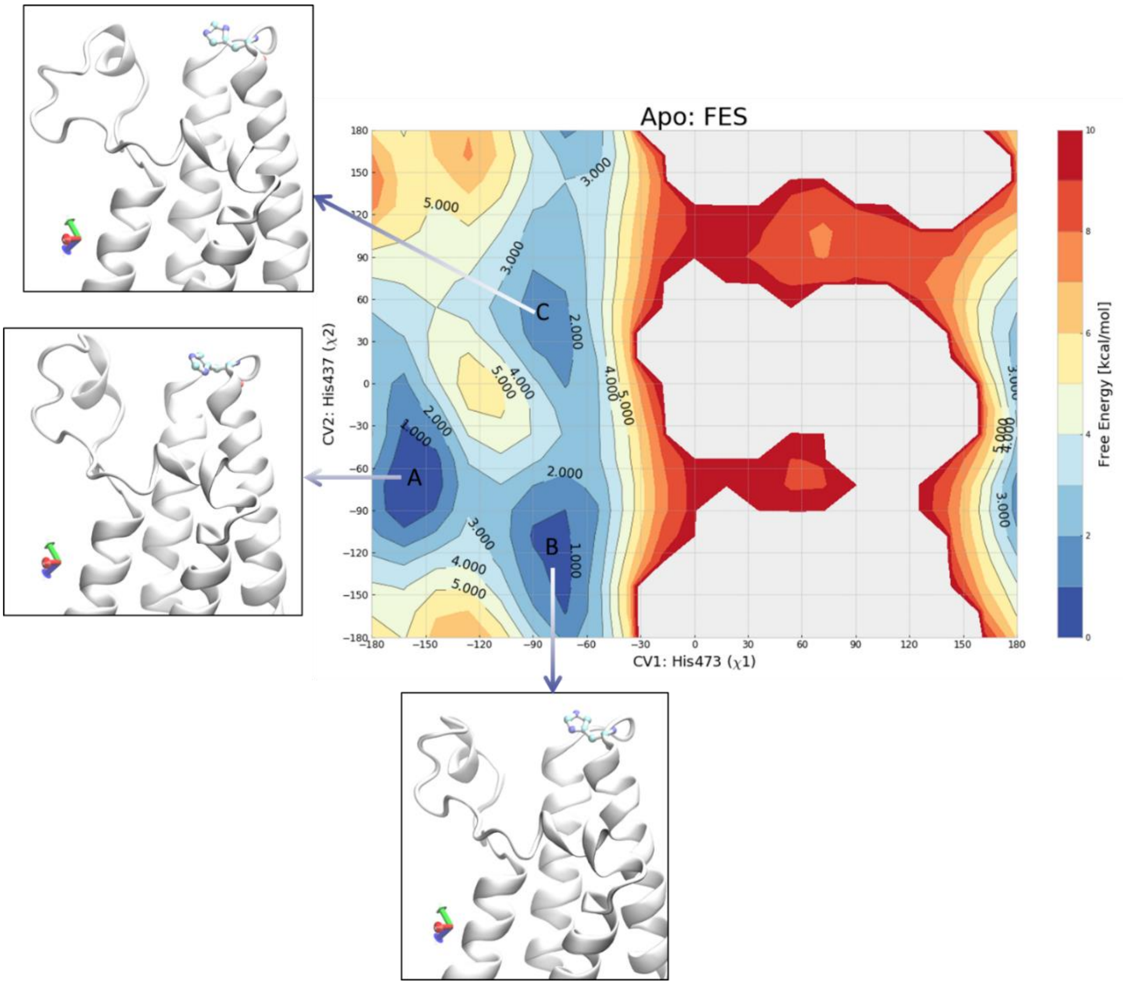


Figure 5-15 Free-energy surface for the apo form obtained from MetaD simulations as a function of CV1 and CV2 which are the χ1 and χ2 torsional, respectively. The free energy is displayed every 1 kcal/mol, the contour levels are shown up to 5 kcal/mol. A representative conformation of minima A, B and C is reported.

In the presence of **1**, the FES (Figure 5-16) as function of χ_1 and χ_2 shows the energy minima in the same positions as in the apo form, which correspond to the open and closed conformation, but the closed conformation (minima *A*, Figure 5-16) is now 1 kcal/mol ± 0.9 kcal/mol higher in energy than in the MetaD of the apo protein and the energy barrier between *A* and *B* has increased from 3 kcal/mol ± 0.6 kcal/mol to 4 kcal/mol ± 0.9 kcal/mol.

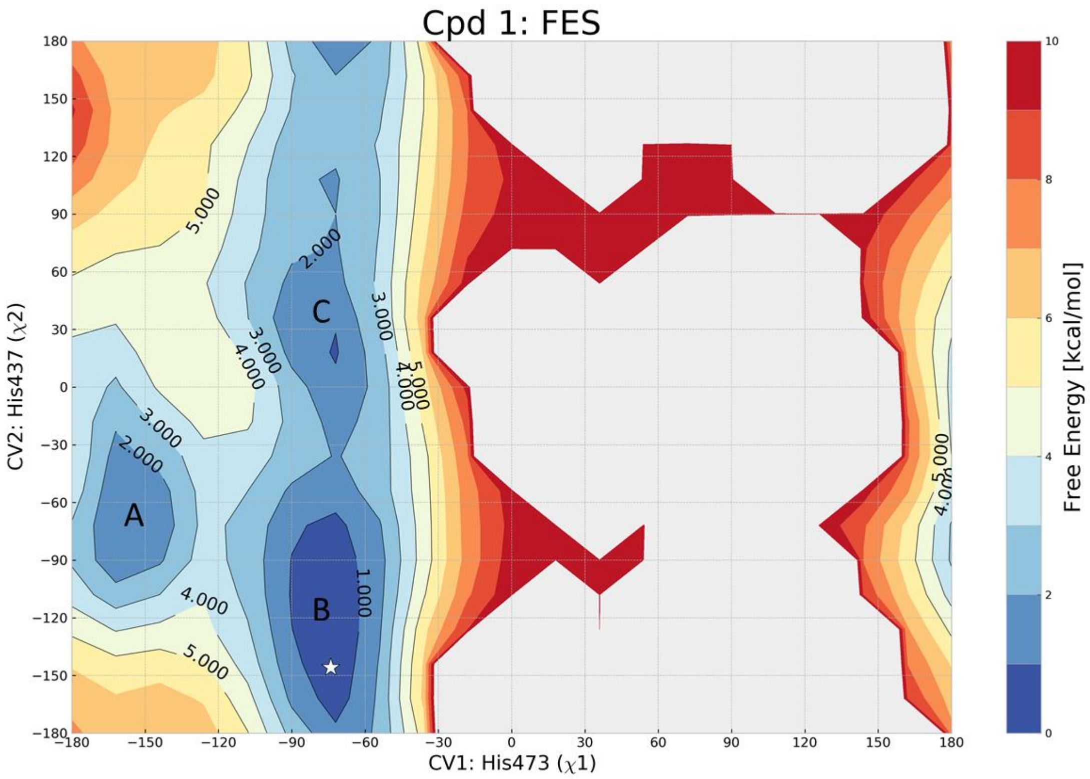


Figure 5-16 Free-energy surface for compound **1** obtained from the MetaD simulations as a function of CV1 and CV2 which are the χ_1 and χ_2 torsional, respectively. The free energy is displayed every 1 kcal/mol, the contour levels are shown up to 5 kca/mol. The white star indicates the χ_1 and χ_2 torsional at the starting of the simulation.

During the MetaD of **1** it could be observed that while His437 is in the closed conformation, the acetyl capped piperidine in the R1 position has to move away from its starting conformation. This indicates an incompatibility between the histidine in the closed conformation and the R1 group.

In conclusion, in the apo form both the open and closed conformations of the histidine are equally accessible. This suggests that favouring only the open conformation of His437 is not going to give BD1 preference, therefore, it should not be regarded as the main strategy to achieve higher selectivity towards the BD1 domain.

5.6.3 Drug Design: expansion towards the Asp144/His437

The MetaD simulations (Figure 5-15 and Figure 5-16) showed that selectivity cannot be achieved by simply forcing the His in one conformation. Therefore, the design was focussed on targeting with a direct interaction the side chain of Asp144 in BD1 (Figure 5-17).

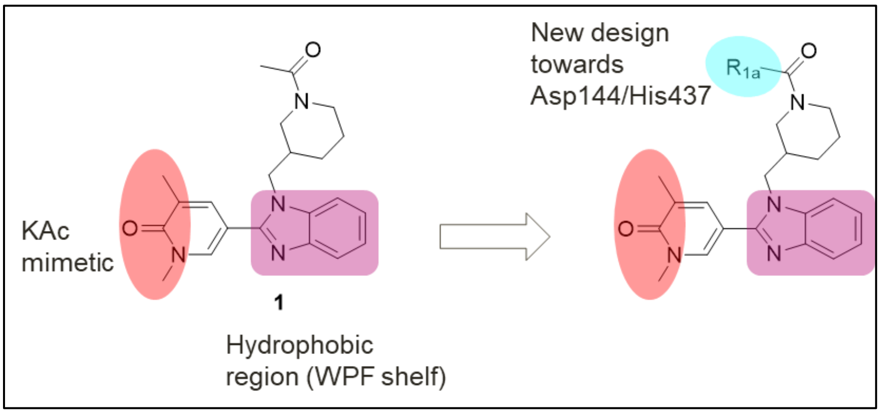
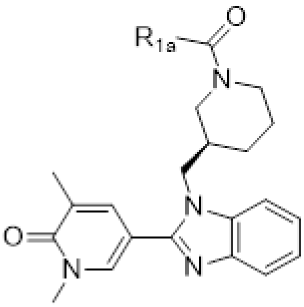
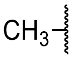
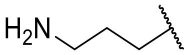
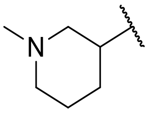
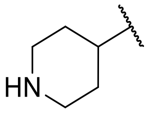
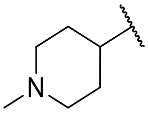
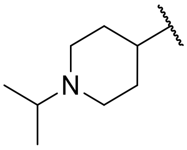
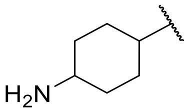
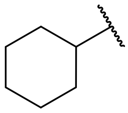


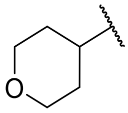
Figure 5-17 Schematic representation of the newly designed BRD4 inhibitors with BD1 over BD2 selectivity.

Several compounds were proposed with expansion from the acetyl group of the piperidine in R1, here referred to as R1a position. The design was done using as template the crystal structure of **1a**. In the R1a position were proposed substituents with the adequate length to reach Asp144 and a terminal positive charge to favour the formation of a hydrogen bond. The compounds were firstly docked into the BD1 and BD2 domains following the protocol as described in 5.5.2. The synthesis of those compounds was carried out by Erin Bradley, a PhD student who was also enrolled in the GSK/University of Strathclyde PhD-Programme. Given the initial SAR analysis, the desired modifications towards His437 were inserted into the lead compound **1**, as reported in Table 10.

Table 10 Optimization of the R1a position towards the Asp144/His437 domain difference.



R1a	pIC_{50} BD1/BD2	BD1 selectivity	Cmpd
	8.0/6.2	1.8	1
	7.8/5.6	2.2	13
	6.8/5.1	1.7	14
	7.5/4.9	2.6	15
	7.6/4.8	2.8	16
	7.8/4.8	3	17
	7.9/5.1	2.8	18
	6.8/4.9	1.9	19

	6.7/4.7	2	20

All of the proposed R1a substituents (Table 10) are directed towards the Asp144/His437 mutation present in BD1 and BD2, respectively. It was postulated that the compounds would show preference for the BD1 domain given the insertion of a positively charged group which would make a hydrogen bond interaction with Asp144. The insertion of the 3-carbon linker with a basic group (**13**) did not give the boost in potency one should expect for a direct interaction with Asp144 in BD1; this could be attributed to both the desolvation penalty of incorporating an explicit charge on the ligand and to the displacement of a stable water molecule (see infra). The flexible amine chain was then converted to a closed ring (**14-18**) because in BD1 there is enough space to accommodate such a transformation and at the same time a larger group should be less tolerated in BD2 as observed in the structural analysis of His437, see Figure 5-13. All of the compounds in Table 10 (**14, 19** and **20** excluded) are able to maintain BD1 potency and display a decrease in BD2 potency, which results in a higher selectivity towards the BD1 domain. The removal of the basic nitrogen in **19** and **20** or its placement in the meta position (**14**) shows a drop in BD1 potency, likely caused by the impossibility to form a hydrogen bond with Asp144; on the other hand, **14**, **19** and **20** show consistently low BD2 potency, suggesting that the positively charged group is not required for decreasing BD2 potency.

In conclusion, growing in the Asp144/His437 region has proven to be a successful strategy that led to the identification of compounds **16** and **17** which showed 1000x fold selectivity towards BRD4-BD1 satisfying the objective of this project but further modelling was needed to better understand the causes of the selectivity.

5.6.4 WaterMap simulations

The synthesised compounds in Table 10 showed increased selectivity towards BD1 with **16** and **17** satisfying the desired selectivity of about 3 log unit. It was noted that the boost in potency for BD1 was not obtained and such effect was investigated by studying the water molecules surrounding the Asp144/His437 region in more details.

From close inspection of the apo crystal structure of BD1 (PDB 2OSS) it was observed that Asp144 is hydrogen bonding with the NH backbone of Lys121 and also with Asn140 via a water bridged interaction. In the apo model of BD2 (PDB 2OUO) the His437 residue in the closed conformation is interacting with the NH backbone of the gatekeeper residue (Val439), and the water bridged interaction with Asn429 (Asn140 in BD1) is not observed. The difference in the water network around the Asp144/His437 could be crucial for understanding the BD1/BD2 potency, therefore, it was decided to computationally explore this further.

In order to study the thermodynamic properties of the water molecules in both BD1 and BD2 ligand binding site and explain the observed trend for the synthesised ligands, WaterMap calculations were run as described in section 5.5.5. The values calculated from a WaterMap simulation correspond to the average excess enthalpy, entropy and free energy that a water molecule, located at the hydration site, possess (see Section

2.4). The excess energies are measured relative to bulk water. In general; it is favourable to displace a water molecule that has a positive free energy (ΔG > 0).

From the WaterMap results of the apo form of the BD1 and BD2 domains, a water molecule is identified in proximity to the backbone-NH of the gatekeeper Ile146/Val439. Such a water molecule (water site *d* in Figure 5-18 and *c* in Figure 5-21) has a predicted free energy of 2.4 kcal/mol (BD1) and 2.8 kcal/mol (BD2). The enthalpic contribution for this water site is minimal (-0.3 kcal/mol and 0.1 kcal/mol) meaning that the entropic contributions are predominant and its release in the bulk is beneficial for binding. Therefore, its displacement could explain the boost in potency observed for the acetyl capped piperidine ring when firstly inserted in R1 (Table 1, compound **1**).

From the WaterMap calculations of the apo of BRD4-BD1 protein (2OSS), three water sites (*a*, *b* and *c*) are reported in the region where the R1 substituents are extending with predicted free energies of -1.3 kcal/mol, 1.2 kcal/mol and 0.2 kcal/mol (Figure 5-18). The negative enthalpic contribution of those water sites (Table 11) suggests that this is a rather polar portion of the ligand-binding site; such finding is in line with the residues in this location.

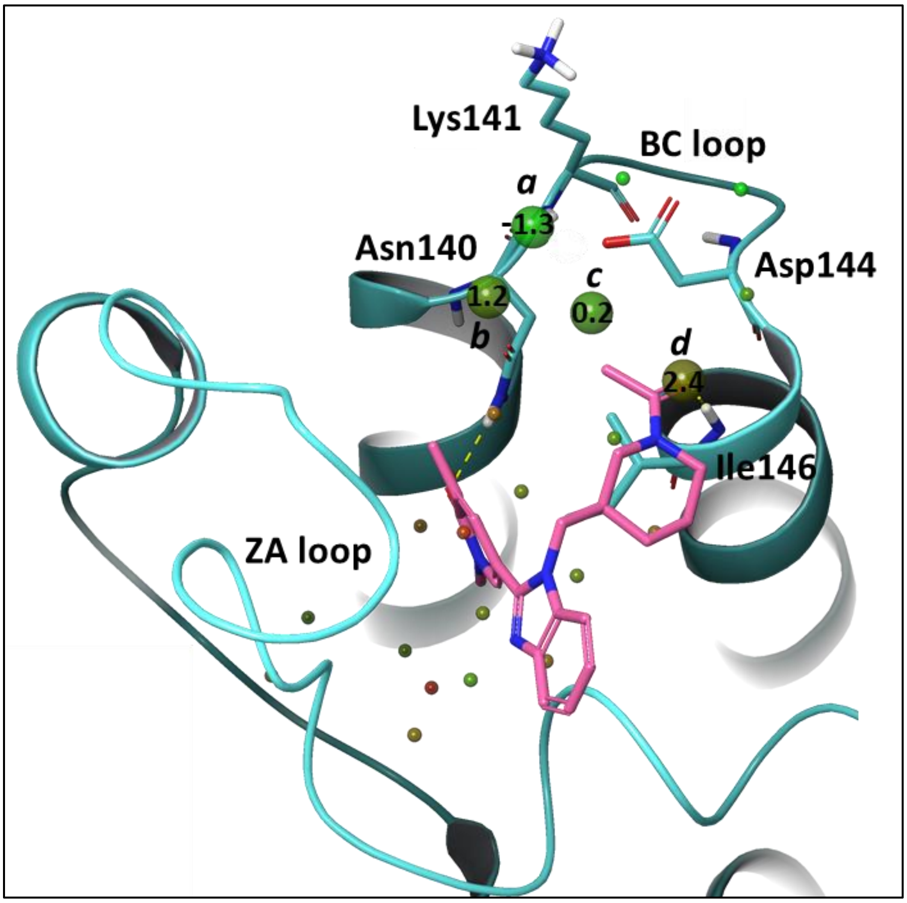


Figure 5-18 WaterMap of the apo-BD1 form (PDB: 2OSS), **1** is displayed only for reference. The water sites are reported in spheres, color coded by free energy (green= stable to red= unstable) and labelled by predicted free energy in kcal/mol. The water sites *a*, *b* and *c* are located in the region where R1a substituents are extending and water site *d* is close to the gatekeeper and displaced by the acetyl capped piperidine group.

Table 11 Predicted thermodynamics properties in term of occupancy, enthalpy (ΔH), entropy (-TΔS) and free energy (ΔG) of the water sites (a, b and c) in the apo BD1 domain.

Water site	*ΔH*	*-TΔS*	*ΔG*	*occupancy*
a	-3.7	2.4	-1.3	0.7
b	-3.0	3.2	0.2	0.8
c	0.4	0.8	1.2	0.4

The WaterMap simulations were also run on BRD4-BD1 in complex with **1**, **13** and **16** (Figure 5-19). Compared to the apo model, the water sites *c* and *d* are displaced in all the cases; water site *b* is displaced with **13** and **16** and in presence of **1**, the predicted free energy increases from 1.2 kcal/mol to 1.9 kcal/mol (Figure 5-19).

Interestingly, the water site *a* found in the apo form with a ΔG of -1.3 kcal/mol, is predicted to have a free energy of -3.9 kcal/mol in the presence of **1** (Figure 5-19). This suggests that when **1** is bound, water site *a* is even more stabilized. Moreover, in **1** an additional water molecule at 2.9 Å from the apo water site *c* is predicted to have a free energy of -0.8 kcal/mol (Figure 5-20). Therefore, compared to **1**, when **13** and **16** are present, the R1a groups are inserting in a region composed of stable water molecules whose displacement is unfavourable, especially in the case of water site *a* (Figure 5-20).

In summary, the stable water network in the BC loop of BD1 may explain why the boost in BD1 potency from **1** ($pIC_{50\ (BD1)}$ = 8) to **13** ($pIC_{50\ (BD1)}$ = 7.8) and **16** ($pIC_{50\ (BD1)}$ = 7.6) by adding a positively charged group able to interact with Asp144, is not observed.

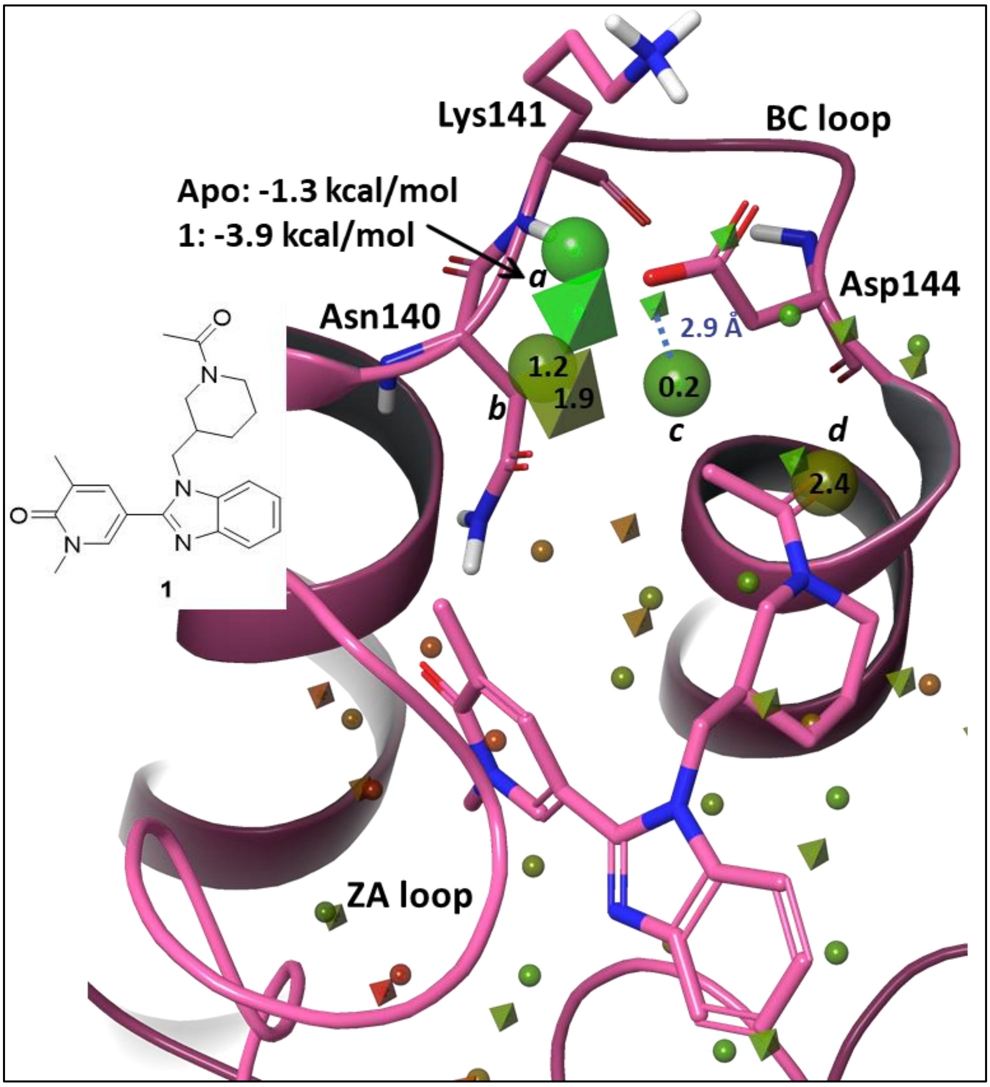


Figure 5-19 WaterMap water sites predictions of BRD4-BD1 in complex with **1** reported as spheres superposed to the one obtained for the apo form, reported in pyramidal shape. The numbers on top of the highlighted water molecules represent the value of the free energy (ΔG) in kcal/mol. Water site *d* and *c* are displaced in presence of **1**. Water site *a* is more stable in 1 than in the apo form. Note that the ZA loop is not displayed for clarity.

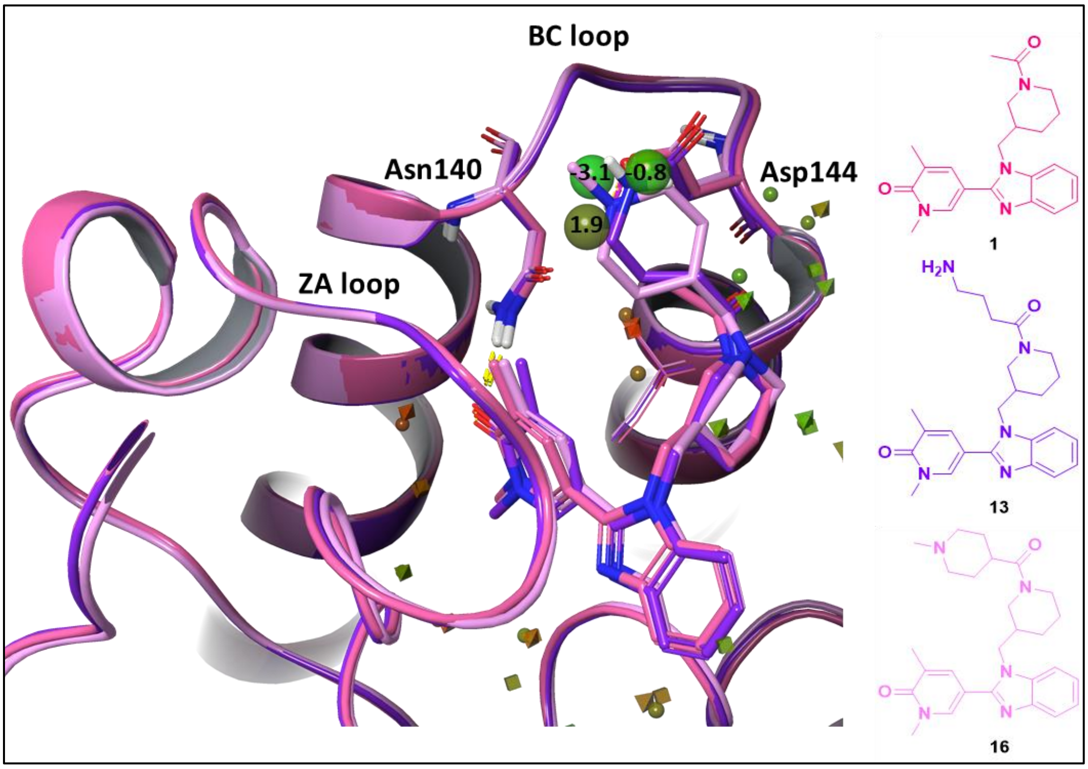


Figure 5-20 WaterMap water sites predictions of BRD4-BD1 in complex with **1**, **13** and **16**. The free energy in kcal/mol of the water site predicted in **1** is reported.

In the apo BD2 model, the His437 residue in a closed conformation is interacting directly with the backbone of the gatekeeper, Val439, and with a water molecule with predicted free energy of 0.6 kcal/mol (Figure 5-21). His437 was also modelled in the open conformation and from analysis of the WaterMap simulations, a water molecule was reported bridging the interaction between His437 and Asn429 which is occupying the space of the His437 when in the closed conformation (Figure 5-21, water site *b*). This water molecule has a predicted free energy of 0.6 kcal/mol, hence, its displacement, when **13** and **16** are bound, is not considered to be associated with an energy penalty. Such a finding is in line with the observation that the His437 can be identified in closed and open conformations but it does not explain the BD2 potency drop observed for **13** and **16** (Table 10).

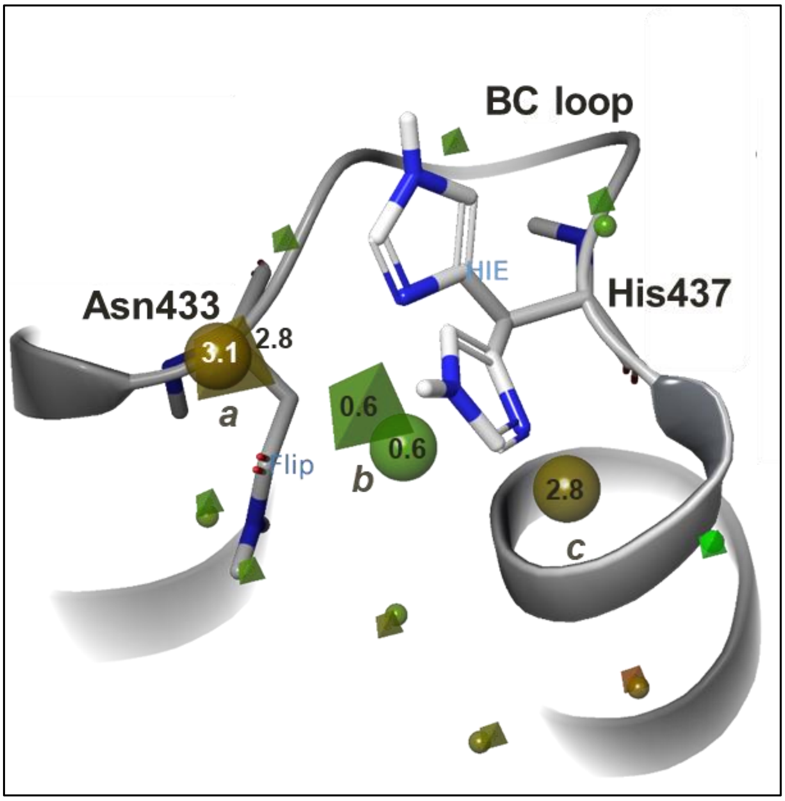


Figure 5-21 WaterMap of BD2-apo form (PDB: 2OUO) with His437 in open and closed conformation. The predicted free energy in kcal/mol is reported for the highlighted waters close to the His437. Spherical water sites refer to the His-open conformation, pyramidal water sites to the His-closed.

In both BD2 open and closed conformations, there is a water molecule in proximity to Asn429, water site *a* in Figure 5-21, which appears to be unstable with a ΔG of 3.1 kcal/mol (open conformation) and 2.8 kcal/mol (closed conformation). In the presence of **1**, this water molecule has a comparable free energy of 2.5 kcal/mol (Figure 5-22) suggesting that it should favour the ligand binding making the water molecule less unstable. The thermodynamic partitioning indicates that water site *a* is unstable in the apo form both for enthalpic and entropic reasons (Table 12) while in the case of **1**, it seems to be more entropically constrained.

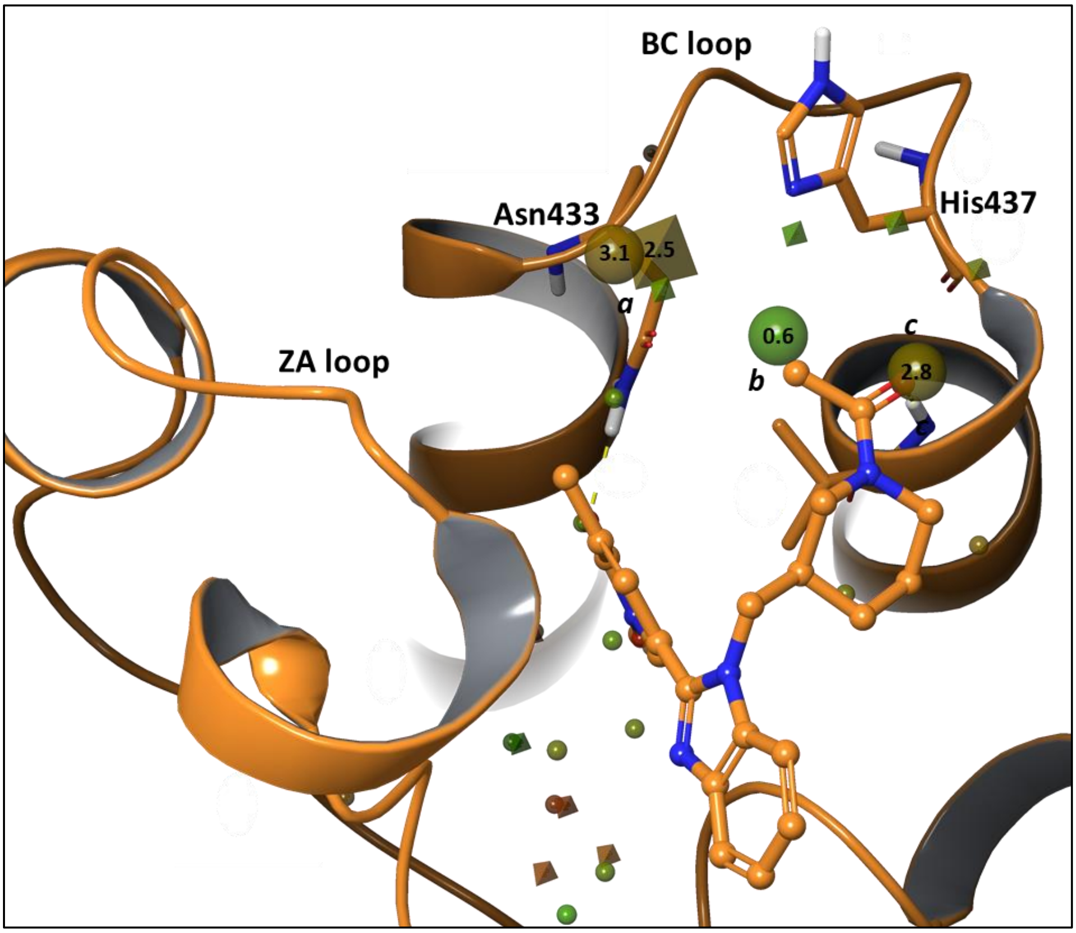


Figure 5-22 WaterMap water sites predictions of BRD4-BD2 of the apo form and in complex with 1. The free energy in kcal/mol of the water site is reported in spheres for the apo and in pyramidal shape for **1**.

Table 12 Predicted thermodynamics properties in term of enthalpy (ΔH), entropy (-TΔS), free energy (ΔG) and occupancy of the water sites (a, b and c) in the BD2 domain of apo form and in complex with **1**.

Water site	*Entry*	*ΔH*	*-TΔS*	*ΔG*	*occupancy*
a	Apo-His-open	1.9	1.2	3.1	0.5
a	1	-1.9	4.6	2.5	1
b	Apo-His-open	-0.9	1.5	0.6	0.5
c	Apo-His-open	0.1	2.7	2.8	0.9

The results from the WaterMap simulations of BD2 do not explain the loss in BD2 potency observed for **13** and **16**, on the contrary they seem to suggest that ligand binding in the region of His437 should be favourable. It appears evident that the BD1 and BD2 water network in the Asp144/His437 region is different; in fact, while in BD1 there are stable water sites (negative free energy), in BD2 all the water sites have positive free energies. This may indicate a potential limitation of the WaterMap calculations in describing the BD2 domain and/or that the reason for the BD2 potency loss is not mainly driven by the water network.

5.6.5 Molecular Dynamics simulations

To elucidate the drop on BD2 potency, MD simulations were run. The BRD4 BD1 and BRD2-BD2 proteins in complex with **13** and **16** were simulated in replica of 3 as explained in Section 5.5.3. The aim of running the MD simulations was to compare the dynamics of **13** and **16** in the BD1 and BD2 domains separately and in relation to each other.

Firstly, the effect of ligand binding on the overall dynamics of the BD1 and BD2 proteins was investigated. The Root Mean Square Fluctuation (RMSF) of the alpha carbon of the protein during the course of the simulations in the presence of **13**, **16** and without ligand were compared (Figure 5-23). In the case of BD2, the presence of the ligand (**13** or **16**) decreases the fluctuation of the ZA and BC loops; the same effect was observed in a recent MD-based study[181] in which the dynamics of Olinone, a BD1 selective ligand was investigated. On the contrary, the RMSF fluctuation in BD1 is not significantly affected by the presence of the ligand as compared to the apo protein

(Figure 5-24). In contrast, in the same work of Rodriguez *et al*; Olinone is also found to decrease the ZA loop movement in BD1. However, they suggest that the change in mobility of the ZA loop is larger in BD2 than in BD1 and that this could in part contribute to explain the higher selectivity for BD1.

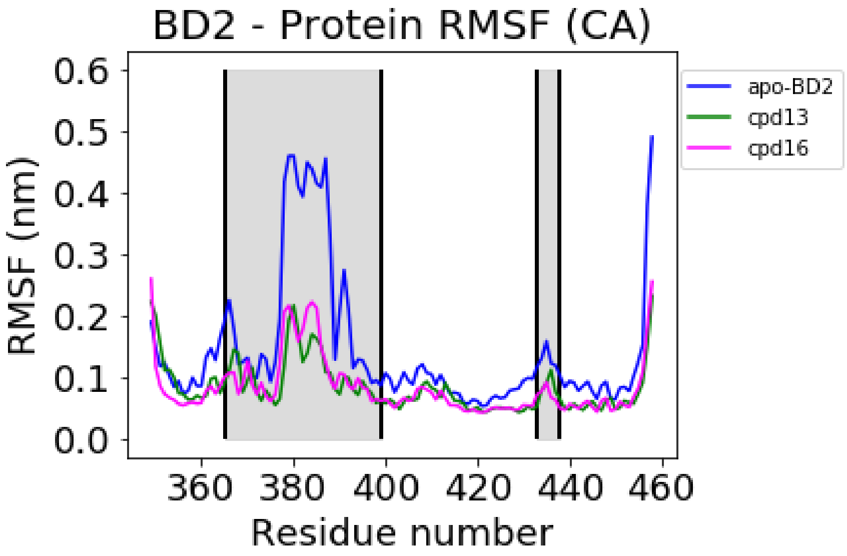


Figure 5-23 The RMSF fluctuation of the alpha carbon of the BRD4-BD2 protein without ligand (blue) and BRD2-BD2 with **13** (green) and **16** (magenta) is displayed. The areas in light-grey colour refer to the ZA and BC loop, respectively.

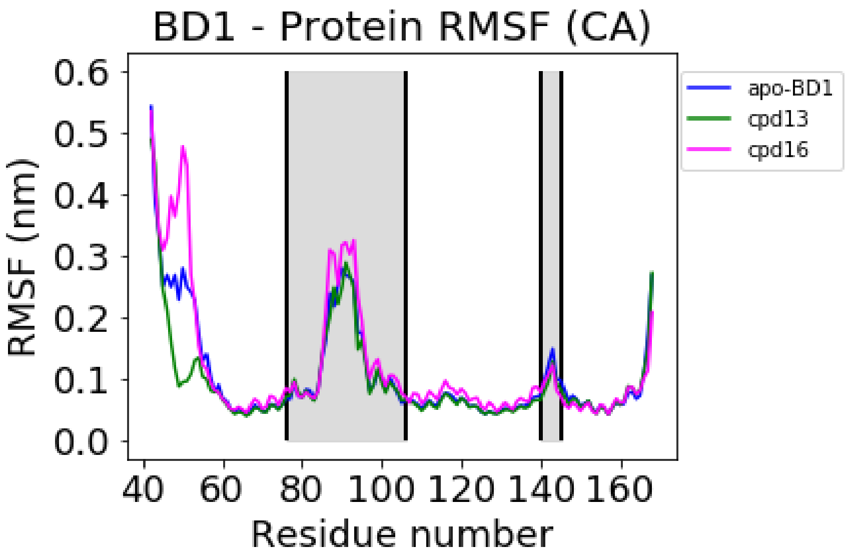


Figure 5-24 The RMSF fluctuation of the alpha carbon of the BRD4-BD1 protein without ligand (blue), with **13** (green) and **16** (magenta) is displayed. The areas in light-grey colour refer to the ZA and BC loop, respectively.

Next, a more detailed investigation into interactions between the ligand and the binding site was carried out. For all the simulations, two measurements were monitored:

1) The RMSD of ligand heavy atoms, using the starting conformation as reference.
2) Distance between the ligand (positively charged nitrogen) and Asp144 (oxygen) in BD1 or His437 (Nδ) in BD2.

In the following section the MD simulations of **13** and then of **16** will be discussed.

In BD1, **13** is found close to its docked geometry, occasionally adopting conformations that differ from the starting one but the overall interactions with Asn140 and Asp144 are maintained (Figure 5-25,a). Importantly, the distance distribution between **13** and Asp144 is maintained very close to 0.25 nm for the vast majority of the simulation (Figure 5-25,b), suggesting the presence of a key hydrogen bond stabilizing the ligand binding mode in BD1. The flexibility of the amine side chain is reflected in the presence of a long tail in the distance distribution stretching up to 1.2 nm. Conversely, in BD2 the amine chain of compound **13** is highly mobile (Figure 5-25,c), the peak at 0.25 nm is absent showing that the hydrogen bond is not formed between the amine and the His437. This suggest that the interaction between **13** and the histidine is not key for the binding in the BD2 protein.

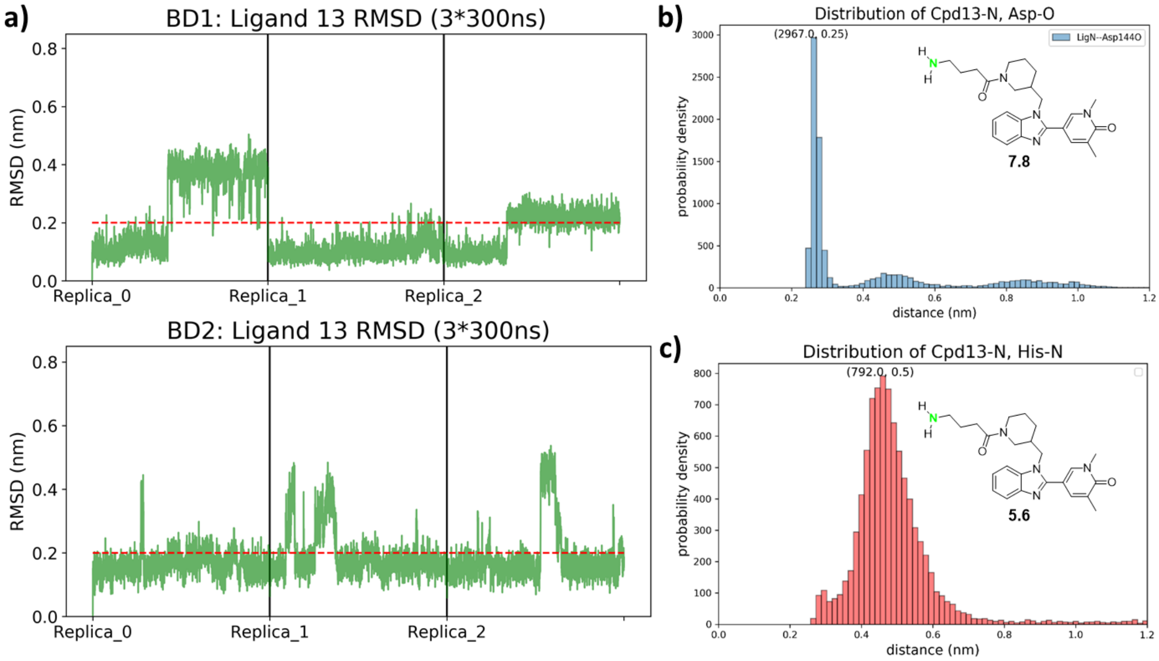


Figure 5-25 Measurements monitored during the 3*300 ns replicas of **13** in the BRD4-BD1 and BD2 domains. a) RMSD of the ligand heavy atoms in nm; b) density distribution between the positively charged nitrogen of the ligand, highlighted in green in the 2D representation, and the negatively charged oxygen in Asp144 (BD1); c) density distribution between the positively charged nitrogen of the ligand, highlighted in green, and the Nδ in His437 (BD2). The pIC50 in BD1 (b) and BD2 (c) is reported below the 2D-representation of the ligand.

In BD1, **16** shows a stable RMSD (Figure 5-26, a). In Replica_0, a RMSD increase at around 0.3 nm is observed which is maintained until the end of the simulation. In this case, the ligand is found to be less deep in the ligand-binding pocket but the interactions with Asn140 and Asp144 are maintained. The distance distribution between **16** and Asp144 is comparable to the behaviour observed in the case of **13** but there are some differences: the peak is broader, and at a longer distance and it has less of an extended tail suggesting that the piperidine ring has less ability to move within the pocket (Figure 5-26, b). In BD2, the ligand RMSD of **16** is stable and conformations close to the starting position are mainly explored (Figure 5-26, a). The distance distribution from the histidine residue is comparable to the one observed for the same ligand in BD1 with a maintained peak at 0.3 nm (Figure 5-26, c). The piperidine ring is more rigid in comparison to **13** and probably less able to move due to steric restrictions in the pocket. By analogy to compound **13** it can be assumed that the hydrogen bond between **16** and His437 is unlikely to be key to the binding mode in BD2 and therefore is probably observed as a consequence of the ligand being sterically restrained in close proximity to the histidine.

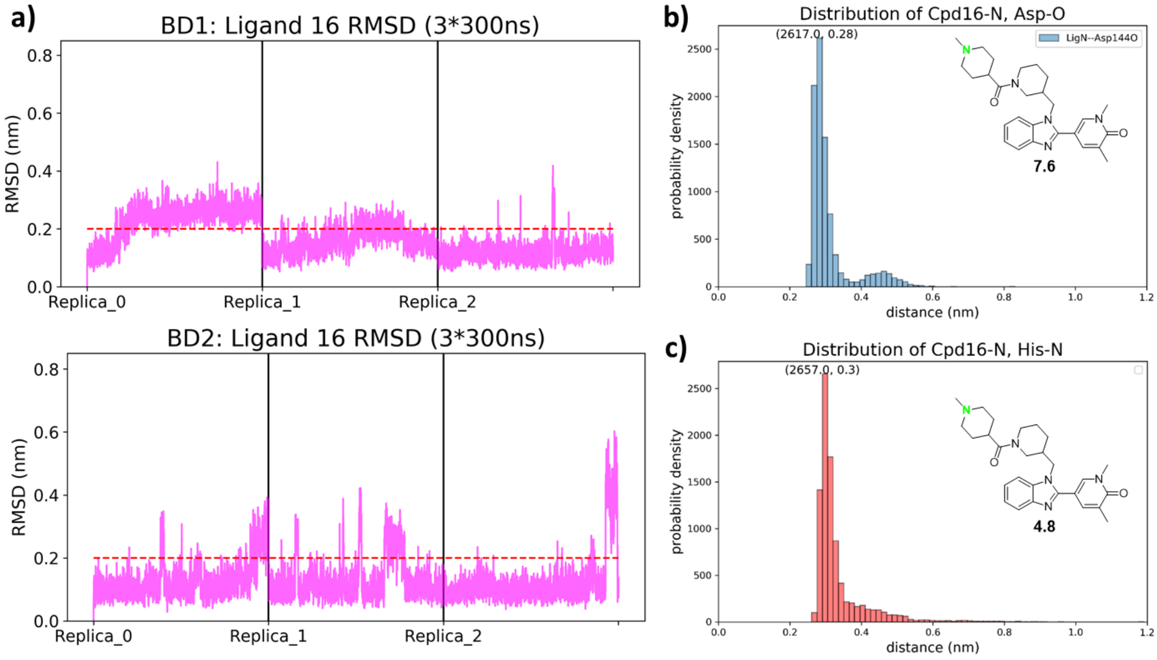


Figure 5-26 Measurements monitored during the 3*300 ns replicas of **16** in the BRD4-BD1 and BD2 domains. a) RMSD of the ligand heavy atom in nm; b) density distribution between the positively charged nitrogen of the ligand, highlighted in green in the 2D representation, and the negatively charged oxygen in Asp144; c) density distribution between the positively charged nitrogen of the ligand, highlighted in green, and the Nδ in His437. The pIC50 in BD1 (b) and BD2 (c) is reported below the 2D-rapresentation of the ligand.

The ligand Root Mean Square Fluctuation (RMSF) reported in Figure 5-27 shows that in general the amine chain in **13** is more flexible than the piperidine ring in **16**. The high flexibility of the amine side chain in **13** initially seems to contradict the sharp peak seen for the amine and the Asp144 interaction in Figure 5-25, b. However, closer investigation reveals that the amine and the acid move in conjunction maintaining the interaction (Figure 5-25). Furthermore, the observation that the same amine is less mobile in the BD2 protein supports the hypothesis that the BD2 pocket is more sterically restricted. In both BD1 and BD2 the piperidine ring of **16** is rigid (Figure 5-27) and fixed in a conformation which seem to favour the formation of an interaction with Asp144/His437. This supports the theory that the hydrogen bond in BD2 is a consequence of the enforced close proximity rather than a key driving force behind the binding mode. In fact, in **16** the piperidine ring is either pointing towards His437 or, at the end of the simulations, moving towards the WPF shelf suggesting both a poor shape complementarity between the ligand and the ligand binding site of the protein and the incapability to adjust in the binding site. Interestingly, this is also consistent with the fact that a crystal structure of the BD2 protein bound to **16** was requested but could not be obtained.

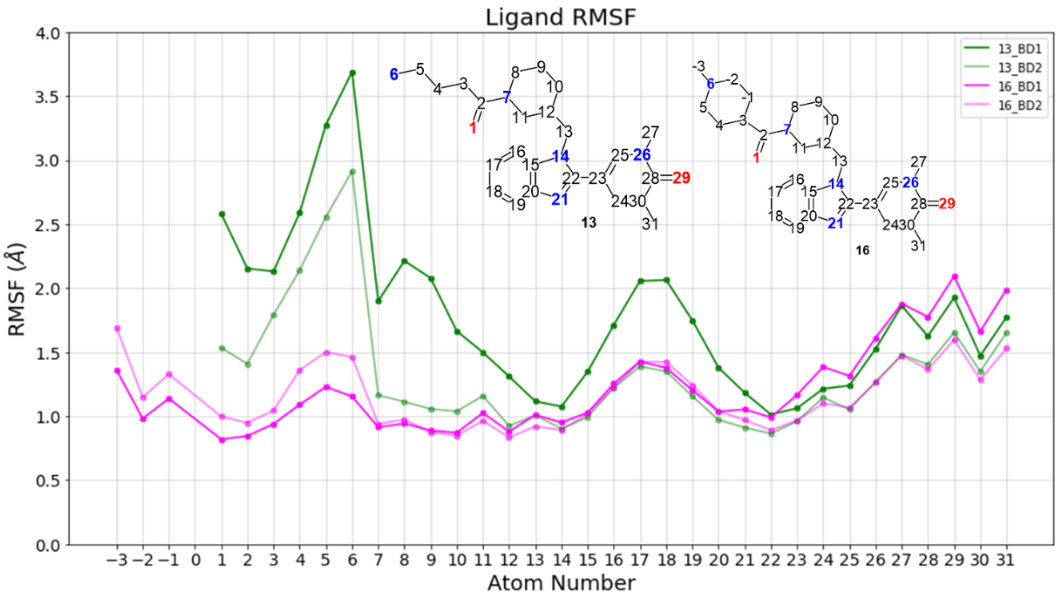


Figure 5-27 Ligand RMSF for 13 and 16 during the course of the MD simulations. The RMSF is broken down by atom number which is reported on the ligand 2D structure.

Overall, MD simulations have been able to provide some hypothesis for the reduced binding affinity of compound **16** to the BD2 protein but have not been able to produce a completely compelling evidence. In a prospective situation, it is unlikely that **16** could have been identified as a weak BD2 binder. This is one of the reasons why the routine application of MD simulation to drug discovery projects remains limited.

5.7 Discussion

The goal to obtain a probe for the BET-BD1 protein was achieved with both compounds **16** and **17**. The BRD4-BD1 potency is not increasing significantly with the addition of the R1a substituents due to the presence of stable water molecules (negative free energy from WaterMap simulations) close to Asp144. At the same time, a decrease in BD2 potency was consistently observed in all the newly synthesised compounds (Table 10).

Compounds **13** and **16** were the subjects of MD studies with the aim to elucidate the potency loss observed for the BD2 domain. The reasons why both **13** and **16** are poor BD2 binders are likely to originate from subtle effects which are challenging to isolate. In fact, there are a lot of aspects linked to the ligand binding events that need to be considered when designing or optimizing compounds, e.g. protein and/or ligand flexibility, solvation and desolvation penalties, the enthalpic/entropic effects[182]. Trying to decompose the contributions that are responsible for small changes in potency and selectivity is computationally difficult to address in a reliable way, especially in this case where the potency window is restricted to a small range.

The ring closure of a flexible side chain has been used as a strategy to improve the ligand binding affinity by pre-organizing the ligand and preventing any entropy loss during the binding event[183]. On the other hand, this could also decrease or simply maintain the binding affinity if, for example the ligand is locked in a conformation which is different from the binding one. It is also common to attempt to make optimal interactions (electrostatic, H-bonding, etc.) with the protein to increase the enthalpic contribution. However, it is often observed that adding functional groups does not necessarily improve potency. The difficulties in increasing potency are often attributed to the enthalpy-entropy compensation[184, 185] which refers to the phenomenon that the entropy losses in response to the enthalpic change often result in no affinity gains. For example, structuring of the protein-ligand complex with the formation of interactions will be associated with loss in conformational entropy, overexposure of non-polar groups will result in a loss in solvation entropy.

Another crucial aspect to consider in the ligand binding event is the displacement of water molecules. The release of a water molecule from a protein cavity can be enthalpically favourable/unfavourable or entropically favourable/unfavourable. It is very difficult to predict which category a specific water molecule will fall into because it can be dependent on the interactions with the protein, ligand and other water molecules, as well as on the dynamics of the protein and of the molecules that were previously surrounding both the ligand and the protein. Assuming that the displacement of water molecules will always be favourable is too reductionist. The field of water modelling has progressed significantly in the last decade but being able to quantify the energetic contribution of water molecules to design compounds still seems to be a major challenge[186].

Lastly, the change in conformation and electrostatic perturbations induced by ligand binding may result in changes in ionization states and therefore, it is important to consider how the surrounding environment could affect both ligand and protein protomeric states when calculating the binding free energy. Preliminary simulations were performed with His437 in the HIE, HID and HIP protonation states (Appendix A). The ligand has a different behaviour in the presence of the different protonation states. It was not possible to identify the state that clearly dominates but the HIE protonation state was selected to be consistent with the other simulations and because the model appears to be more stable.

5.8 Conclusion

A BD1-selective BET inhibitor with 1000-fold selectivity was designed starting from the benzimidazole series. The exploration of the SAR combined with an X-ray crystal structure of compound **1a**, allowed the identification of the R1 position, which extends towards His437/Asp144, as the most promising vector to obtain selectivity.

Metadynamics simulations showed that His437 can adopt multiple conformations in the apo form and that the closed and open forms are equally observable (both are in a global minimum). WaterMap simulations revealed the presence of a water molecule located close to the gatekeeper residue which was predicted to be unstable, therefore, its displacement with the presence of the correct group, e.g. acetyl-capped piperidine, (**1**), could explain why growing the molecules in this region is providing a boost in potency.

Several compounds were proposed with the intention of interacting with Asp144 in BD1 with a hydrogen bond and at the same time to clash with His437 in BD2. The strategy was successful in boosting the BD1 selectivity to the target level. The increased selectivity was achieved primarily by decreasing BD2 affinity whilst maintaining BD1 potency.

The WaterMap simulations in BD1, showed that the water network surrounding the Asp144 is stable. Therefore, this could explain why the designed compounds do not display a boost in potency from the addition of a positively charged group that should form a hydrogen bond with Asp144.

The MD simulations provided insights on the selectivity of **13**. The drop in BD2 potency could be attributed to the fact that while in BD1 a stable hydrogen bond is formed with Asp144, in BD2 the same interaction is not formed with His437. MD simulations of **16** showed the formation of a stable hydrogen bond with Asp144 in BD1 and with His437 in BD2. The flexibility of **13** in BD2 shows that the ligand preference is to not interact with the histidine. In **16** the basic group is rigid and therefore, is constrained in proximity to the histidine, which appears to be unfavourable, given the observation with **13**. This is consistent with the weaker binding of **16** to BD2.

It is important to note that the ranges of BD1 and BD2 potencies are limited and consequently the range of selectivity is also narrow (8- 6.8 for BD1 and 6.2 to 4.8 for BD2). This small range of data combined with the fact that a number of subtle factors are likely to be combining to achieve selectivity mean it has not been possible to identify a specific cause of the achieved selectivity. A series of computational experiments could not conclusively identify an obvious reason for the loss of potency caused by the extension of the R1 substituent.

5.9 Future Work

It would be interesting to use the Free Energy Perturbation (FEP) method to investigate whether the range in pIC_{50} would be correctly predicted. In FEP all of the effects that comes into play when the ligand and protein bind together are taken into account. As it is believed that the potency for BD1 is mitigated by ligand desolvation and water displacement and the potency drop for BD2 is instead due to a combination of multiple factors including steric, desolvation, protonation states, with the usage of more advanced simulation techniques it may be possible to better predict potency and selectivity.

Chapter 6 Investigating the dual binding mode of a BRD4-BD1 ligand

6.1 Overview

In a SBDD project generating an X-ray crystal structure of the protein-ligand complex of interest is one of the most critical steps. Once a crystal structure is obtained, the key interactions between the ligand and its target can be determined and used to propose ligands modifications to further optimize the interactions with the protein by assuming that the binding mode is retained. This strategy has been proven successful for several cases, for example for the HIV protease[187] (Amprenavir), neuraminidase[188] (Relenza), thymidylate synthase[3] (Raltitrexed) and Abl tyrosine kinase[189] (Gleevec).

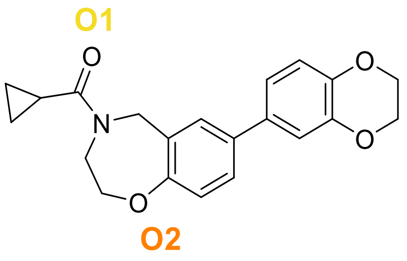
In general, a single crystal structure is obtained of a given protein-ligand complex and the binding pose identified is assumed to be the one contributing to the measured binding affinity. However, in principle more than one energetically similar pose could exist for a given protein-ligand complex. In some instances[190, 191], X-ray crystallographic screening has allowed the discovery of ligands with multiple binding modes and the clarification of the initially inconclusive SAR. How often such cases arise is unclear, primarily because multiple structures of the same complex are rarely obtained. In a recent GSK drug discovery program, the same protein-ligand complex was solved multiple times revealing two distinct binding modes depending on the specific conditions under which the structure was generated. This raises the question of whether computational methods could be used to predict the presence (or absence) of additional binding modes from a crystal structure of a protein ligand complex.

The first part of this chapter is devoted to structural analysis of these crystallographic data to elucidate the main difference between the two binding modes, called BM1 and BM2, which are then investigated using molecular simulation.

As well as being interesting in their own right, the crystal structures provide a stringent test for molecular simulation methods. Being able to reliably identify secondary binding modes using computational methods would be invaluable in SBDD projects, especially when experimental evidence is limited.

In the second part of this chapter, experiments are reported to assess whether computational chemistry methods could have identified: (i) BM1 and BM2 in the absence of data about either complex but with previous knowledge regarding other ligands interacting with BRD4 proteins; (ii) BM1 from knowledge of BM2; (iii) BM2 from knowledge of BM1.

The GSK compound referred to as **Ligand A** (Figure 6-1), showed a pIC_{50} equal to 4.9 in a Fluorescence Polarization assay developed for BRD4-BD1 and was co-crystallized with BRD4-BD1 under different crystallisation conditions. In the following section, BRD4-BD1 will be referred to as BRD4.



Ligand A

Figure 6-1 BRD4 co-crystallized **Ligand A** that shows a dual binding mode with BRD4 pIC_{50}= 4.9 measured by Fluorescence Polarization (FP) assay. The oxygen atoms, O1 and O2, involved with the hydrogen bond with Asn140 O1 are highlighted.

Structural analysis of the obtained crystal structures indicated the presence of two different binding modes of **Ligand A** (Figure 6-2), here called Binding Mode 1 (BM1) and Binding Mode 2 (BM2). In particular, out of the 11 crystal structures, 9 showed the **Ligand A** with BM2 and only 2 presented BM1.

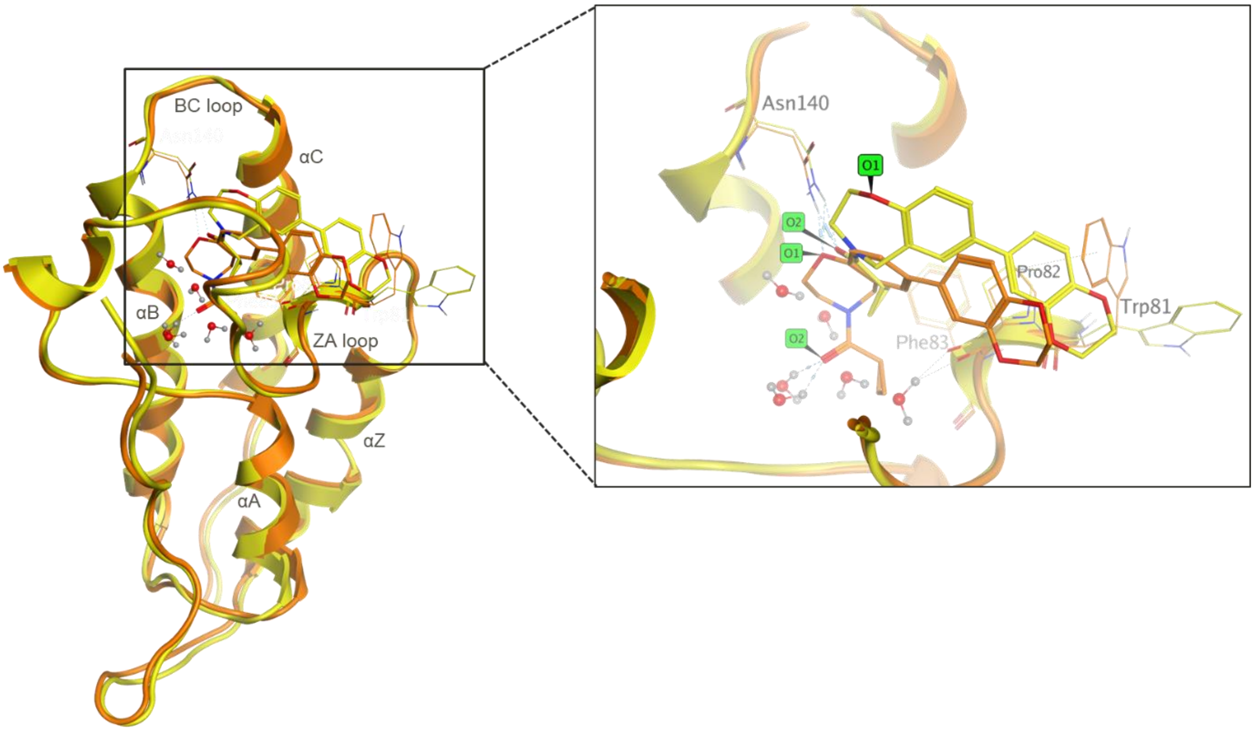


Figure 6-2 Ribbon representation of BRD4 with **Ligand A** in BM1 (yellow) and BM2 (orange). The ligand binding site is reported in the box on the right. The hydrogen bond with Asn140 is displayed with dashed line. The residues in the lipophilic shelf (Trp81, Pro81 and Phe83) are reported in stick as well as the water network at the bottom of the pocket. The oxygen atoms in the ligand are labelled as O1 and O2.

The heavy atom ligand RMSD between BM1 and BM2 is 3.8 Å. In BM1, **Ligand A** establishes the "classical" hydrogen bond interactions seen in BRDs with the side chain amide of the residue Asn140 located in the BC loop, through the oxygen of the ligand amide-like bond, called O1, and with Tyr97 through a water bridge, while the network of four water molecules at the bottom of the pocket is preserved (Figure 6-2). Moreover, in order to accommodate the ligand BM1, the residue Trp81 belonging to the lipophilic shelf assumes a conformation that differs from the usual conformation seen in many other BRD4 crystal structures (this is discussed in more detail later). In BM2, **Ligand A** displaces the water network and binds deeper in the pocket so that the oxygen of the benzoxazepine ring, called O2, is now forming a hydrogen bond with the residue Asn140 (Figure 6-2).

The two distinct binding modes obtained for **Ligand A**, BM1 and BM2, make these complexes an interesting system worth studying in more detail with computational tools. In particular, in order to investigate the different binding modes, and to assess whether interconversion occurs, several methods have been used including MD and MetaD simulations and a combination of Induced-fit docking with BPMD.

6.2 Methods

6.2.1 Schrodinger Software

6.2.1.1 System set up in Maestro

The BM1 and BM2 complexes were prepared with the Protein Preparation Wizard[142] tool in Maestro[143] using default options. The ligand protonation state was determined with Epik program at pH =7 ± 2.0. Hydrogens were added, bond orders were assigned to the ligand, the protein termini were capped, and the crystallographic water molecules were removed apart from the W1-W4 at the bottom of BD1 and W5 and W6 in BM2. The hydrogen bonding network of the receptor was then optimized by reorienting hydroxyl and thiol groups, amide groups in Asn and Gln, and selecting appropriate states and orientation of the His imidazole group (ProtAssign algorithm). Then, a restrained minimization in which the hydrogen atoms are allowed to freely move and the heavy atoms restrained to move by a maximum of 0.3 Å with the Impref module of Impact and the OPLS3e force field was performed. The complexes obtained from this set up were then used for: MD, MetaD, IFD and WaterMap simulations.

6.2.1.2 Molecular Dynamics

MD simulations were carried out using Desmond[146] v.5.6 and the OPLS3e force field. Once the system was fully prepared (6.2.1.1), the System Builder tool was used to create the box of TIP3P water with a truncated octahedral shape. Ions (Na^+ and Cl^-) were added to neutralize the system and reach a concentration of 0.15 M. At this stage, the system was relaxed and submitted to the MD protocol which consists of several stages. The system was relaxed at 10 K by 100 ps Brownian dynamics NVT simulation with restraints on the solute heavy atoms (50 kcal/mol/Å^2). The solvent box was then equilibrated at 10 K by 12 ps of NVT simulation and 12 ps of NPT simulation with restraints on the solute heavy atoms (50 kcal/mol/Å^2). The system was then heated to 300 K and full equilibration was performed in the NPT ensemble for 12 ps with restraints on the solute heavy atoms (50 kcal/mol/Å^2). The last step of the relaxation procedure was a 24 ps NPT dynamics run at 300 K without restraints. Finally, the production simulation was carried out in the NPT ensemble for a length of 1μs. A constant temperature of 300K was maintained using the Nosé-Hoover thermostat algorithm and Martyna-Tobias-Klein barostat algorithm to maintain a pressure of 1 atm. The short-range coulombic interactions were analysed using the short-range method with a cut-off value of 9 Å. The long-range electrostatic interactions were calculated using the particle mesh Ewald method. The nonbonded forces were calculated using a RESPA integrator where the short-range forces were updated every 2fs and the long-range forces were updated every 6 fs. The trajectories were saved every 500 ps for analysis, for a total of 2000 frames for each replica. Trajectories were additionally processed in VMD 1.9.2[180], the protein was centred in the water box by using the pbc tool and the trajectories were aligned on the protein backbone heavy atoms using the RMSD Trajectory tool. The trajectories were loaded in a custom Jupyter notebook and further analysed with MDtraj and Pytraj packages.

6.2.1.3 Clustering of the MD trajectories

The 10000 (2000 * 5) frames obtained from the MD simulations were clustered with the Daura[192] clustering algorithm present in the plugin implemented in VMD 1.9.2. The algorithm is a simplification of the Quality Threshold (QT) algorithm. In summary, a threshold is specified and a candidate cluster C1 is formed with the first frame of the trajectory as the seed. All frames having a RMSD value less than or equal to the similarity threshold when compared to the seed are added to C1. The process is repeated for all the frames in the data set. The frames contained in previously identified candidate clusters are available for future candidate clusters. At the end, there will be as many candidate clusters as number of frames in the data set and many overlaps. The biggest candidate is then selected, and all its members marked as a cluster and removed from further consideration. The process is completed when all the frames are either assigned to a particular cluster or marked as unclustered.

6.2.1.4 Induced-fit Docking

The **Ligand A** was extracted from the crystal structure and prepared with the LigPrep[178] tool using the default settings. The apo crystal structure of BRD4-BD1 with PDB: 2OSS was used as reference receptor to perform docking calculations of the ligand and prepared following the default setup using the Protein Preparation

Wizard[142] in Maestro[143] v.2018.04. From the Induced-Fit Docking panel, the Extended Sampling protocol was selected which consist in several steps:

1) Initial Glide docking with a soft potential and removal of side chains. Representative poses are generated.
2) Prime side-chain prediction for each protein-ligand complex, followed by minimization.
3) Glide re-docking of each protein-ligand complex structure.
4) Scoring of each pose with the IFDScore.

It is important to note that in order to predict poses similar to BM1, the **Ligand A** was superposed to the apo crystal structure and used as reference to build the docking grid from the Receptor tab. In general, the default parameters were used. A grid length of 10 Å was chosen as measure of the ligands to be docked; no constraints between protein and ligands were used. From the Ligands tab, the ligand ring conformational sampling was selected with an energy window of 2.5 kcal/mol. Finally, in the Prime Refinement tab, the residues within 5 Å of a ligand pose were selected to be refined. The OPLS3e force field was used and the binding energy of the final poses was estimated with the IFDScore.

6.2.1.5 BPMD

The complexes were prepared using the Preparation Wizard Tool. The prepared systems were then submitted to the same procedure as described in the Section 4.2.4 relative to Binding Pose Metadynamics (BPMD).

6.2.1.6 Metadynamics

The prepared apo form of BRD4 was submitted to the metadynamics simulation using Desmond[146] software v 5.6 with the GPU implementation and the OPLS3e force field. The CVs used in this calculation were the χ_1 and χ_2 torsional angle of the Trp81 as defined in Section 6.3.4. The default parameters were used, gaussians were deposited every 0.09 ps with a hill height of 0.03 kcal/mol and width of 5.0°. The convergence of the MetaD simulation was checked by comparing the FES at different simulation times. The error associated with the estimation of the FES was about 0.6 kcal/mol.

6.2.2 Funnel MetaD

6.2.2.1 System Setup

The complexes were solvated with TIP3P water, the net charge was neutralized by adding ions and then a concentration of 0.15 M was reached with both Na+ and Cl-. Ligands were parametrized with the Generalized Amber Force Field (GAFF) and charges were computed with the AM1-BCC semiempirical method as implemented in Antechamber in Ambertools16[120]. The protein, ions and water molecules were parametrized with the AMBER14SB force field. Simulations were run using GROMACS[193] version 5.1.4 with the PLUMED[194] plugin version 2.4.1 with a 2-fs integration step in the NPT ensemble.

6.2.2.2 Equilibration and Production

The energy of the simulated system was minimized using a steepest descent integrator until the maximum force exerted on the atoms dropped below 1000 kJ/mol/nm. The system was heated from 100 K to 300 K for 1 ns applying positional restraints on the α carbons of the protein at constant volume using a using a velocity rescale thermostat.

The system was equilibrated without restraints in the NVT ensemble for 1 ns and then for 2 additional ns in the NPT ensemble. The frame with the density closest to the average density was then used for the production run in the NVT ensemble.

The production was generated using GROMACS simulation package with the PLUMED 2.4.1 plugin with a 2-fs integration step. Constant volume conditions were employed with the velocity rescale thermostat. Bond lengths were constrained using LINCS, while van Der Waals interactions were treated with a cutoff of 10 Å. Electrostatics interactions were computed using the PME method with the direct sum cutoff of 10 Å and Fourier spacing of 1.2 Å.

6.2.2.3 MetaD

The metadynamics simulations were run using a parallel tempering scheme in the 300–310 K temperature range.; in the production run exchanges were attempted every 1000 steps.

During the production metadynamics runs Gaussian hills were deposited every 2 ps in the well-tempered scheme with a bias factor of 8. The Gaussian width was set to 0.05 nm for the Z-projection and XY-projection and the initial height was 1 kJ/mol.

The production runs of PT-metaD in the well-tempered ensemble were performed for each system consisting of 6 replicas; in this way, a Gaussian potential was deposited in the collective variable space every 2 ps with the height $W = W_0 e^{-V(s,t)/(f-1)T}$ where W_0= 1 kJ/mol, *V(s,t)* is the bias potential at time *t* and CV value *s*, the bias factor *f* was set to 8 and *T* is the temperature of the replica.

The collective variable used in the study are the pocket-ligand distance vector projected on the Z-axis and XY plane of the funnel. The pocket-ligand distance vector was calculated by using the distance between the nitrogen N1 of the **Ligand A** and the center of mass between of the Cα of Gly108, Phe133, Met132 and Leu153, located in the αA, αB and αC loop at the bottom of the pocket (Figure 6-3). The ligand-binding site is positioned perpendicular to the XY-plane such that the opening of the binding site is parallel to the Z-axis.

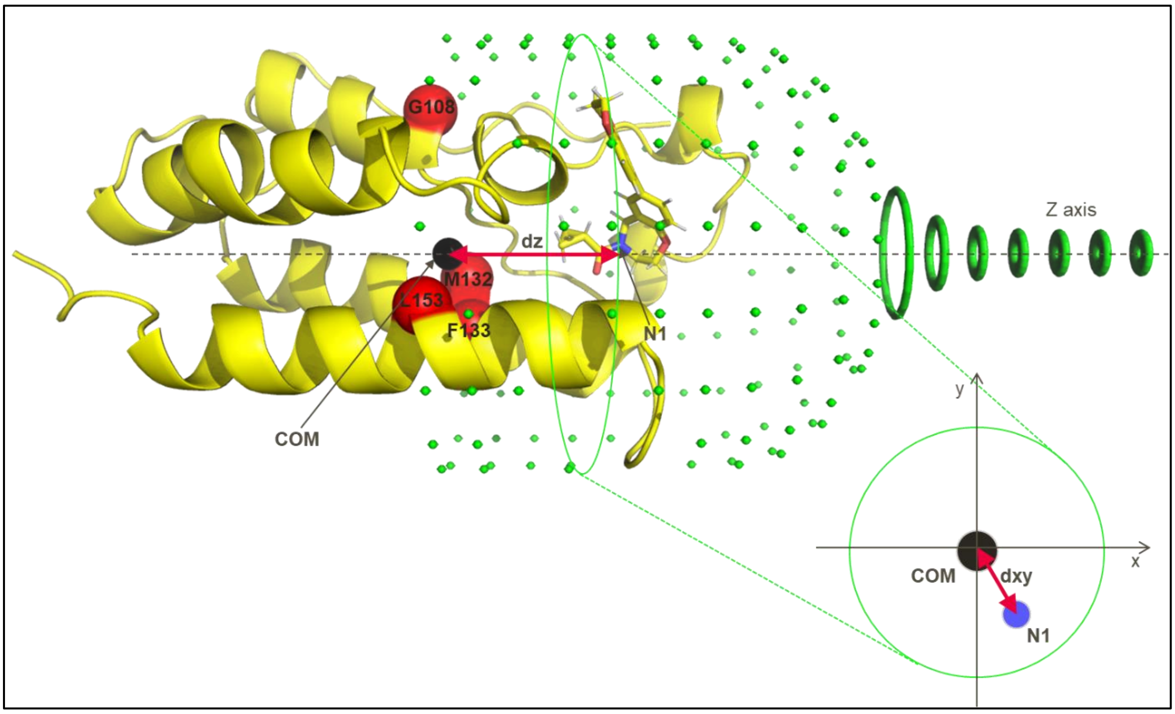


Figure 6-3 Funnel restraint potential applied to BM1 in green. The Cα atoms of the residue Gly108, Met132, Phe133 and Lys153, reported in red spheres, were used to calculate the protein centre of mass (COM). The projection of the vector between the N1 atom of the ligand and the COM on the Z-axis (dz) of the funnel and on the XY-plane (dxy) were used as CVs.

The funnel-like restraint was constructed with a sigmoid function in the space of the XY-projection CV:

$$r = \frac{h}{1 + e^{s*(z - z_0)}} + b$$

Where *r* is the dxy CV, *h* is the funnel width which was set to 1.2 nm, *s*= 5 nm^{-1} controls the steepness of the function, *z* is the value of the dz CV, z_0= 2.5 nm is the inflection point and b= 0.12 nm is the minimum width. A quadratic repulsive potential was applied at the boundaries of the funnel.

Using the FM, the ligand-protein binding constant, K_b, can be computed as follows accordingly to ref[195]:

$$K_b = \pi R_{cyl}^2 \int_{site} \mathrm{d}z \, \mathrm{e}^{-\beta[W(z) - W(ref)]}$$

Where R^2_{cyl} is the surface of the funnel used as restraint potential; the potential for the bound W_z and unbound state, W_{ref}, are obtained through metadynamics calculation. β is constant and equal to $1/(k_B T)$, k_B is the Boltzmann constant and T is the temperature of the system.

The equilibrium binding constant, K_b, is directly related to the absolute protein-ligand binding free energy, ΔG_b^0, through this relation:

$$\Delta G_0^b = -\frac{1}{\beta} ln(C^0 K_b)$$

Where C_0 is the standard concentration at 1M and is equal to 1/1600 Å^{-3}.

6.3 Analysis of the crystallographic data

In the following sections an overview of the crystallographic experiments (6.3.1), the BM1 (6.3.2) and BM2 (6.3.3) crystal structures and their differences (6.3.4), and the comparison with the available crystal structures of BRD4 (6.3.5) are discussed.

6.3.1 Crystallographic experiments

A total of 11 co-crystal structures of **Ligand A** in BRD4 were obtained with different crystallization conditions, as reported in Table 13.

Table 13 Crystallization conditions used to obtain the co-crystals of **Ligand A** with BRD4 protein. Abbreviations: PEG= Polyethylene Glycol; MES= 2-(N-morpholino)ethanesulfonic acid; Tris= tris(hydroxymethyl)aminomethane; MMT= Malic acid, MES and Tris; HEPES= 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid; n/a= not available.

MODE	*Crystal symmetry*	*Resolution (Å)*	*Conditions*
BM1	P1	1.73	20% PEG6000, 0.1M MES 0.2M $MgCl_2$ pH 6
BM1	P1	1.72	20.00% PEG3350, 0.20M NaK phosphate
BM2	$P2_12_12_1$	1.5	20% PEG6000, 0.1M Tris-HCl pH 8
BM2	$P2_12_12_1$	1.85	25% PEG1500, 0.1M PCB pH 9
BM2	$P2_12_12_1$	1.68	25% PEG1500, 0.1M MMT pH 9

BM2	$P2_12_12_1$	1.64	25.00% PEG1500, 0.10M MMT buffer (1M pH 6.0)
BM2	$P2_12_12_1$	1.66	30.00% PEG3000, 0.20M Li_2SO_4 0.10M Tris-HCl (pH 8.5)
BM2	$P2_12_12_1$	1.62	n/a
BM2	$P2_12_12_1$	1.7	10.00% PEG8000, 10.00% Ethylene Glycol, 0.10M HEPES (1M pH 7.5)
BM2	$P2_12_12_1$	1.64	n/a
BM2	$P2_12_12_1$	1.6	n/a

The resolution of the crystals varied from 1.5 Å to 1.85 Å. The pH at which the crystals were obtained varied from basic to neutral. The main difference between BM1 and BM2 crystal structures is the space group; the two BM1 crystal structures were resolved in the same crystal form (triclinic space group P1) whereas the 9 crystal structures of BM2 were obtained from a crystal of a different space group, $P2_12_12_1$. As previously observed[196], two structures of the same protein resolved in different crystal forms are likely to be more different that two structures resolved in the same crystal form and the same consideration is applied for the the crystal contacts of structures derived from different crystal forms.

Overall, there is no clear dependence between observation of BM1 or BM2 and crystallographic conditions.

6.3.2 BM1

The ligand binding mode observed in BM1 follows the usual way in which ligands interact with the BRD4 protein. Specifically, the O1 (carbonyl group, Figure 6-1) of **Ligand A** in BM1 is forming two hydrogen bonds, one to Asn140 and the other to the conserved water molecule at the base of the pocket. In addition, the water network composed of five molecules[63] (W1 to W5 in Figure 6-4) at the bottom of the pocket is preserved.

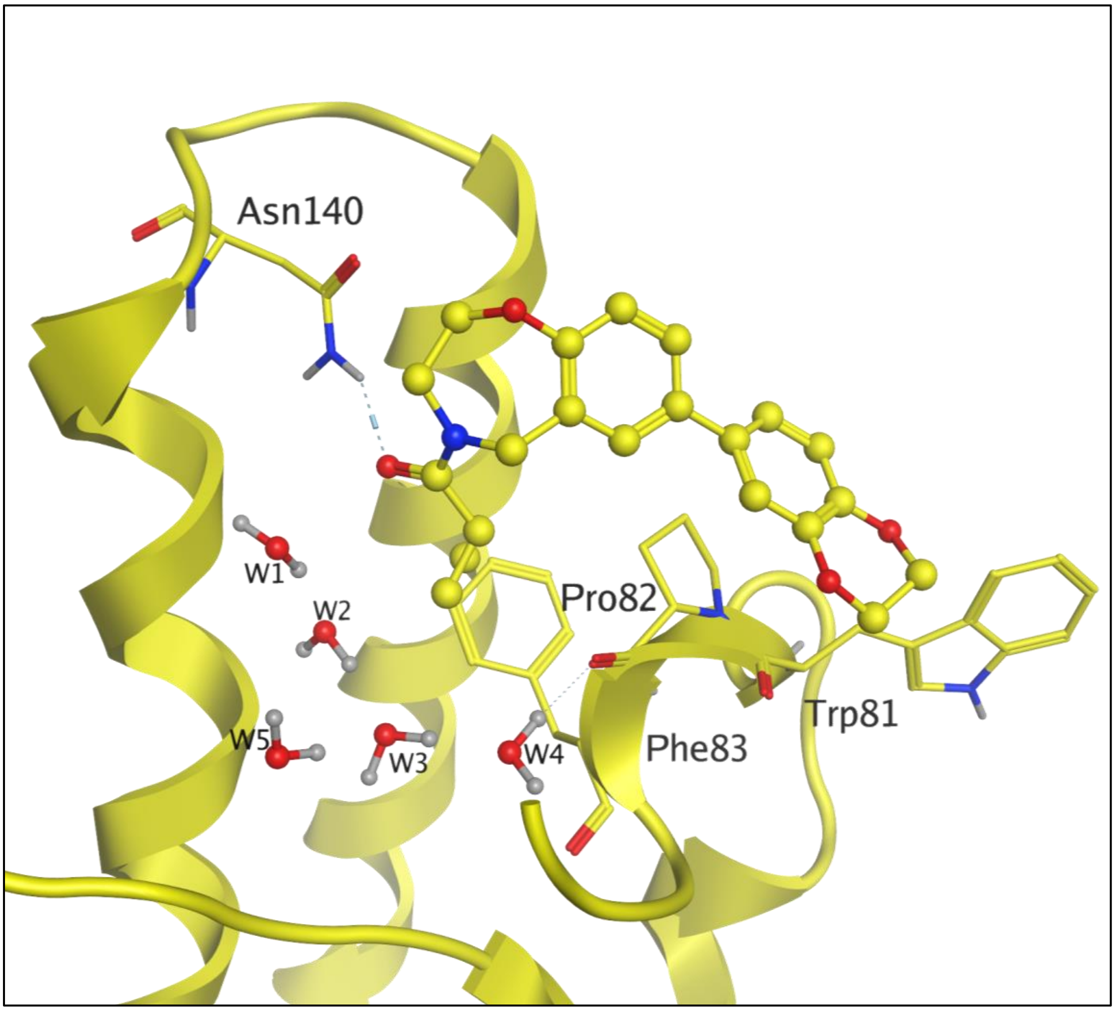


Figure 6-4 **Ligand A** co-crystallized in BRD4 assuming the so called BM1 conformation. The hydrogen bond interactions are displayed in dashed lines. The most important residues are displayed in stick representation.

6.3.3 BM2

In the BM2-like conformation, **Ligand A** is found to be deeper in the binding pocket such that the O1 (oxygen of the acyl group) is interacting through a water bridged hydrogen bond with both Asn135 and Met105 (Figure 6-5). In this conformation, interaction with Asn140 is established with the oxygen of the benzoxazepine ring (O2).

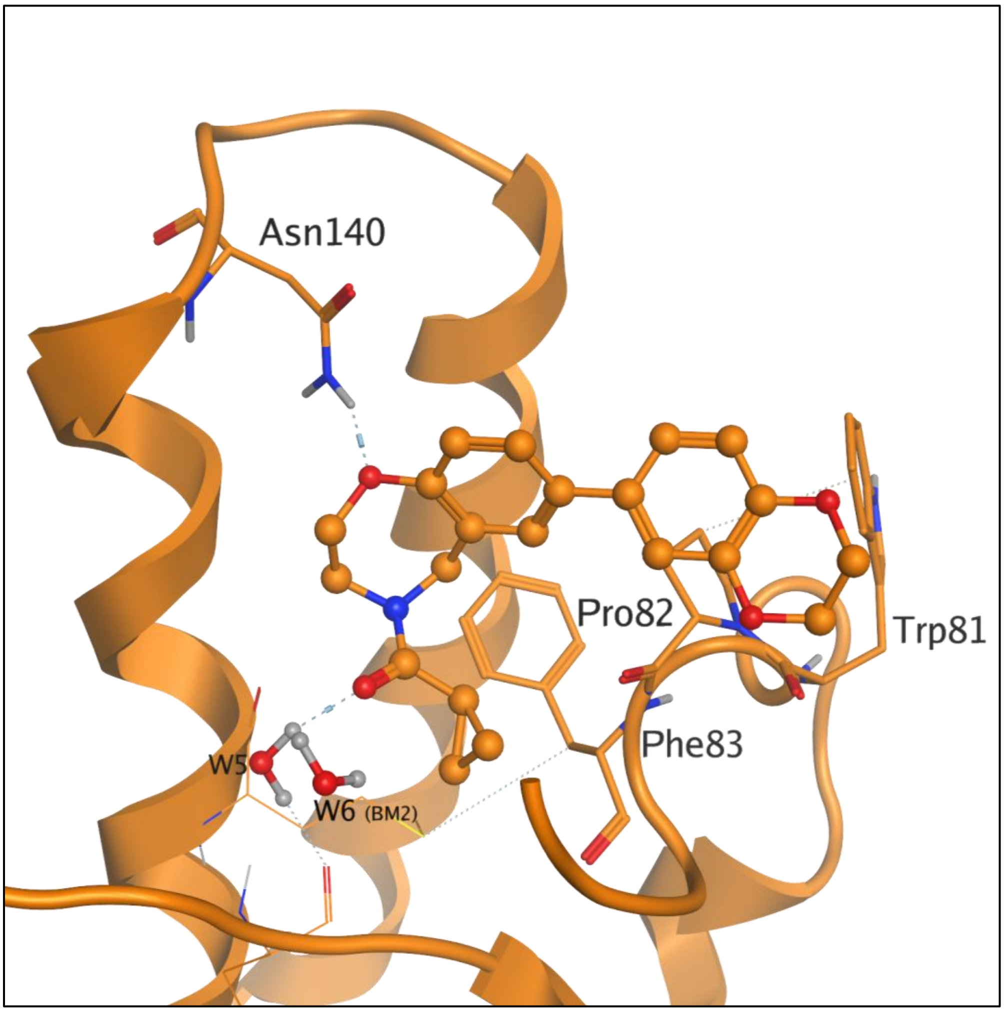


Figure 6-5 **Ligand A** co-crystallized in BRD4 assuming the so-called BM2 conformation. The hydrogen bond interactions are displayed in dashed lines. The key residues involved in ligand binding are displayed in stick representation. The Trp81 sidechain in the lipophilic shelf is assuming the "typical" conformation.

6.3.4 Differences between BM1/BM2: Trp81

One of the key differences between BM1 and BM2 is the orientation of Trp81, which belongs to the lipophilic shelf. This residue has been shown to be important in ligand binding by forming additional π–π interactions[168].

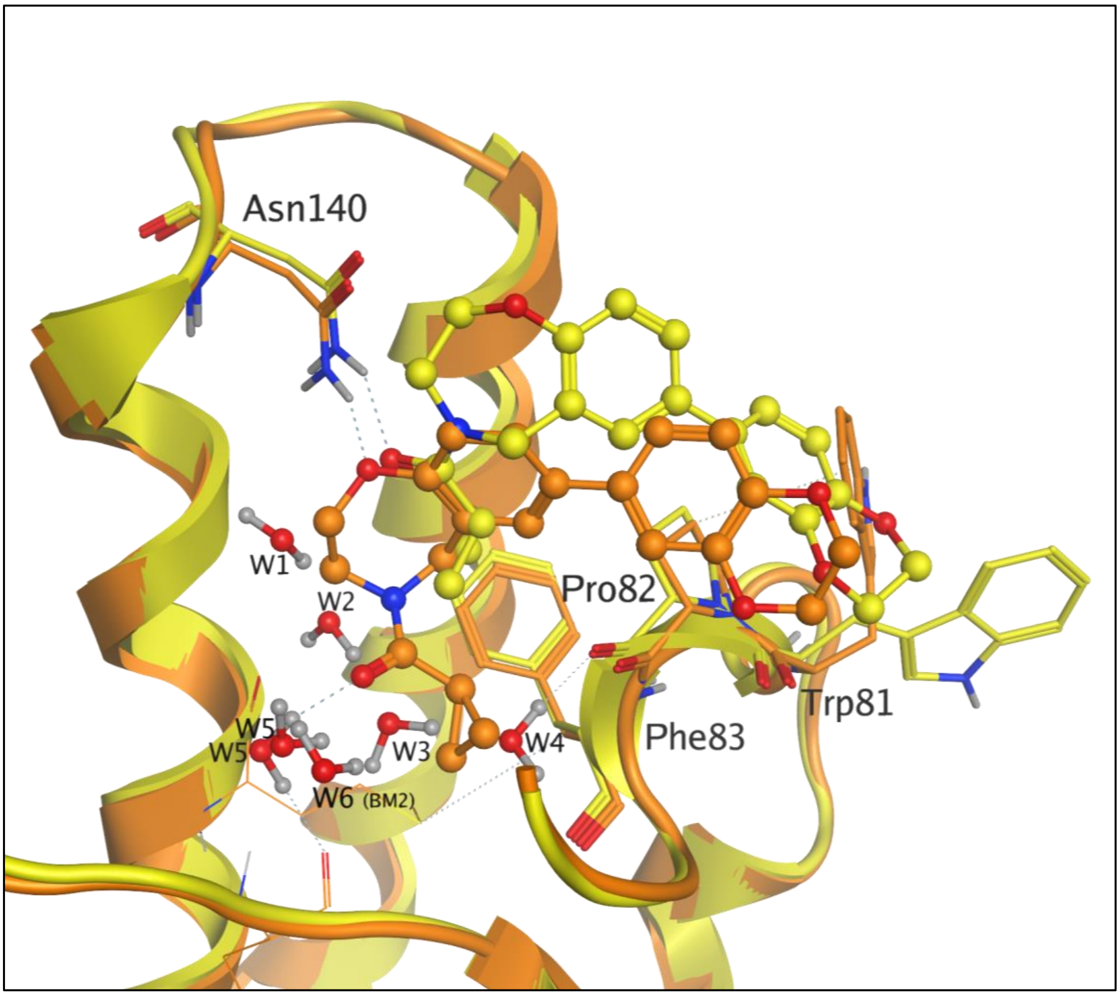


Figure 6-6 Superposition of the BM1 and BM2 binding modes. The residues in the lipophilic shelf (Trp81, Pro82, Phe83) are displayed in stick representation. The hydrogen bond between O1 (BM1) or O2 (BM2) and the amide side chain of Asn140 is displayed. The water network W1-W6 is displayed in ball and stick representation.

In order to accommodate **Ligand A** in the BM1 binding mode, the Trp81 sidechain adopts a conformation that points outward from the ligand binding site. The χ1 and χ2 angles are equal to -71.1° and 18.3°, respectively. On the other hand, when **Ligand A** is in the BM2 binding mode, the Trp81 sidechain is found to point towards the binding site with χ1 and χ2 angles of 60.6° and -89.4°, respectively. The χ1(N-CA-CB-CG) and χ2 (CA-CB-CG-CD1) angles in the two binding modes are depicted in Figure 6-7

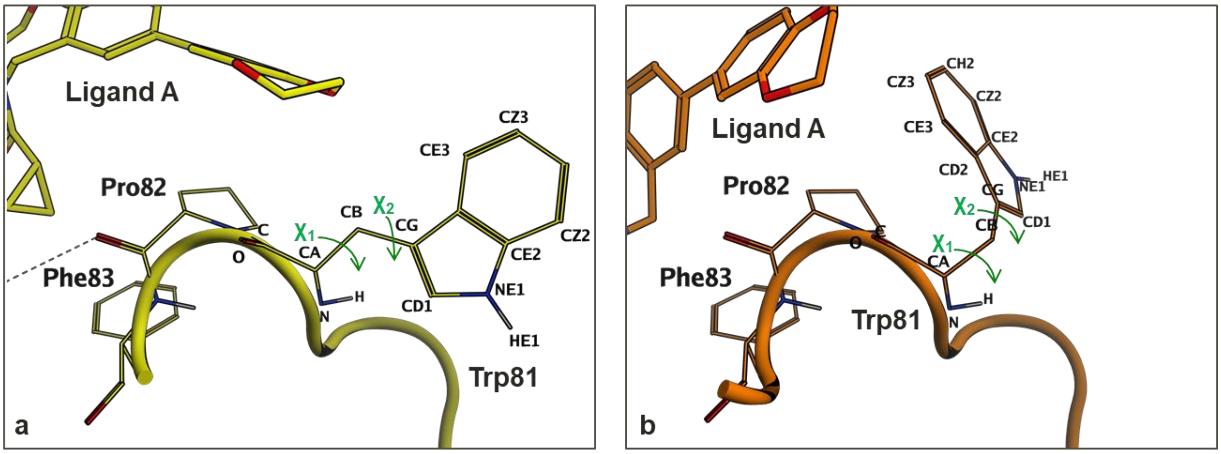


Figure 6-7 Shelf residues Trp81, Pro82, Phe83. The atom type and the torsional (χ_1 and χ_2) of Trp81 are reported for both BM1 (a) and BM2 (b).

6.3.5 Trp81 conformation in BRD4 proteins

To understand the relevance of Trp81, the conformations it adopts in BM1 and BM2 were compared to those in other BRD4 structures. To do so, all the X-ray crystal structures of BRD4 proteins found in the in the PDB (136) and in the GSK database (316) were analysed by calculating the dihedral angles χ1 and χ2 of Trp81 as illustrated in Figure 6-8.

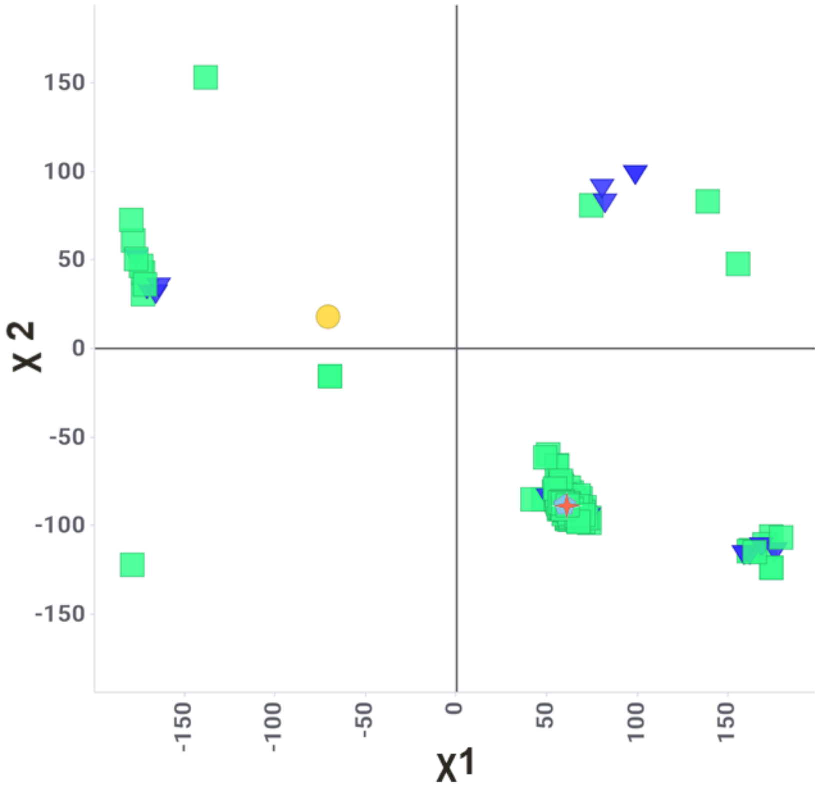


Figure 6-8 Trp81 conformations of 452 crystal structures represented by means of χ1 and χ2. The most populated region can be found around χ1~ 60○ and χ2~ -90○ (as seen in BM2). Each crystal structure is colour and shape coded by the sources from where it was retrieved either PDB (blue triangle) or GSK (green square), in addition BM1, BM2 and apo are highlighted with different shape (circle, star and hexagonal) and colour (yellow, orange and azure).

From the analysis of the 452 crystal structures (369 unique PDB), one large population of 415 crystal structures was identified (Figure 6-8) with χ_1~ 60° and χ_2~ -90°. Moreover, two additional clusters of crystal structures are found in the regions defined by χ1~ 160° and χ2 ~-115° (14 crystal structures) and by χ1~ -170° and χ2~ 45° (11 crystal structures)

The conformation adopted by Trp81 in BM2 falls within the most represented cluster in Figure 6-8. On the other hand, in BM1 Trp81 adopts a conformation not found amongst the other crystal structures (as illustrated by the yellow circle in Figure 6-8).

This might suggest that the Trp81 conformation observed for BM1 is in an energetically unfavourable conformation. The apo and BM2 structures are found in the most represented cluster.

Inspection of the crystal packing reveals that when Trp81 is in the preferred conformation (most observed cluster) it is also interacting with the proline residue located at position 95 of the next unit (Figure 6-9). Such architecture is not seen in BM1; the presence of the ligand seems to force the Trp81 in an "open" conformation which does not create any interaction with the residues in the next unit (Figure 6-9, a). Therefore, from this observation it could be hypothesised that the crystal packing exhibiting the ligand binding mode BM1 is not as favourable as the one reporting BM2.

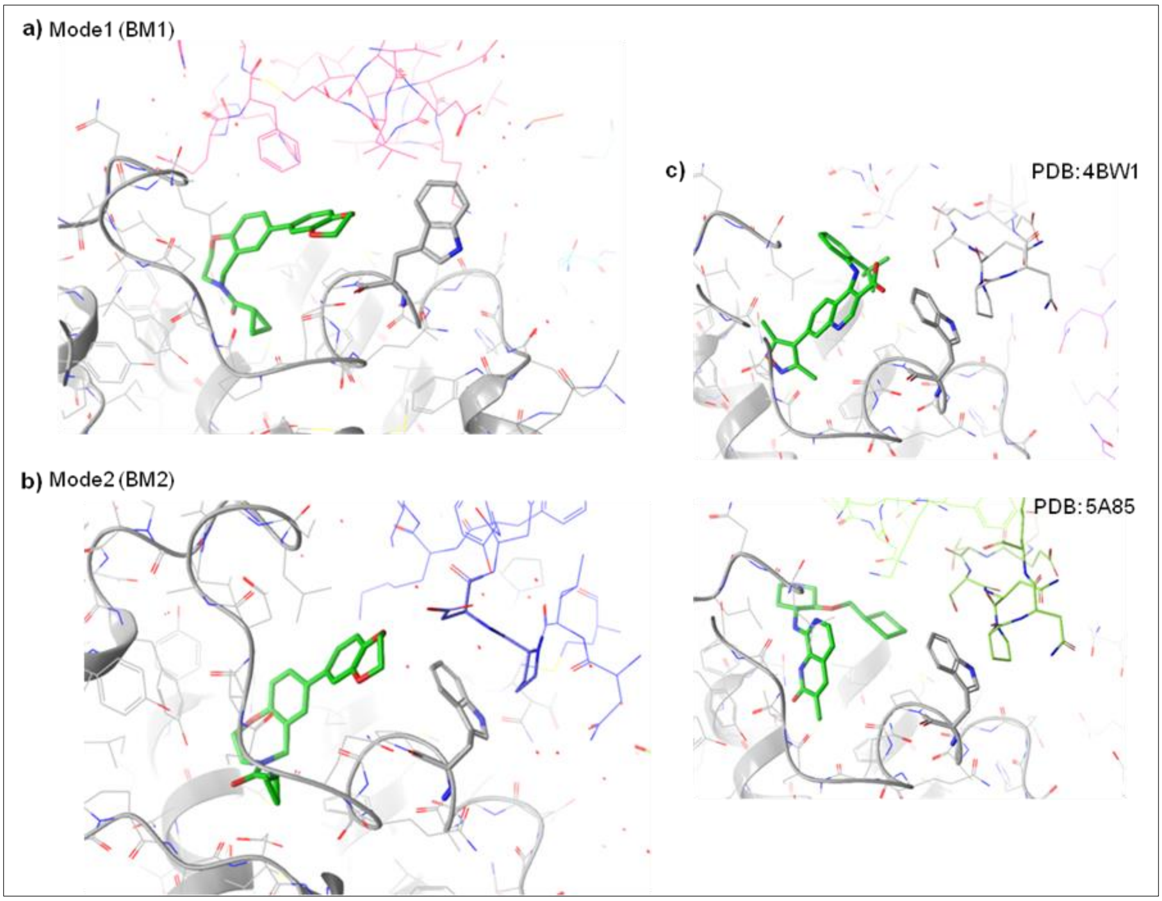


Figure 6-9 a) Crystal structure of **Ligand A** (green) in BM1 in grey with crystal contacts in magenta. b) Crystal structure of **Ligand A** (green) in BM2 with crystal contacts in purple. Trp81 has χ_1= 65.5 and χ_2= -89.4. c) Two representative PDB crystal structures in which Trp81 assumes an orientation similar to the one in BM2. In 4BW1, χ_1= 62.1 and χ_2= -87.1; in 5A85, χ_1= 65.7 and χ_2= -90.7.

The space group which describes the symmetry of the crystals was also studied in the crystal structures (from PDB and GSK) as illustrated in Figure 6-10.

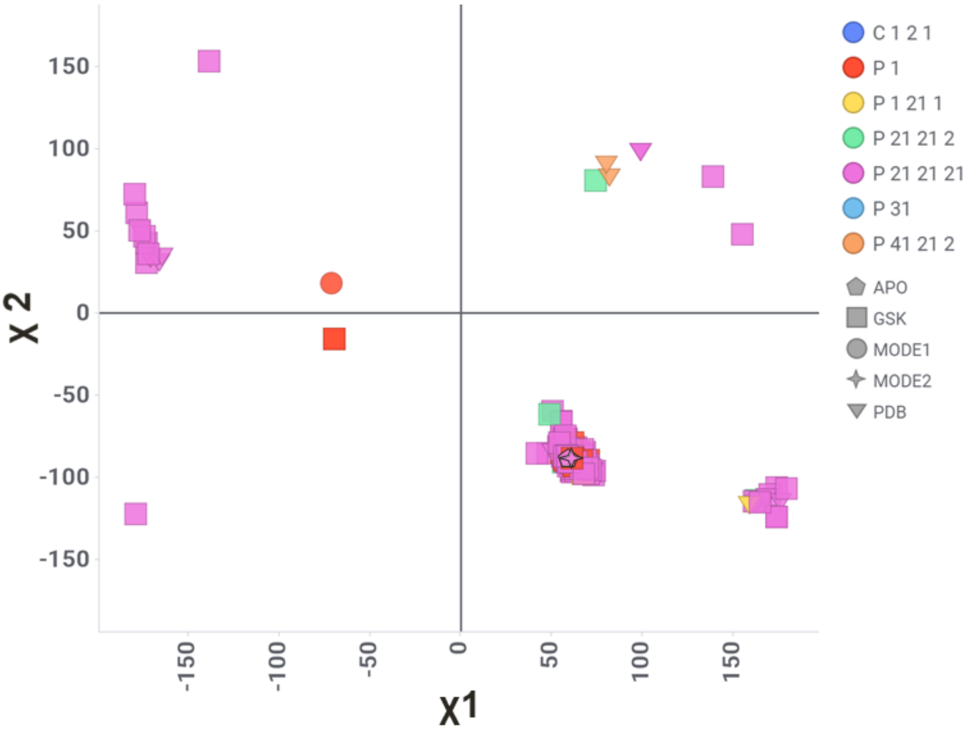


Figure 6-10 Trp81 conformations of 452 crystal structures represented by means of χ1 and χ2. The most populated region can be found between χ1~ 60○ and χ2~ 90○ (as seen in BM2). Each crystal structures is colour coded by the space group and shaped accordingly from the source they were retrieved either PDB (triangle) or GSK (square); in addition BM1, BM2 and apo are highlighted with different shape.

In general, the majority of the structures have a $P2_12_12_1$ symmetry group including the apo form and BM2. The $P2_12_12_1$ symmetry group is the most observed in the structures solved at GSK (261/316) or deposited in the PDB (115/136) and it those structures the χ1 and χ_2 of Trp81 shows different values. Moreover, the most populated cluster shows some alternatives from the $P2_12_12_1$ symmetry such as P1 and $P2_12_12$ suggesting that the conformation of Trp81 might not be entirely driven by the packing or vice versa.

In summary, from the analysis of the available crystal structures alone it could be hypothesised that the Trp81 sidechain has a preferred conformation in the crystal structures that may play a role in the formation of the crystal structures. A further important note is that no additional crystal structures are found in the Trp81 conformational space identified in the crystal structure of BM1.

6.4 Exploring the structure and dynamics of the binding pocket in BM1 and BM2

To investigate the structure and dynamics of the **Ligand A** in both BM1 and BM2 MD and MetaD simulations were carried out following the protocol reported in Section 6.2. In the following sections, the protein flexibility especially of the ZA loop is firstly analysed. Then, particular attention is given to the dynamics of Trp81 and to the water network at the bottom of the pocket with the aim to understand their implication in the observation of BM1 and BM2.

6.4.1 MD simulations initiated from the crystal structure of BM1

The folded state of the BRD4 protein is stable over five replicas of 1μs each as reported in the RMSD of the protein in Figure 6-11 with the only exception of Replica 4 in which two main different states are sampled.

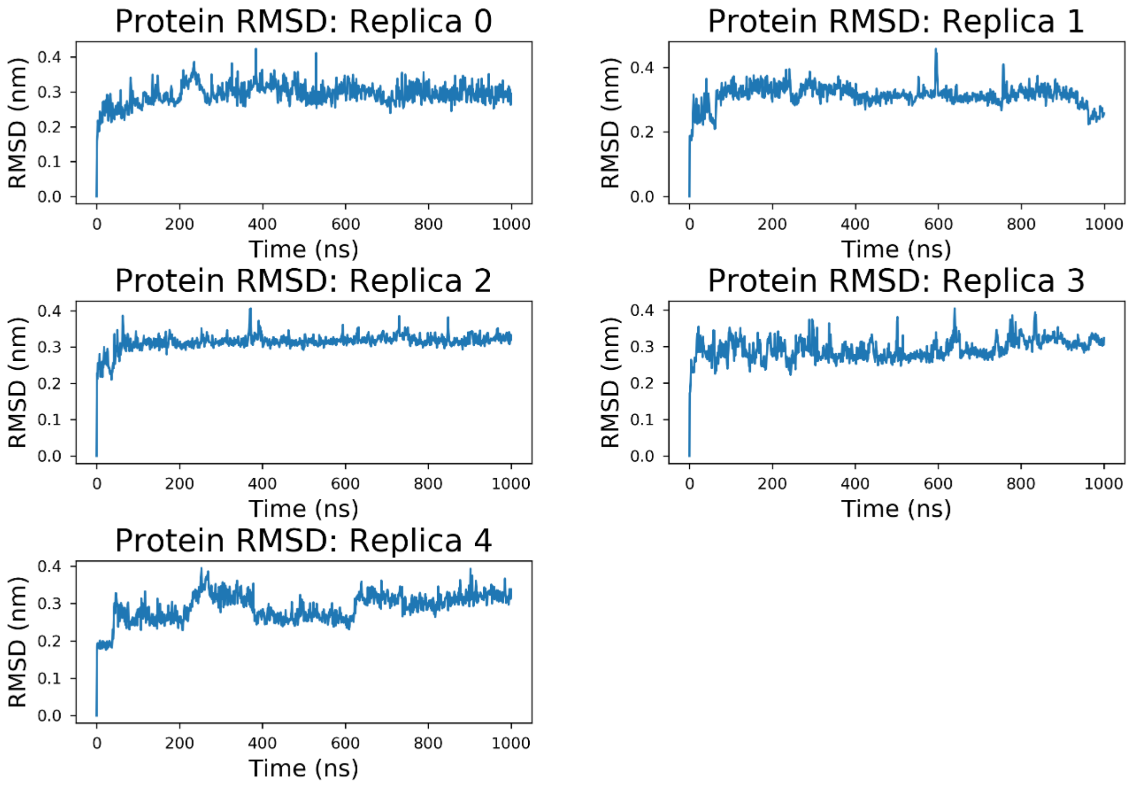


Figure 6-11 RMSD of the protein backbone during the course of the BM1 simulations.

The Root Mean Square Fluctuations (RMSF) of the alpha-carbons indicates that the fluctuations of the ZA loop and of the termini are the largest followed by a moderate fluctuation of the BC loop (Figure 6-12). The fluctuations of atoms in the ZA loops is up to 0.4 nm and the fluctuation of the individual atoms is not consistent during the replicas.

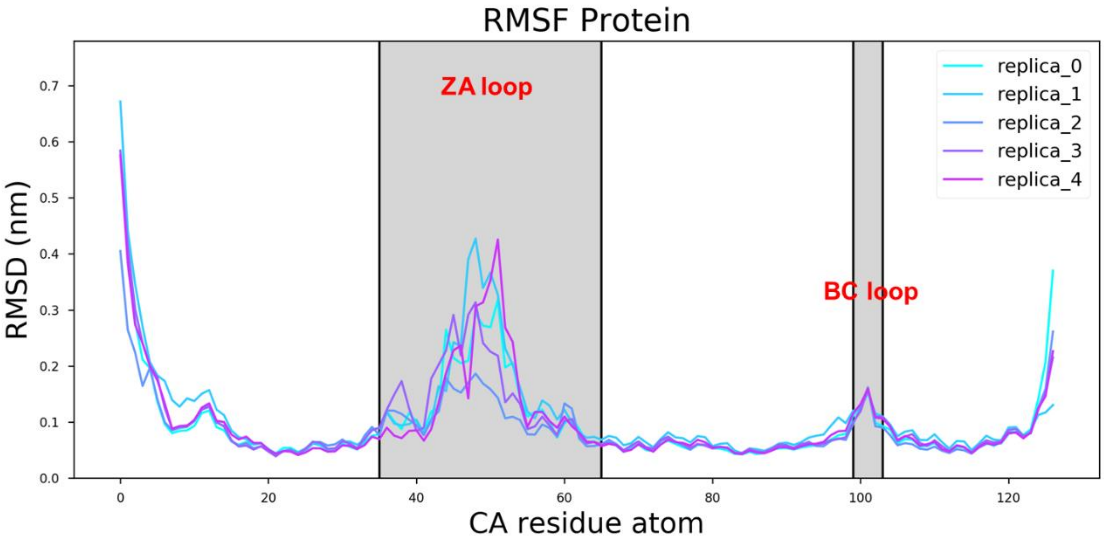


Figure 6-12 Root Mean Square Fluctuation (RMSF) of each Cα atom in the five MD replica of BM1.

In all of the replicas the ZA loop is accessing different conformations which make the pocket more open in comparison to the starting conformation (Figure 6-13) and consequently the ligand moves more freely in the binding site. This behaviour is also reported in the protein backbone RMSD for both the entire protein and for the ZA loop (Figure 6-11 and Figure 6-14).

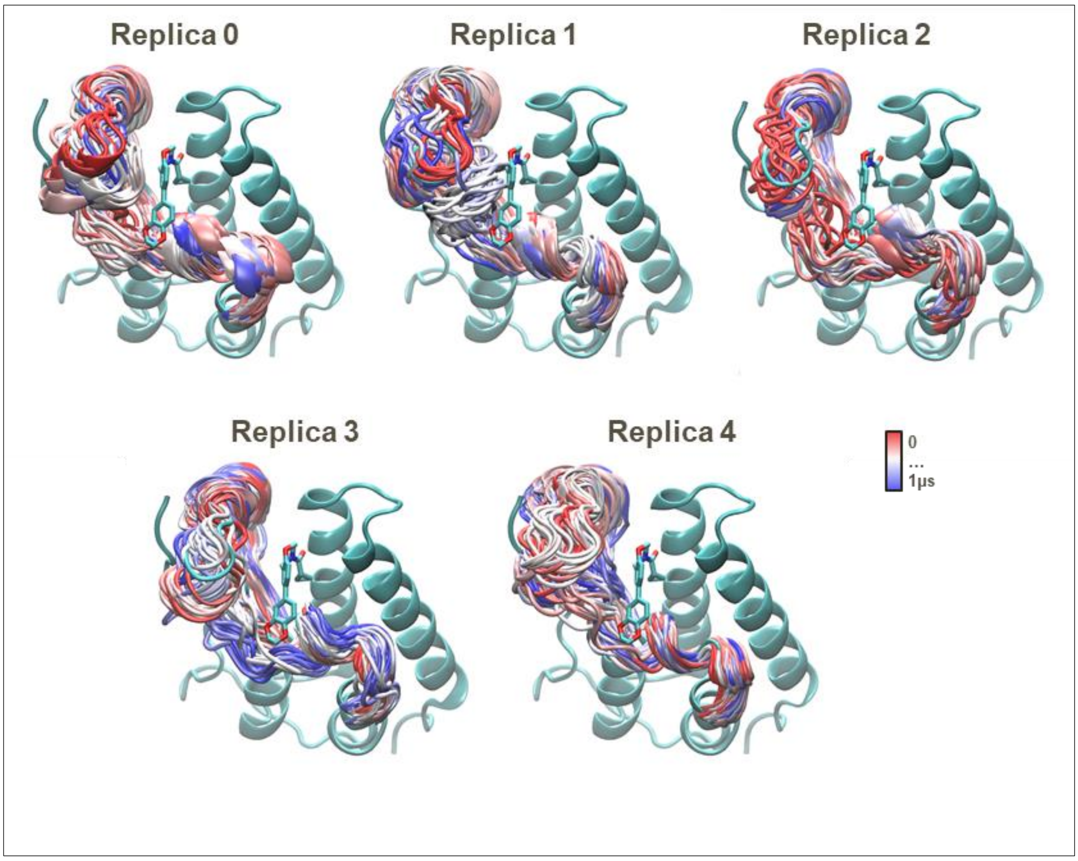


Figure 6-13 Overlay of the ZA loop of Replica 0 to 4 every 1.2 ns colour coded by timestep (red represents the beginning of the simulation, white the middle and blue the end). The structure in cyan represent the starting conformation of BM1 to which all the frames are overlaid; the ZA loop is fluctuating significantly during the course of the simulations.

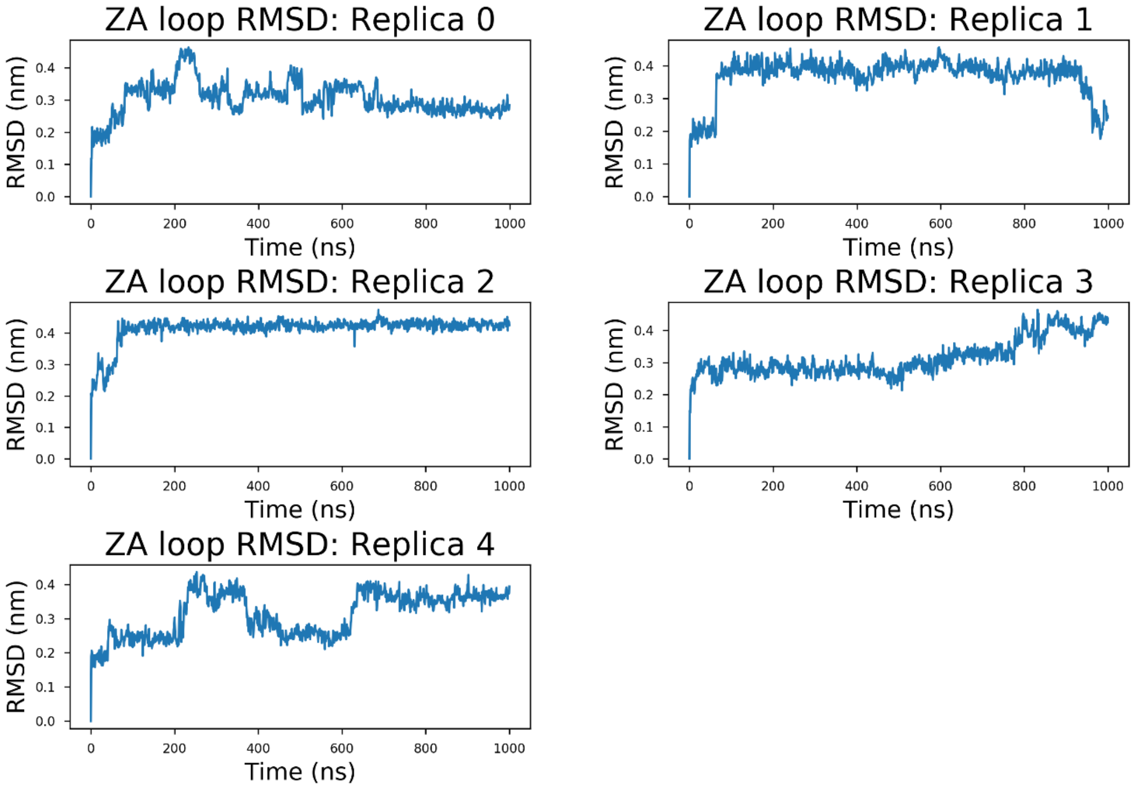


Figure 6-14 RMSD of the ZA loop (residue 77 to 106) during the course of the BM1 simulations.

6.4.2 MD simulations initiated from the crystal structures of BM2

In the five replicas of 1 μs each started from BM2 the folded state of the protein is stable as shown by Figure 6-15. The overall behaviour of the protein RMSD observed during the BM2 replicas is comparable to the one observed in BM1 (Figure 6-11). In contrast to BM1 simulations, the BM2 replicas equilibrate to a value of around 0.35 nm only after about 150 ns.

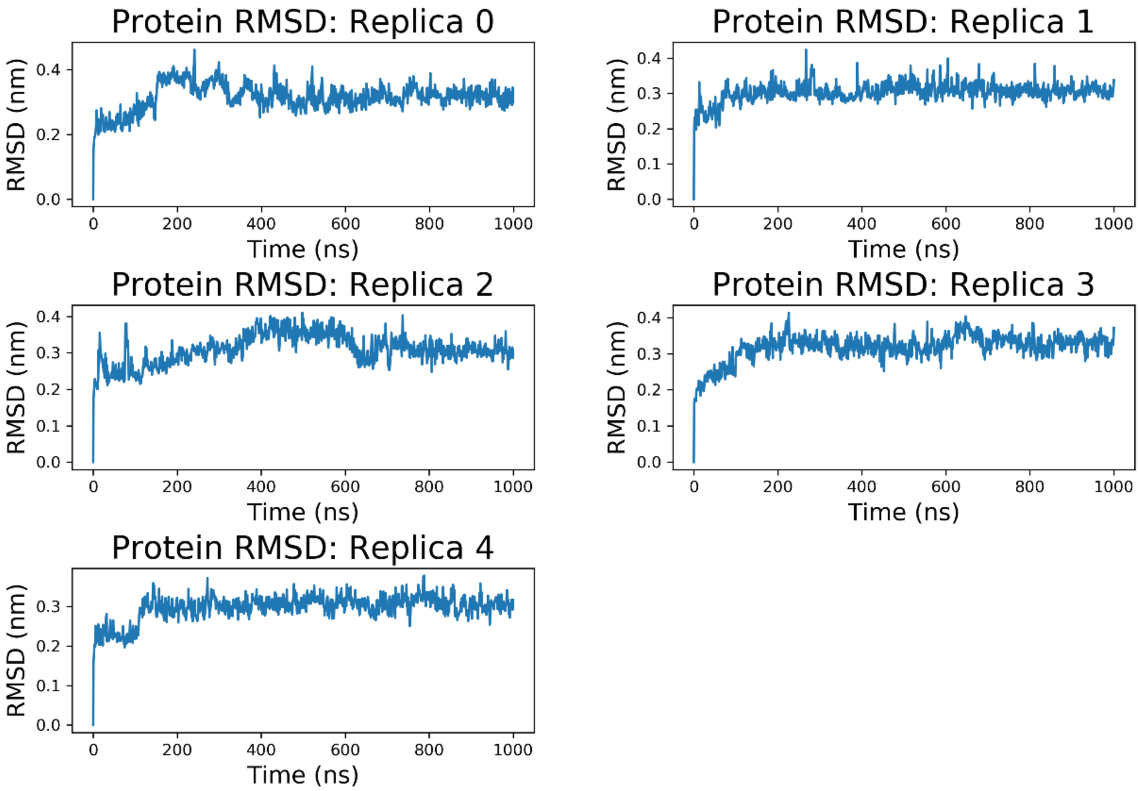


Figure 6-15 RMSD of the protein backbone during the course of the BM2 simulations

By analysis of the fluctuations of the protein Cα atoms, the loops (ZA and BC) as well as the tail are the most flexible portions (Figure 6-16). The behaviour of the loops is consistent in the five replicas of BM2 as opposed to the BM1 replicas (Figure 6-12). The RMSF of the atoms in the ZA loop reach a value up to 0.4 nm while the BC loop fluctuates up to 0.2 nm.

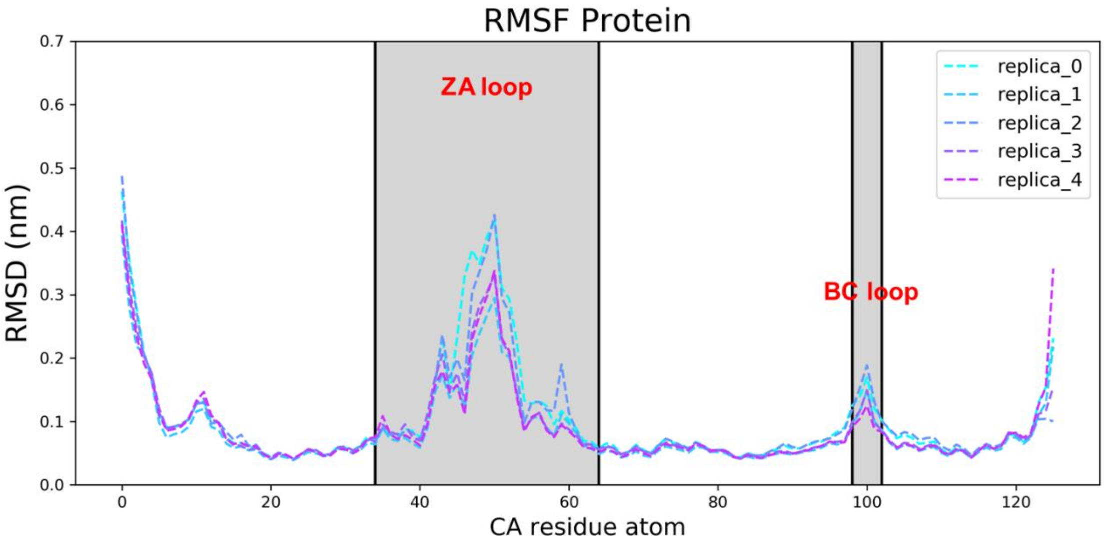


Figure 6-16 Root Mean Square Fluctuation (RMSF) of each Cα atom in the five MD replica of BM2.

6.4.3 Conformational analysis of Trp81

The Trp81 residue of the lipophilic shelf has a different orientation in BM1 and BM2 (Figure 6-6). Although crystal structures can provide a static overview of the preferred conformation of the residues (Figure 6-8), only calculations that involve dynamics can provide insights into the time dependent nature of the observed conformations as well as on the potential interchange between them.

The MD simulations started from both BM1 and BM2 following the protocol as described in Section 6.2, were also interrogated on the dynamics of Trp81 to understand whether its change in conformation is important for the observation of the two ligand binding modes. The results are reported in Appendix B. In summary, the conformations explored by Trp81 differs from the conformations obtained in the crystal structures analysis (Figure 6-10). This could signify the sampling was not adequate for the configuration space of Trp81, therefore, to enhance the sampling of Trp81 MetaD simulation of the apo form were run.

To better study all the available conformations for Trp81, the Free Energy Surface (FES) of Trp81 was reconstructed by means of MetaD simulations by using the apo protein of BRD4 (PDB 2OSS). With MetaD, the Trp81 is allowed to explore and move from one free-energy minimum to the others overcoming the free energy barriers. The collective variables used to drive the transition between the Trp81 conformations were the χ_1 and χ_2 torsional angles of this residue. The FES profile obtained from the simulations of the apo protein revealed several minima which correspond to the observed states during the MD simulations and the deposited crystal structures (Figure 6-17). Minimum *b* corresponds to the Trp81 conformation observed in the crystal structure of BM2 and minimum *c* to that observed in the crystal structure of BM1 (Figure 6-18). According to the MetaD results, minimum *b* and minimum *c* have ~1 kcal/mol ± 0.6 kcal/mol difference from minimum *a*, which appears to be the global minimum. It is important to note that the conformations observed in minimum *a*, are the ones observed most frequently in the BM2 MD simulation (Appendix B).

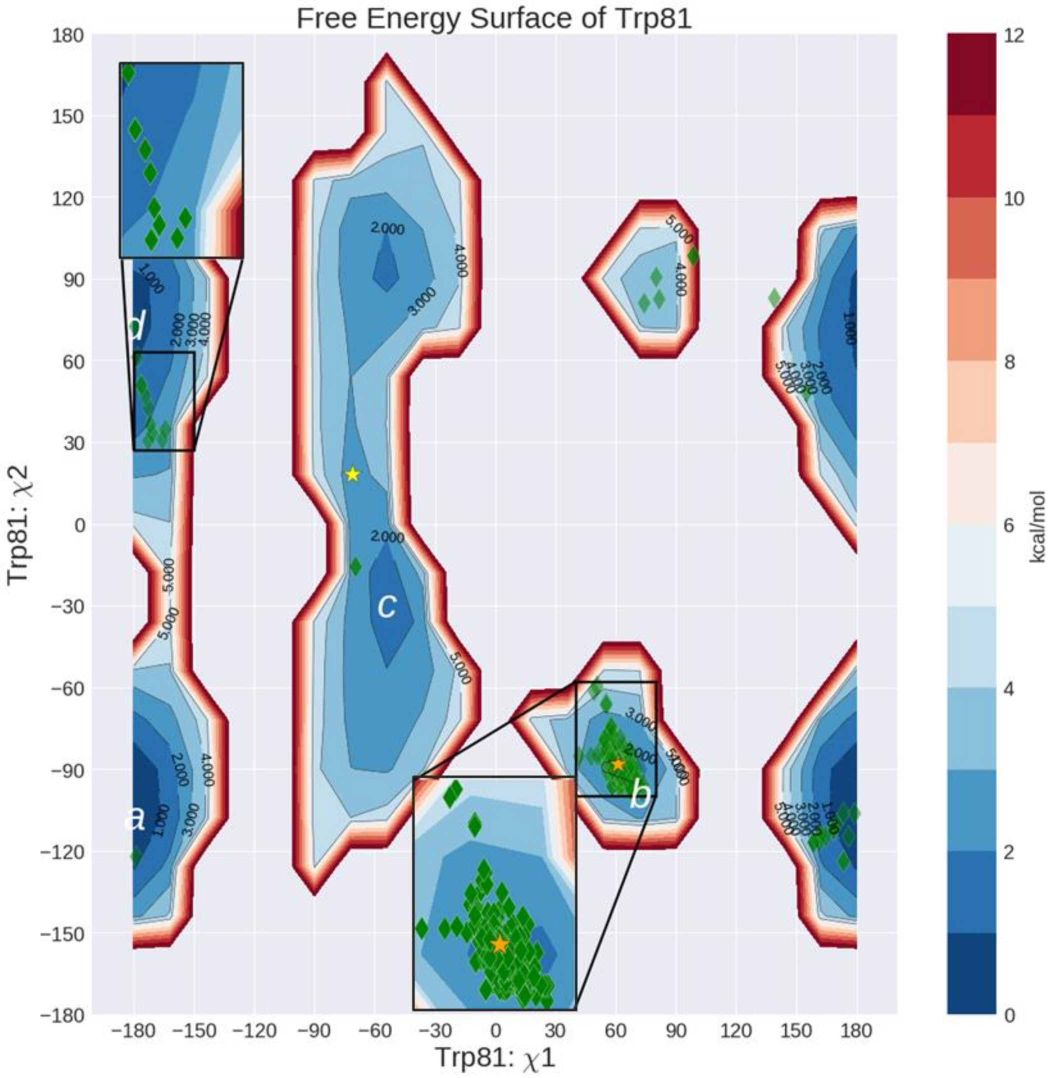


Figure 6-17 FES of Trp81 in apo crystal structure (PDB: 2OSS). The green diamonds represent the conformations of the analysed crystal structures. The yellow and orange stars represent the conformation of Trp81 in BM1 and BM2 respectively.

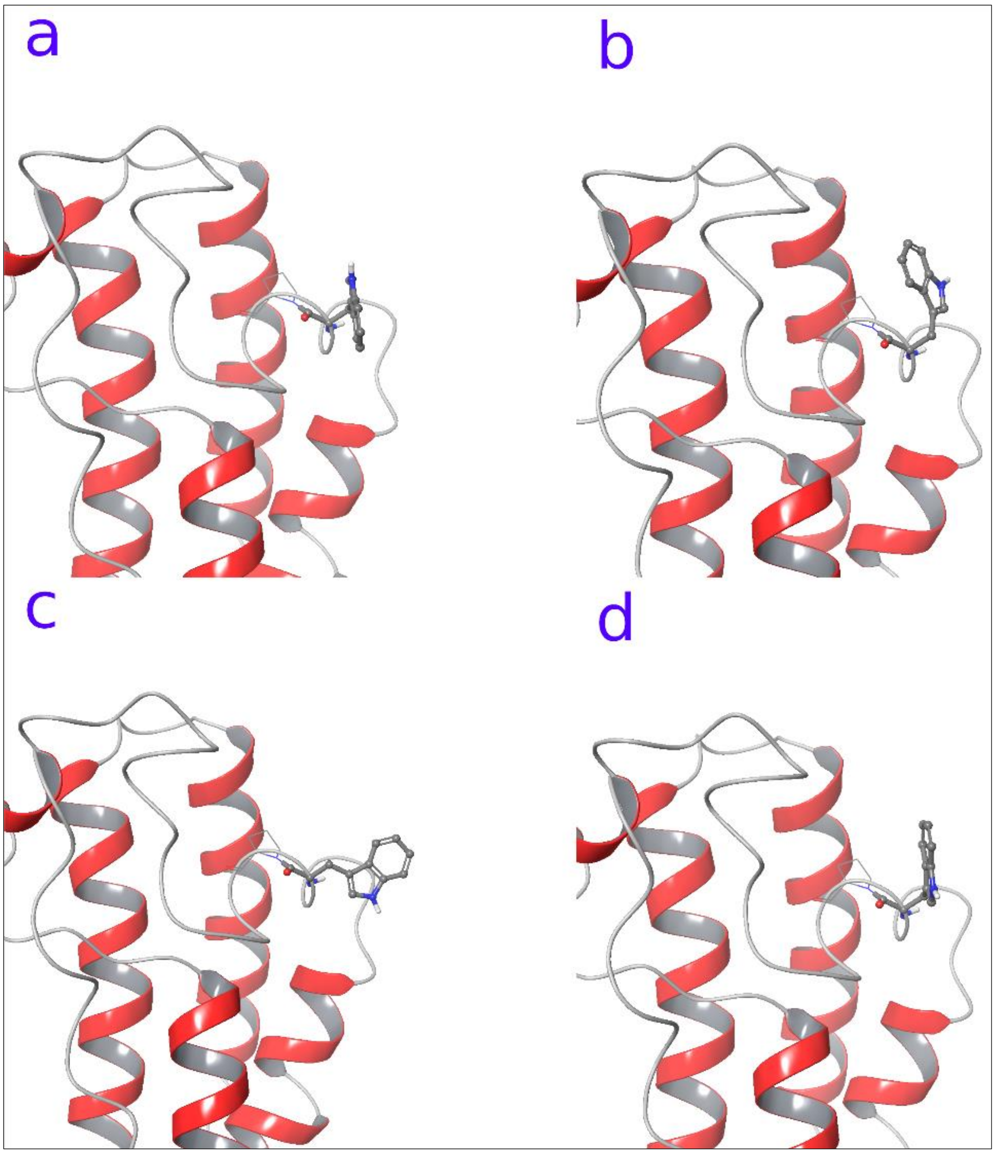


Figure 6-18 The minima explored by Trp81 during the MetaD simulations are reported.

Despite the fact that one conformation dominates in the crystal structure data (Figure 6-18, b), Trp81 was observed to change between several different conformations in the simulations in solution-phase, regardless of whether the ligand was in BM1, BM2 (Appendix B) or not present. The results from the MetaD simulation indicates that the observed conformations in the crystal structures do not correspond to the global minimum in the solution-phase. Several factors could be responsible for this result and it is not obvious which one is the most important. First of all, the conformation of Trp81 could be stabilized by the presence of the ligand in the analysed crystal structures. Secondly, crystal packing could be somewhat responsible for the shift in population towards conformation *b* from Figure 6-18. And lastly, the accuracy of the Force Field might not be sufficient to represent the true conformational landscape of Trp81.

6.4.4 Water sites in the BRD4 protein: what is known?

In BM2, **Ligand A** is inserted deeper in the binding pocket, displacing what is thought to be a stable water network (W1 to W4), whereas in BM1 the water network is preserved. From the crystal structures (Chapter 3) and studies in the literature[186, 197, 198], the water network in BRD4 is considered to be tightly bound to the BRD4 pocket. As discussed in Section 1.4, the transfer of a stable water molecule from the binding site to bulk might be energetically costly especially in the case of tightly bound water molecules as it appears to be the case for BRD4. Studies by Crowford *et al.*[63] reported ligands that were able to displace the water molecules at the bottom of the pocket with a consequent affinity loss for the ligands under investigation. More specifically, the longer aliphatic chain of compounds **4** and **5** (as reported in the original paper) as compared to compound **2** caused a rearrangement and displacement of the water molecules in BRD4-BD1 and TAF1-BD1. This appears to correlate with both affinity loss for BRD4-BD1 and TAF1-BD1 and increased selectivity for the rest of the investigated domains (CREBBP, BRPF1B, BRD9, TAF1-BD2, CECR-BD2 and BRD4-BD2).

In recent work, Aldeghi *et al*[197] showed with Grand Canonical Monte Carlo simulations that the network of waters in BRD4-BD1 as compared to other BET members is the least stable. In addition, they compared the energetic cost to displace the water molecules at the bottom of the ligand-binding pocket with the loss in affinity observed for the compounds described by Crowford[63] and found a good agreement (root mean square error of 0.8 kcal/mol).

Moderate to low affinities are observed ($pIC_{50} \leq 6.5$) for GSK compounds designed to bind to BRD4 that showed displacement of the water network from the X-ray crystal structure.

Therefore, given the available results from the literature and internal GSK knowledge, it is believed that the water network in BRD4 can be displaced with appropriate groups but at an energetic cost that is likely to exceed the gain in ligand binding energy. From this standpoint, the energetic cost required to displace the water network observed in BM2, together with the alternative ligand binding modes (BM1), is in line with the poor affinity for BRD4 of **Ligand A** ($pIC_{50} = 4.9$).

6.5 Assessing Computational Methods for the Prediction of Protein-Ligand Binding Modes

Two different ligand binding modes were observed in the crystal structures obtained for **Ligand A** bound to BRD4. The ability to predict each binding mode using limited experimental information (either one of the two crystal structures or no information) was tested with several computational tools: Induced-Fit docking, BPMD, MD and Funnel-shaped MetaD.

6.5.1 Would induced-fit docking and BPMD have been able to identify BM1 and BM2 in the absence of any information about either binding mode?

In the absence of experimental information, a possible approach to predict ligand binding modes involves the use of docking simulations. To generate solutions of **Ligand A** bound to BRD4 it was decided to use an Induced-Fit Docking (IFD)[98-100] combined with Binding Pose Metadynamics (BPMD) strategy as implemented in the Schrodinger suite release 2018.04 following the workflow in Figure 6-19. The water molecules at the bottom of the pocket were firstly removed. In the poses with a BM1-like conformation, the water molecules were inserted a posteriori, given the knowledge around the water at the bottom of BRD4 complexes (as discussed in Chapter 3). To decrease the number of incorrect ligand binding poses, the solutions were ultimately scored with BPMD which was extensively validated in Chapter 4.

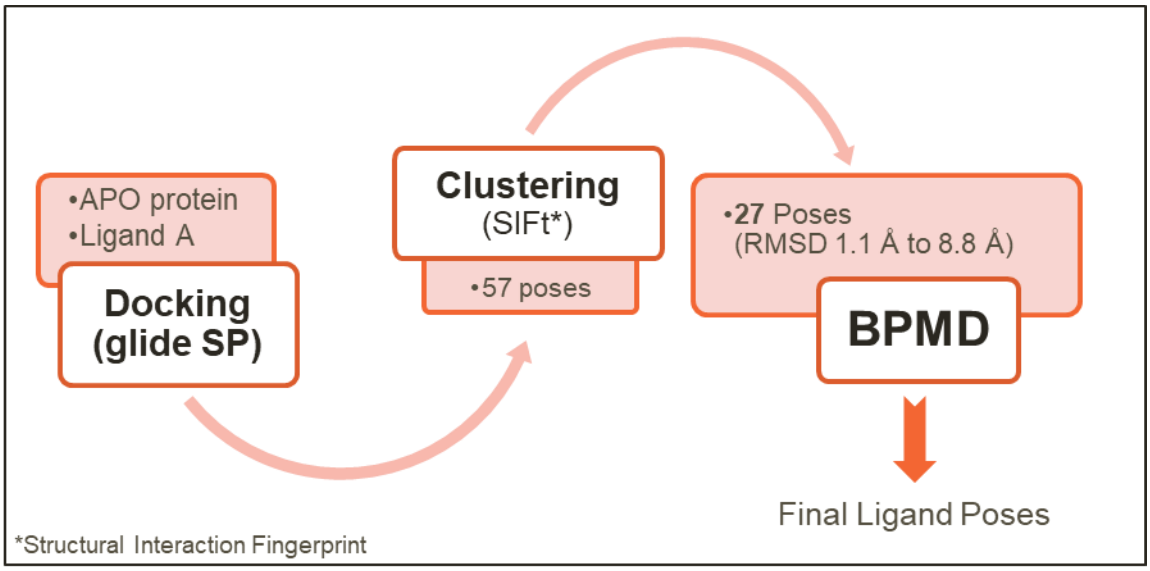


Figure 6-19 Adopted workflow for the prediction of BM1 and BM2 without using crystallographic information.

The 57 poses generated by IFD procedure as described in Section 6.2.1.4 were firstly filtered using the Structural Interaction Fingerprint[199] (SIFt) algorithm to keep only a representative pose among functionally similar poses. The 27 final poses were ultimately submitted to the BPMD tool and ranked by stability with the BPMD scores. It is assumed that the most stable poses are also the ones that should better represent the ligand binding mode. The BPMD score obtained for the prepared crystal structures of BM1 and BM2 were also added to this analysis. In this way, the stability of the complexes as obtained from the crystal structures and docking method were compared. The prepared systems (docking and crystal structures) were submitted to the same procedure as described in the Section 4.2.4 in Chapter 4. The ligand stability was studied by analysing primarily the *PoseScore* obtained from the 10 x 10 ns simulations.

In general, the poses obtained from X-ray crystallography presented a lower *PoseScore* compared to the docking solutions (Figure 6-20). BM1 and BM2 have *PoseScore* of 1.7 and 1.3, respectively. Both binding modes have *PoseScore* below the threshold of 2, suggesting that the ligand conformations in both crystal structure binding modes are stable.

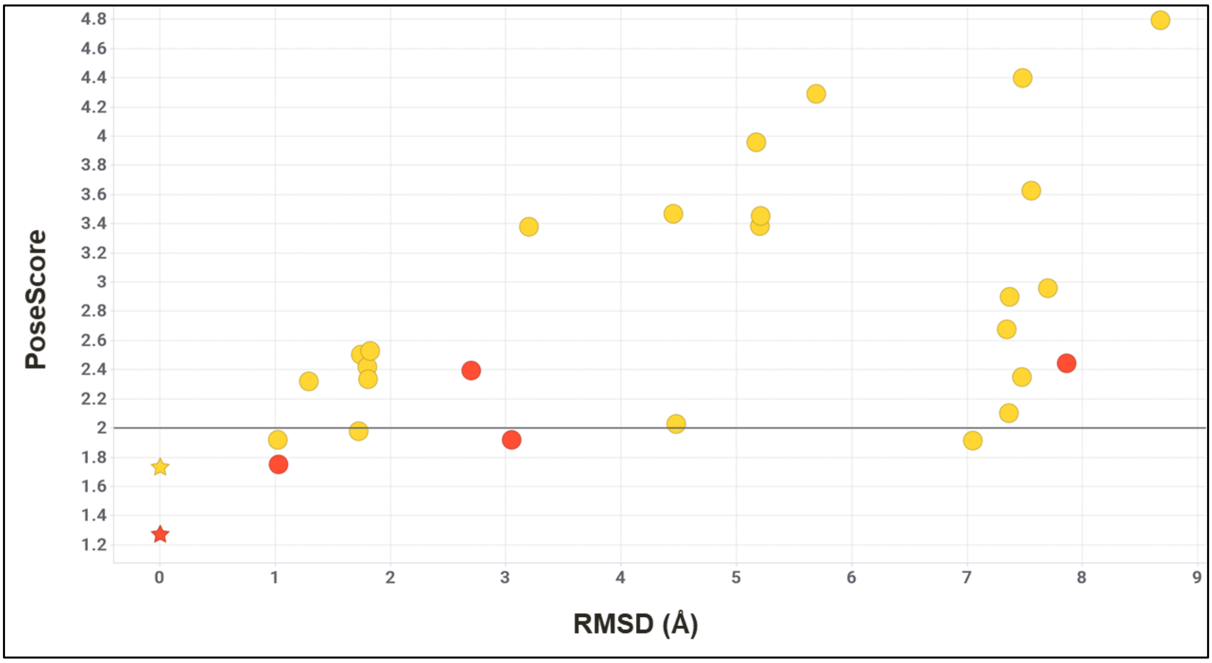


Figure 6-20 Overall *PoseScore* of BM1 (yellow star) and BM2 (orange star) and of the docking solutions with respect to ligand RMSD of both BM1 and BM2. The poses are colour coded by similarity with BM1 (yellow) or BM2 (orange).

Only 2/7 of the docking poses with ligand RMSD< 2 Å from BM1 presented a *PoseScore* below the threshold of 2 (Figure 6-21), meaning that 5 BM1-like poses are not considered to be stable. By close inspection of the docking solutions, it was observed that despite the good ligand RMSD agreement of the docked results, the conformation of the benzodioxine ring (solvent exposed portion) is not well reproduced with respect to the crystal structure of BM1, and Trp81 is adopting a conformation that differs from the one reported in the crystal structure. Those differences might be the cause of an overall less stable ligand conformation and higher *PoseScore*.

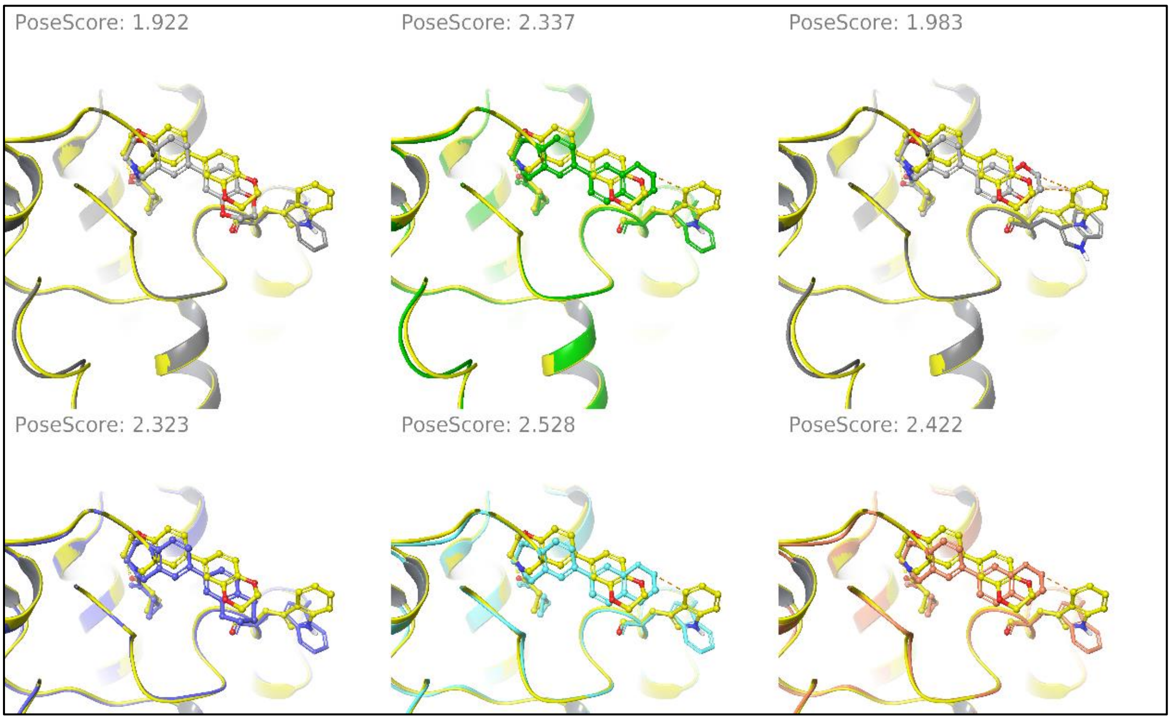


Figure 6-21 BM1 in yellow, superposed on the six docking poses with ligand RMSD < 2 Å. The *PoseScore* is reported in the picture, the docking solutions in grey (first and third image from the left) are the only ones with *PoseScore* below the threshold of 2. The Trp81 is reported in ball and stick in the case of BM1 and in stick from the docked solution.

From the total of the 27 docking solutions only one pose is similar to BM2 with a RMSD of 1.02 Å from BM2. The *PoseScore* for the BM2-like docking solution was 1.75, which is higher than the *PoseScore* of the BM2 crystal structure (*PoseScore* of 1.3). In the case of BM2 there are some important differences in the binding sites. The docked ligand is binding deeper in the pocket such that the oxygen of the amide like bond (O1) is directly interacting with Asn135; moreover, the Tyr97 which usually interacts via a water bridged hydrogen bond with the ligand is instead moved away from the ligand binding site due to the missing water molecules in the ligand binding site (Figure 6-22).

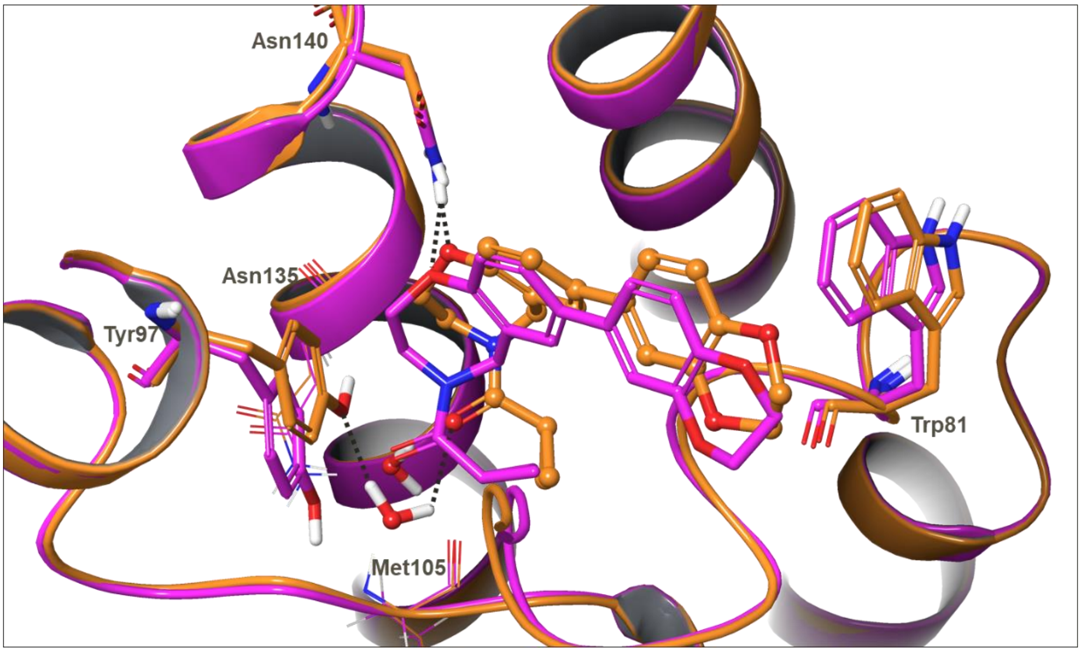


Figure 6-22 Docking solution in magenta superposed to BM2 in orange. The most important residues in the binding site are reported in the figure.

As reported in Figure 6-20, the advantage of using the *PoseScore* is in the possibility to further decrease the number of poses to select, in this case from 27 poses to only 5. To further assess the utility of the BPMD in re-ranking the IFD poses, the poses that would have been selected if only IFDScore was used were also analysed. A poor correlation between *PoseScore* and IFD score is observed with r^2 of .038 (Figure 6-23, a) and the poses with the most energetically favourable IFDScore do not all correspond to the poses with the smallest RMSD to the crystal structures (Figure 6-23, b).

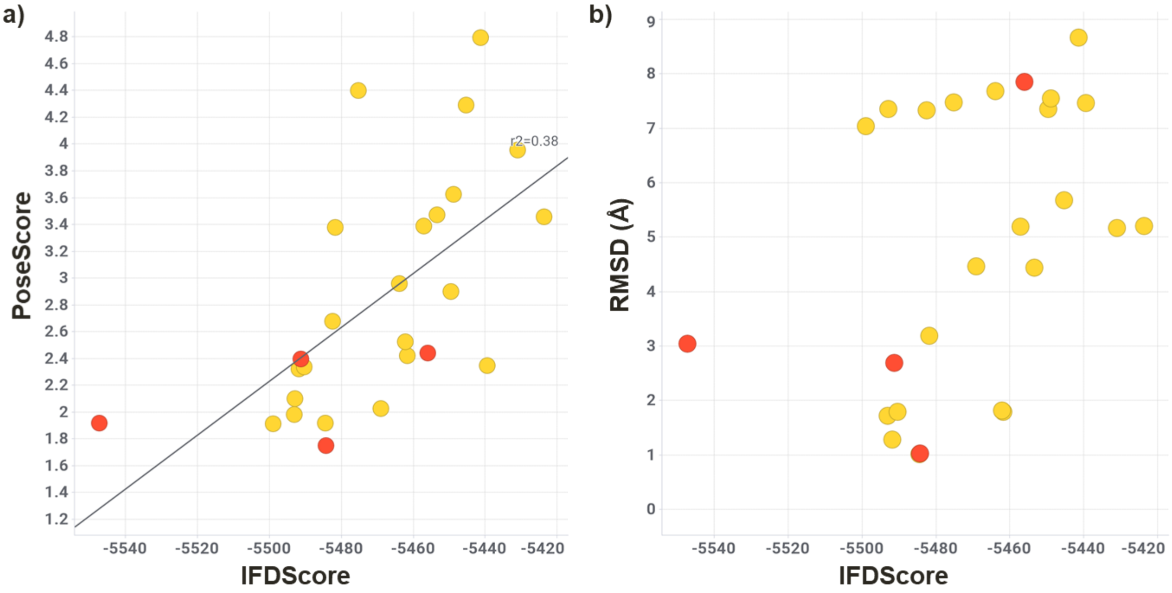


Figure 6-23a) *PoseScore* vs IFDScore of the docking solutions. b) IFDScore of the docking solutions vs RMSD from BM1/BM2; only the closest RMSD is reported and the solutions are color coded by closest RMSD to BM1 in yellow and to BM2 in orange.

6.5.1.1 Conclusions

BM1 and BM2 were successfully reproduced with docking using the apo structure of BRD4. It is important to highlight that some *a priori* knowledge was needed. In fact, in the IFD docking protocol, the insertion or removal of water molecules is not possible. Therefore, to increase the probabilities to reproduce BM2, it was decided to remove the full network of water molecules at the bottom of the ligand-binding pocket. This approach allowed the generation of many poses some of which were similar to either BM1 or BM2. In order to further decrease the number of poses and separate the correct ones, BPMD was used as a filtering tool. The most stable poses were assumed to be the ones with higher probability to better reproduce the correct ligand-binding pose.

In conclusion, by combining a strategy of IFD and BPMD, it was possible to generate several docking solutions and to successfully decrease the number of plausible poses from 27 to only 5 poses. In the 5 poses identified as the most stable from BPMD both BM1 and BM2 were recovered. Moreover, the closest pose to BM2 is scored as the most stable among all the solutions with *PoseScore* of 1.75 and the closest pose to BM1 have a *PoseScore* just below the *PoseScore* threshold of 2.

6.5.2 Would MD and MetaD have been able to identify BM2 from BM1 and vice-a-versa?

6.5.2.1 MD simulations

MD simulations were run to probe the dynamics of the **Ligand A** bound to the BRD4 protein following the protocol as described in Section 6.2.1.2. A key aspect that was monitored is the RMSD of the **Ligand A** during the simulations with the intention to capture BM2 by starting from BM1 and BM1 by starting from BM2. Furthermore, the simulations were clustered as described in 6.2.1.3 to better visualize the ligand conformations assumed during the simulations.

In the following analysis, if not explicitly specified, the ligand RMSD is calculated relative to the portion of **Ligand A** inside the pocket which is responsible for the key interaction with the protein; the benzodioxane ring which is instead solvent exposed is not included in the RMSD calculation.

6.5.2.1.1 BM1

The five replica simulations of **Ligand A** in BM1 showed that the starting conformation of the ligand is consistently lost. The ligand is accessing a range of different conformations, as observed by plotting the ligand RMSD (Figure 6-24). This is suggesting that the original ligand-protein starting structure might not be in a stable conformation. In the five replicas, a consistent finding was that the key interaction reported in the crystal structure between O1 and Asn140 was lost with the only exception of replica 3.

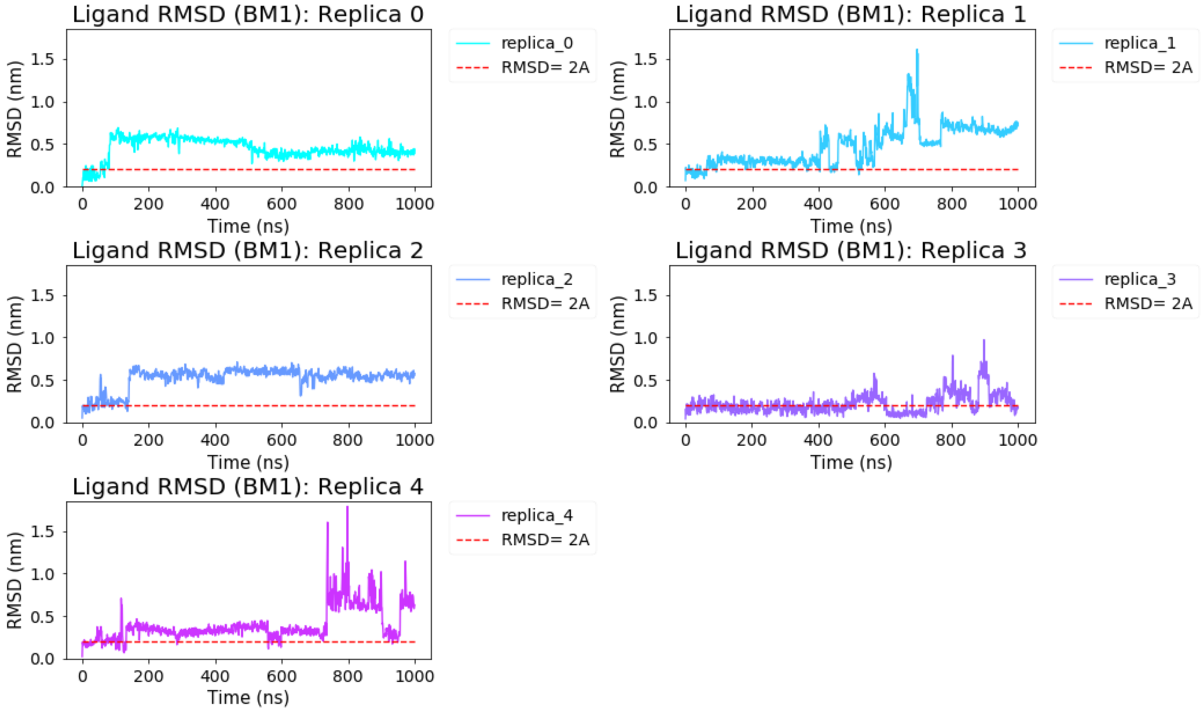


Figure 6-24 Ligand RMSD after superimposing the protein backbone to the reference crystal structure of BM1 in the 5 replicas of 1µs each. The dashed red line indicates a ligand RMSD of 2 Å.

In replica 3 the ligand maintains the starting conformation: the ligand RMSD is below 2 Å for most of the simulation (Figure 6-24), the water molecules at the bottom of the

GSK CONFIDENTIAL INFORMATION – DO NOT COPY

pocket are present and the interaction with Asn140 is maintained (see later, Figure 6-26). Only after around 700 ns is the interaction broken and the ligand explores different conformations. However, towards the end of the simulation the ligand returns to its initial conformation.

The last frame of each simulation is reported in Figure 6-25. In replica 0 the ligand assumes a BM2-like conformation until the end of the simulation; in replica 1 a new conformation is found at around 700 ns and is maintained until the end; in replica 2 after around 150 ns the ligand adopts a new conformation, which is then maintained until the end. This could signify the identification of a third alternative conformation of **Ligand A**. In replica 3 the ligand is mainly found in BM1. In replica 4 after around 700 ns the ligand is progressively leaving the pocket before entering back into the BM1 pose and subsequently leaving the pocket at the end of the simulation.

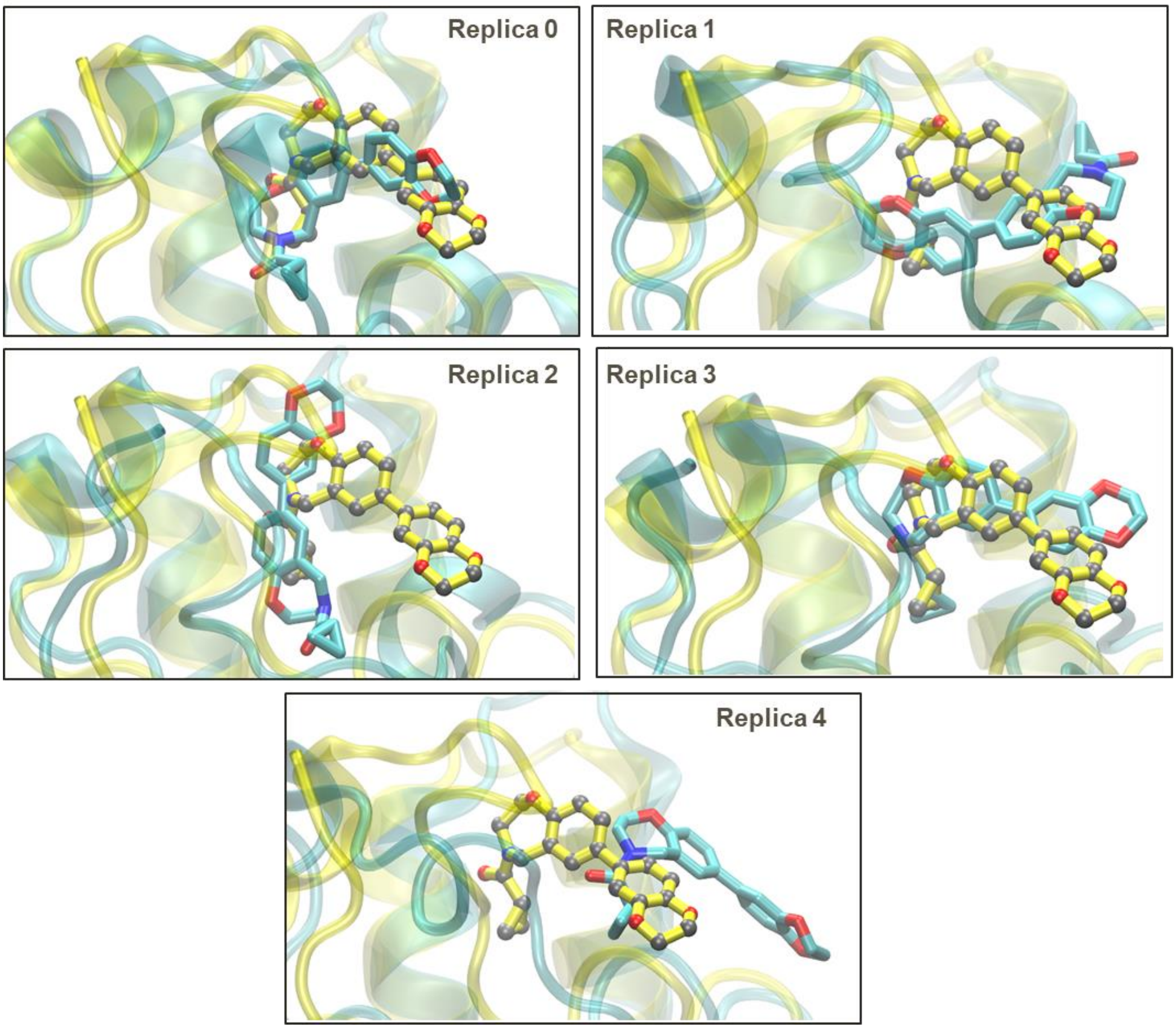


Figure 6-25 The last frame of each BM1-replicas is reported in cyan overlaid to the BM1 crystal structure in yellow. The ligand heavy atoms of the ligand are reported in grey.

The distance between the nitrogen of the sidechain of Asn140 and the oxygen O1 in the amide-like bond of the ligand has been monitored during the replicas and is reported in Figure 6-26. At the same time, the interaction between Asn140 and the oxygen O2 of the benzoxazepine ring is also monitored. It is possible to highlight the frames in which a BM2 like conformation is observed (Figure 6-27) because the O2 and Asn140 distance is expected to be shorter. In fact, in the case of replica 0 the distance between O2 and Asn140 becomes progressively shorter from more than 11 Å to less than 4 Å suggesting that a BM2-like conformation is adopted by **Ligand A**.

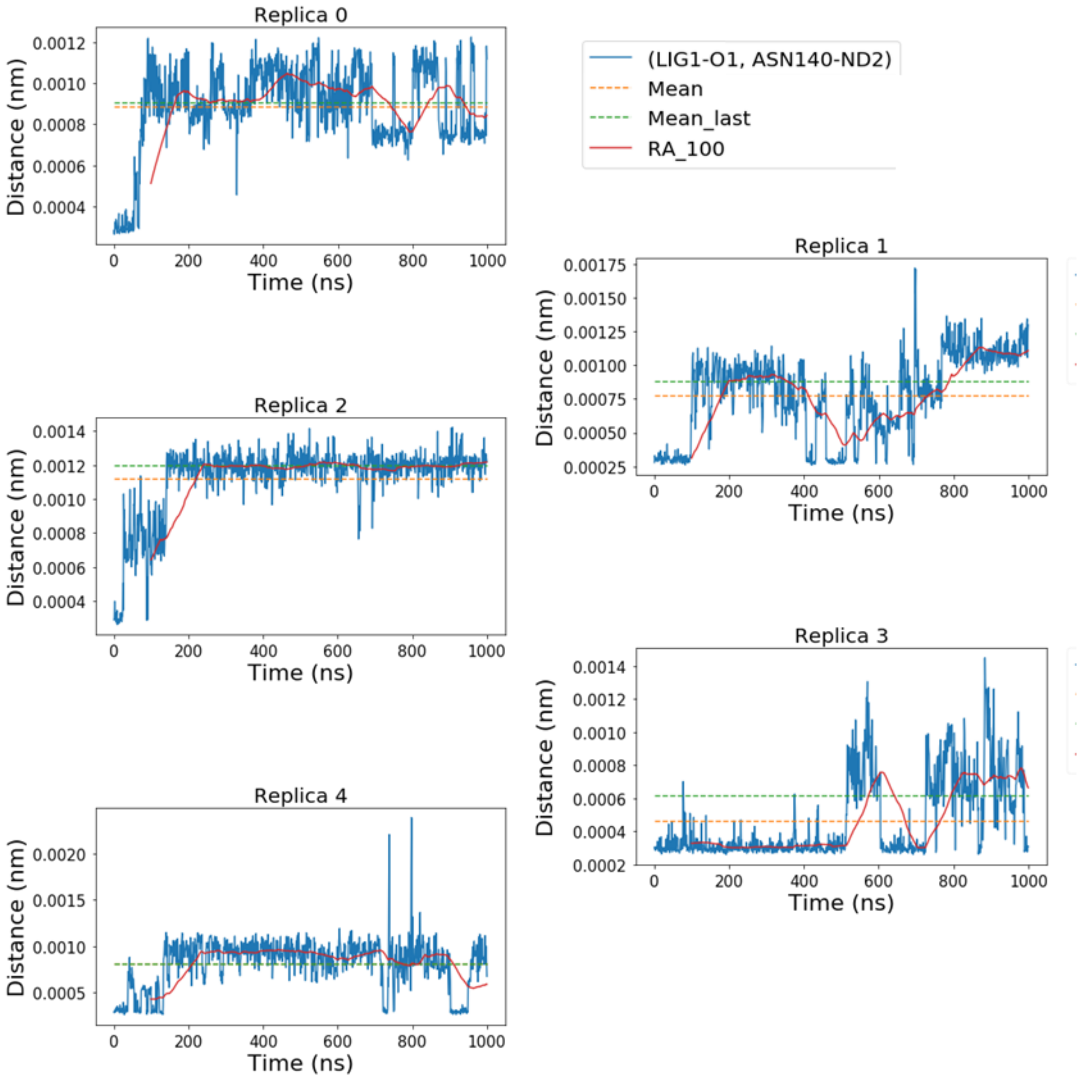


Figure 6-26 Distance between O1 of the ligand and ND2 of Asn140. The mean value over the course of the simulation and over the last 500 ns are reported with orange and green dashed line respectively; the running average, calculated after removing the first 100 ns of the simulation, is reported with red line.

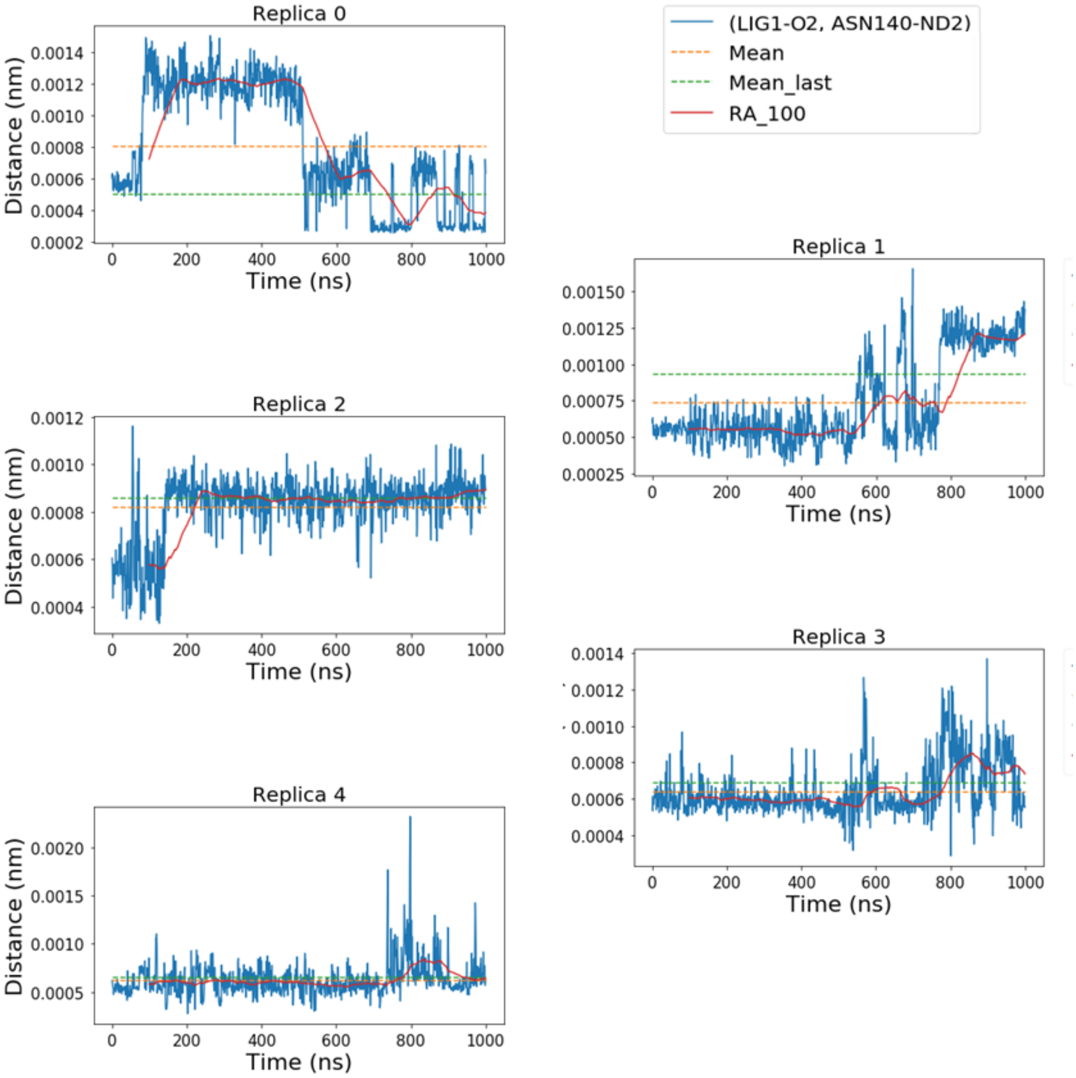


Figure 6-27 Distance between O2 of the ligand and ND2 of Asn140. The mean value over the course of the simulation and over the last 500 ns are reported with orange and green dashed line respectively; the running average, calculated after removing the first 100 ns of the simulation, is reported with red line.

The ligand in BM1 is not only found to be unstable in the ligand binding site but is also adopting a binding mode similar to the one observed in BM2 in replicas 0, 1 and 4 (Figure 6-28). By analysing both the distances from the Asn140 residues and the ligand RMSD using as reference the **Ligand A** heavy atoms in the BM2 conformation (Figure 6-28), it is clear that starting from the BM1 binding mode, the ligand is also exploring BM2-like conformations during the simulations. In those frames it was also noted that while the ligand binds deeper, the water molecules at the bottom of the pocket evacuate the binding site.

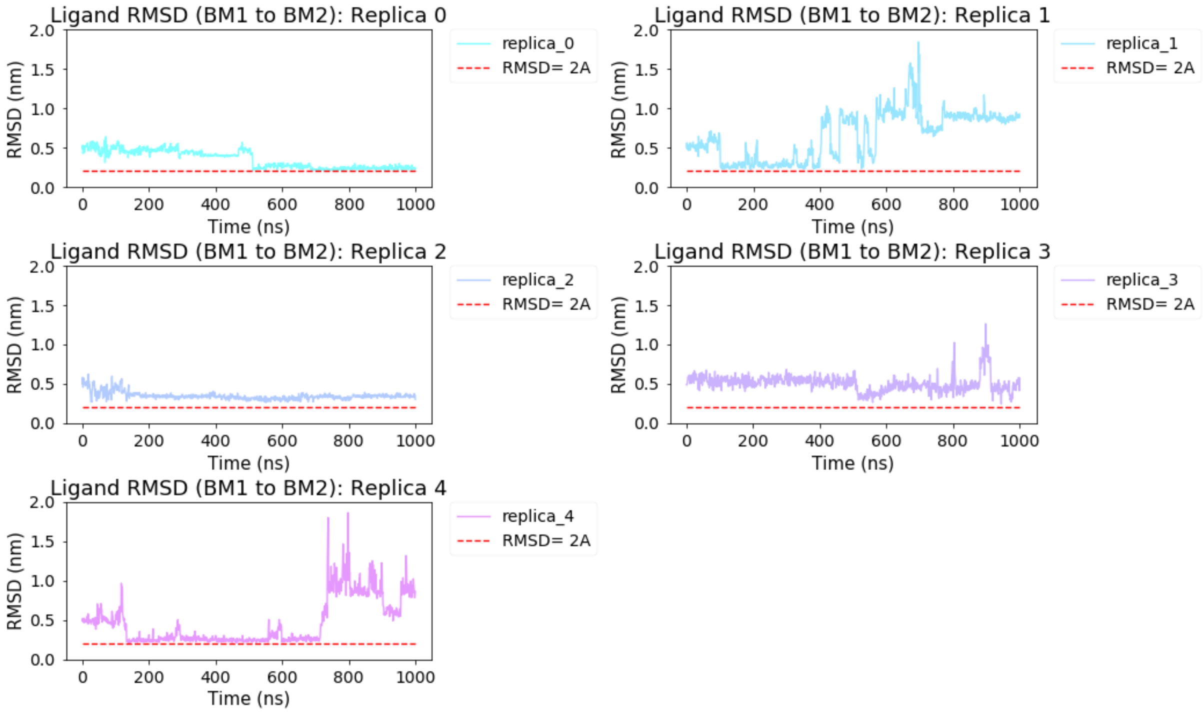


Figure 6-28 Ligand BM1 RMSD relative to BM2 conformation. Replica 0, 1, 2 and 4 present a BM2 like conformation.

The clustering procedure allowed to better visualize the frames in which a BM2 conformations was adopted. The 10000 frames from all the replicas were clustered using the RMSD of the ligand heavy atoms to generate up to 20 clusters. Not all the clusters are reported but only the clusters in which the number of frames was greater than 100. The most populated cluster, cluster 1 (1481 frames) and cluster 4 (742 frames) have a BM2-like conformation; the benzodioxine ring assumes a conformation that differs from both BM1 and BM2. A BM2 conformation is identified in Cluster 5 (641 frames) whose frames belongs mainly to replica 0, 1, 2 and 4 (Figure 6-30).

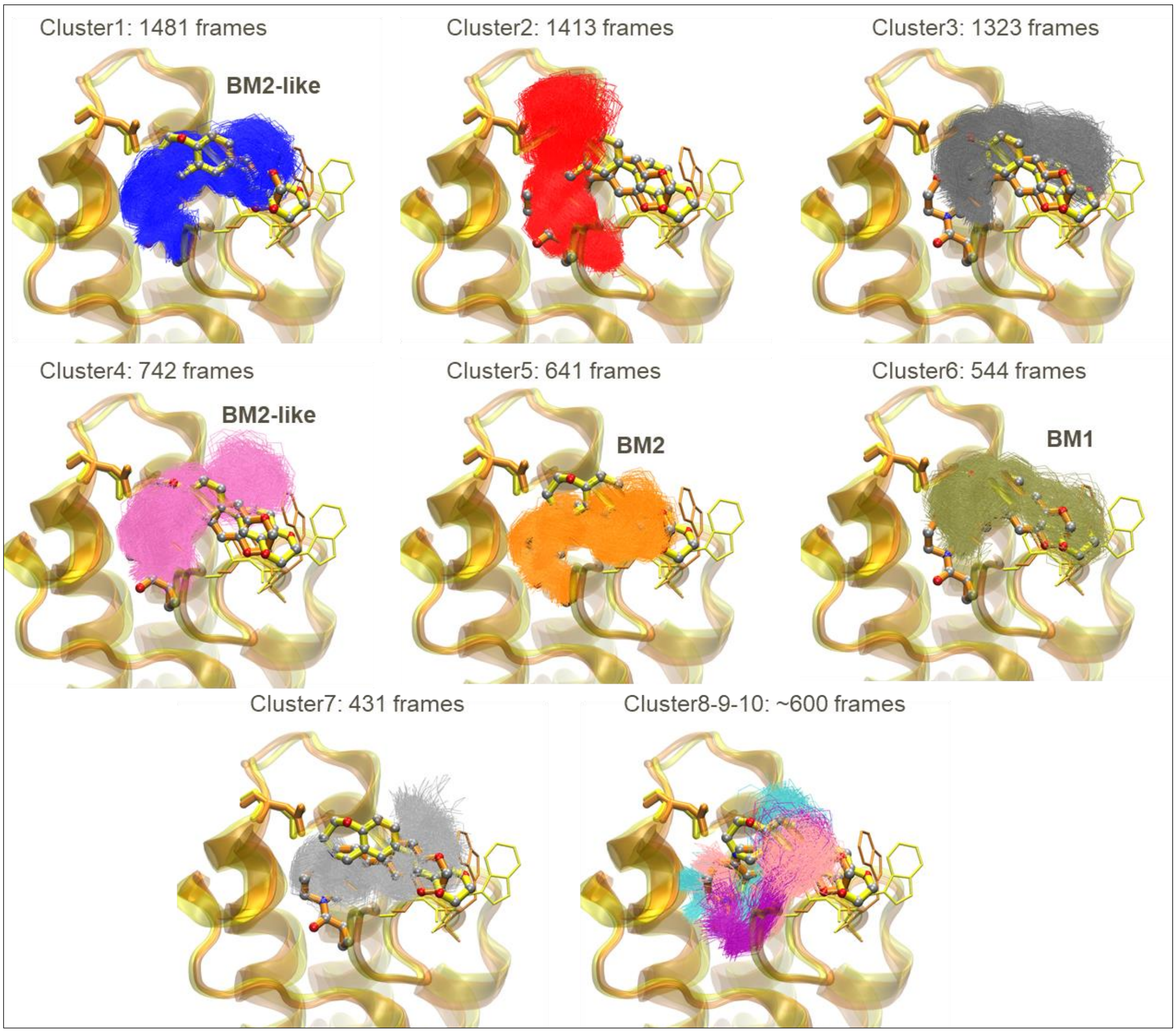


Figure 6-29 The clusters obtained from the BM1 replicas are reported in lines and overlaid to the crystal structure of BM1 (yellow) and BM2 (orange). Asn140 and the residues in the shelf (Trp81, Pro82 and Phe83) are reported in stick.

Cluster 2 is only populated by frames from replica 2 while Cluster 6 shows a BM1 conformation and is populated by frames from replica 3 (Figure 6-30). Cluster 7 contains a flipped conformation of **Ligand A** with the benzodioxine ring inside the ligand-binding site. Lastly, clusters 8, 9 and 10 are variations of a similar conformations in which the ligand is bound deeper in the binding site.

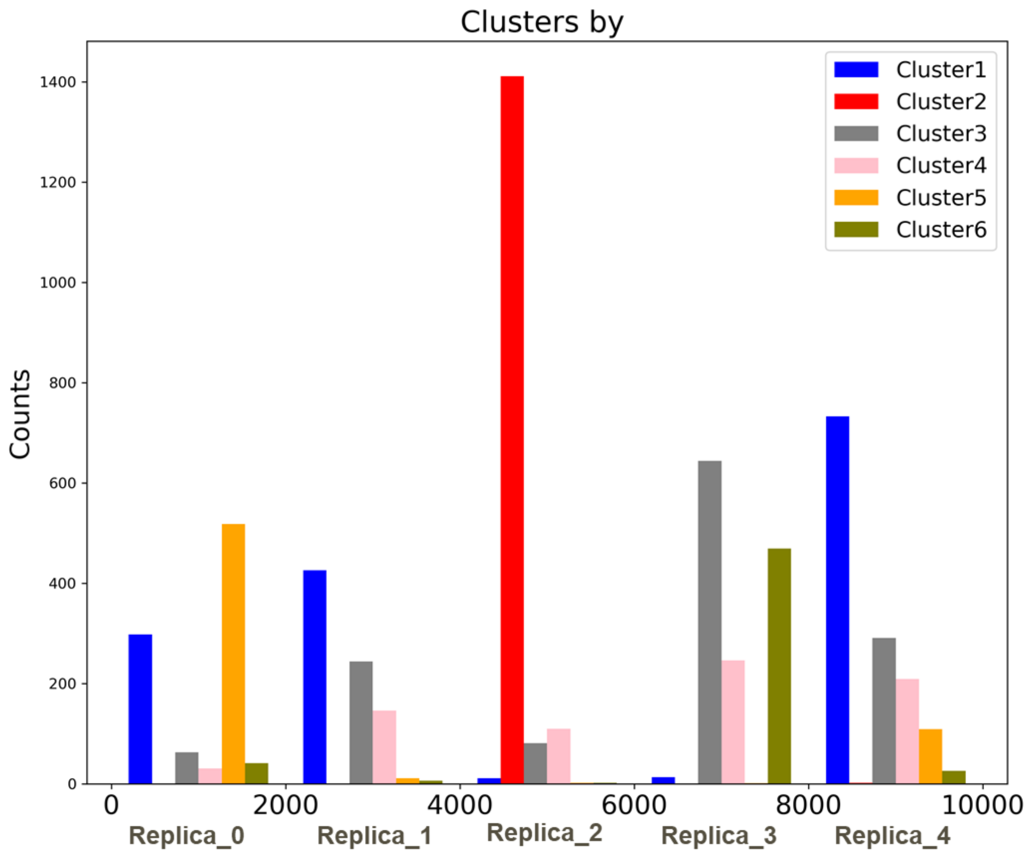


Figure 6-30 Clusters of BM1 identified during the course of the 5 replicas (2000 frames each).

6.5.2.1.2 BM2

The **Ligand A** in the 5 replica simulations of 1 μs was not as flexible as in BM1. The ligand RMSD, relative to the portion inside the binding site is maintained at 2 Å from the starting position (Figure 6-31).

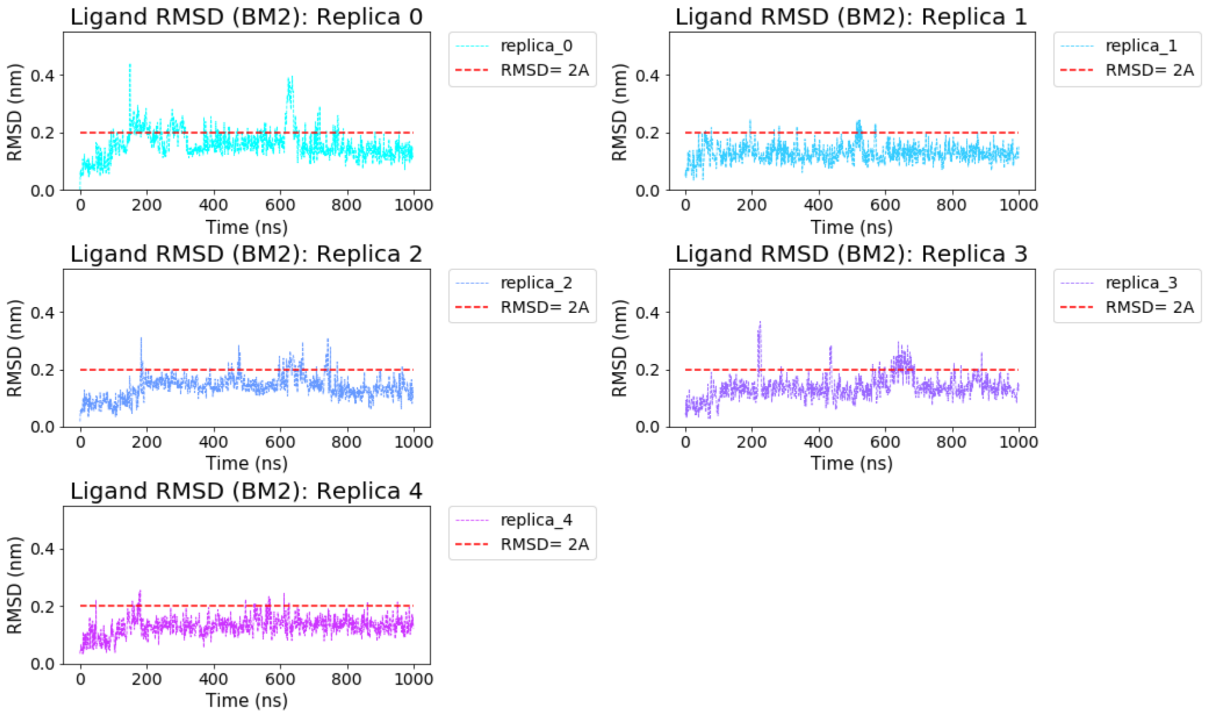


Figure 6-31 Ligand RMSD after superimposing the protein backbone to the reference crystal structure of BM2 in the 5 replicas of 1µs each. The dashed red line indicates a ligand RMSD of 2 Å.

The BM2 ligand maintains the key interactions that were present in the starting conformation of the X-ray crystal structure. Moreover, the water bridged hydrogen bonds formed with Met105 and Tyr97 are consistently present in all of the replicas.

The benzodioxine motif is mobile, adapting its conformation in the solvent and establishing hydrophobic contacts with the residues in the shelf, especially Trp81. This can be observed by plotting the RMSD of **Ligand A** including also the benzodioxine portion and not only the binding site portion, Figure 6-32. Overall the ligand is not experiencing significant changes in the binding mode apart from binding deeper in the pocket, similar to the conformation also observed in BM1 (Figure 6-29, replica 2, cluster 2).

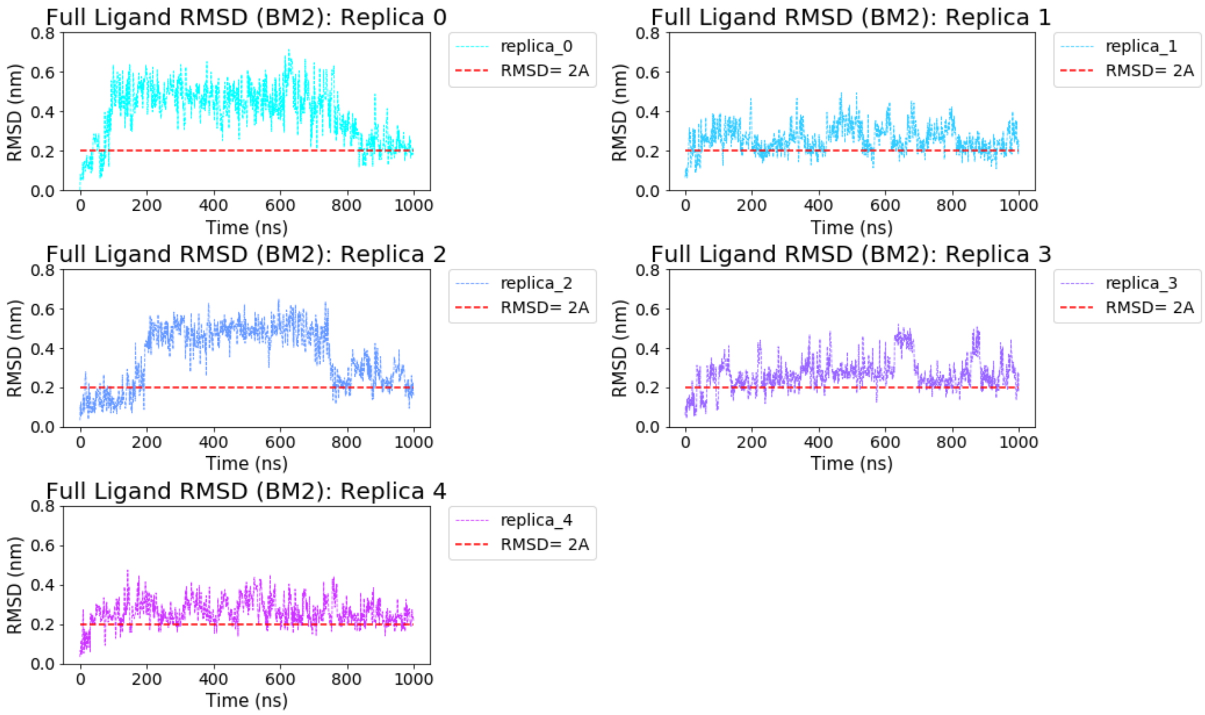


Figure 6-32 Evolution of the **Ligand A** RMSD in BM2 using as reference all the heavy atoms for the 5 replicas of 1 µs each. Note that the RMSD is up to 0.7 nm.

By starting the MD simulations from BM2, the ligand is never adopting a BM1-like conformation in any of the five replicas. This can be visualised by plotting the ligand RMSD using as reference **Ligand A** in BM1 (Figure 6-33).

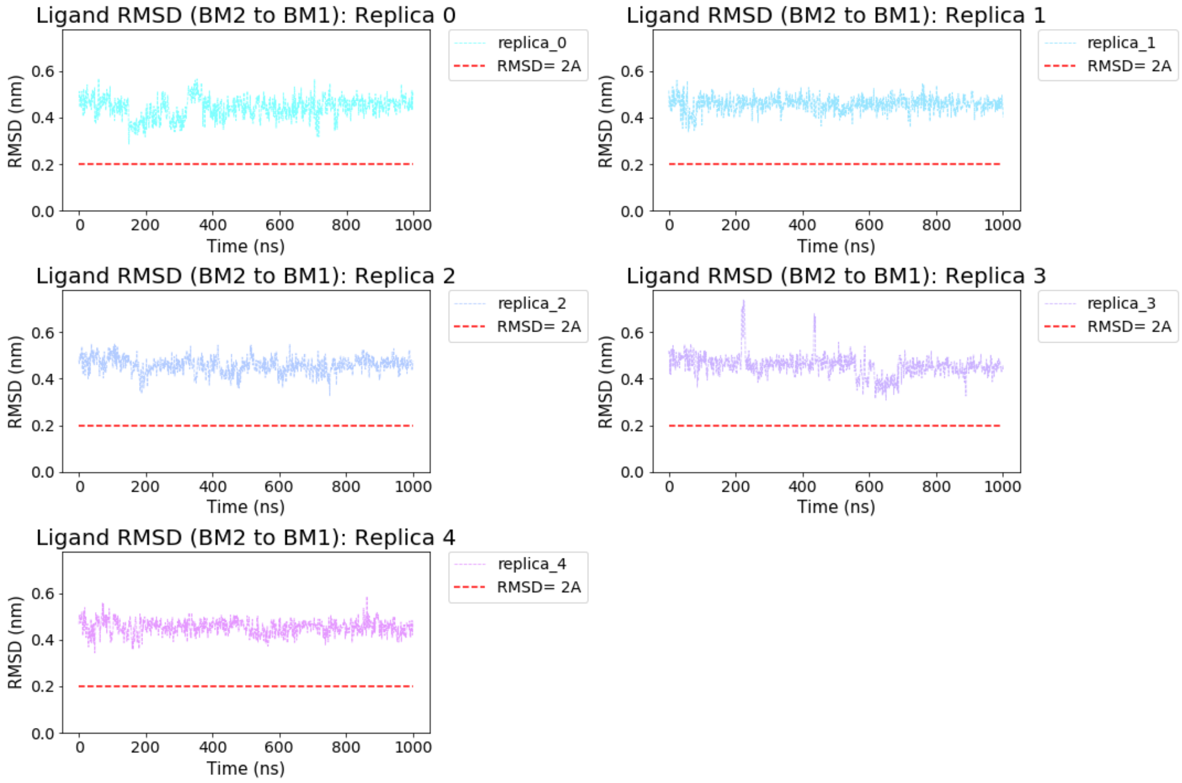


Figure 6-33 Ligand BM2 RMSD relative to BM1 conformation. A BM1-like conformation is never adopted in the replica started from BM2.

From the clustering analysis of the 10000 frames of the BM2 replicas, five clusters could be identified (Figure 6-34). The number of clusters is significantly lower than the one reported for BM1 because the behaviour of the ligand in the BM2 replicas is analogous. As already noted from the RMSD plot, the ligand portion inside the pocket is similar to the BM2 crystal structure, the solvent-exposed portion is flexible. The most populated cluster (6730 frames) shows the ligand in a BM2-like conformation with the benzodioxine ring that is pointing towards the Trp81. In cluster 3 the BM2 conformation is identified. Cluster 4 is similar to the cluster 2 of the BM1 replicas.

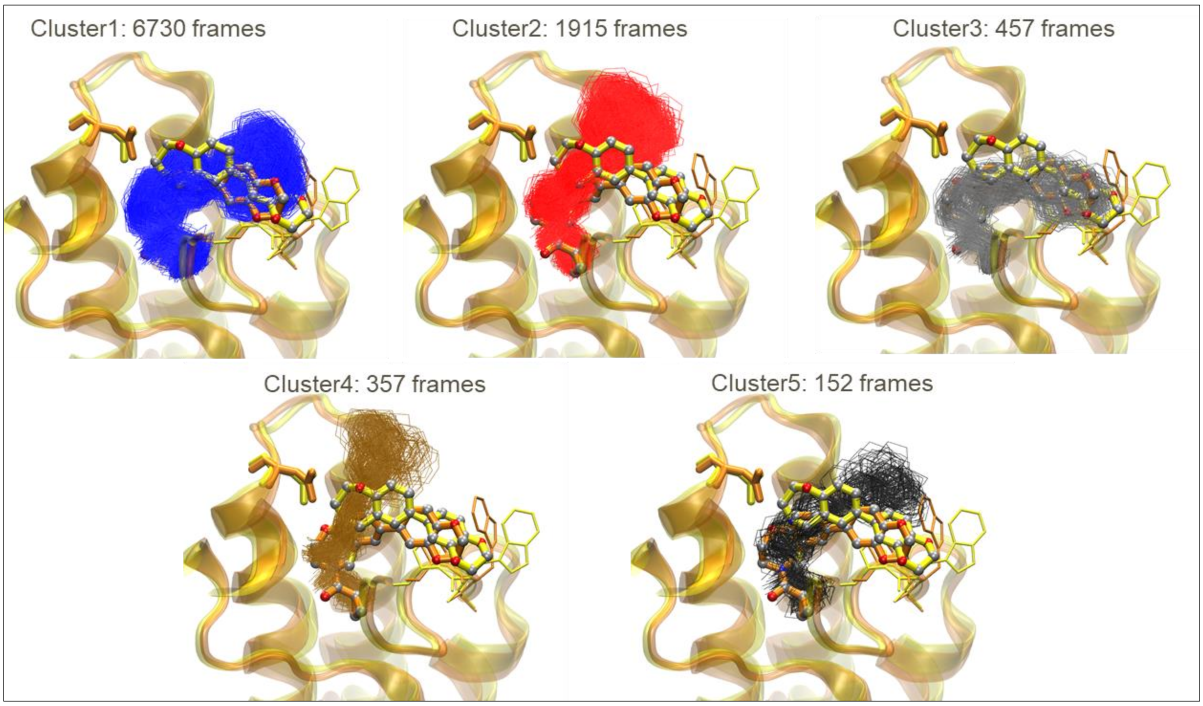


Figure 6-34 The clusters obtained from the BM2 replicas are reported in lines and overlaid to the crystal structure of BM1 (yellow) and BM2 (orange). Asn140 and the residues in the shelf (Trp81, Pro82 and Phe83) are reported in stick.

6.5.2.1.3 Conclusions

From the ten replicas started from both BM1 and BM2 a total of 10 μs were obtained. In general, **Ligand A** is less stable in BM1 than in BM2.It was observed that BM1 is moving freely in the binding pocket as opposed to the replicas started from BM2 in which the ligand is well anchored to the Asn140 via a direct hydrogen bond with O2 (Figure 6-35). In addition, in BM2 the solvent exposed portion, the benzodioxine ring, interacts with Trp81 in the lipophilic shelf and appears to be highly flexible. This might indicate that the interactions with the protein, excluding the one with Asn140, are not particularly strong, which is in line with the low potency of the ligand.

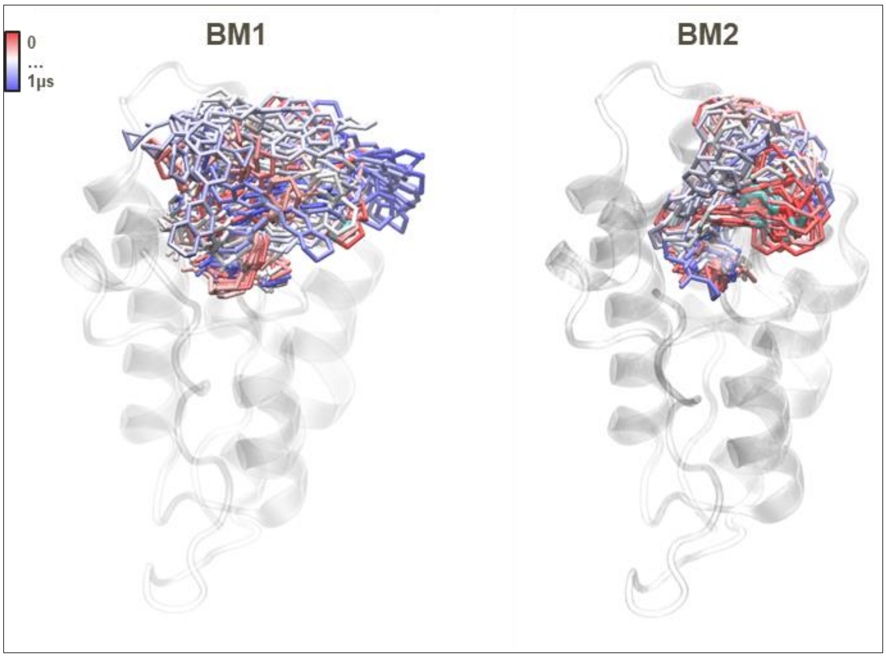


Figure 6-35 Ligand movement during the MD simulations during the 5 replicas of 1μs each of BM1 (left) and BMD2 (right). While BM1 is moving significantly in the ligand-binding pocket, the majority of the fluctuation for BM2 are related to the solvent-exposed portion.

6.5.2.2 Funnel MetaD

The last methodology used to investigate BM1 and BM2 was Funnel MetaD (FM)[195] in which a restraining potential with the shape of a funnel, is applied within the MetaD settings. In FM the ligand is able to freely explore the protein binding site and, at the same time, the presence of the funnel prevents the exploration of all the solvated states outside the binding pocket. In this way, the sampling of the binding and unbinding events should be enhanced and convergence reached in shorter time. As a result, a converged binding FES and an absolute free energy of binding can be obtained.

To quantify the relative free energy of the different poses and compute the full free energy landscape associated with the binding of the ligand, parallel tempering metadynamics simulations (PT-MetaD) with a funnel-shaped restraints potential[195] were run on BM1.

This methodology has been chosen because no a priori knowledge of the ligand binding pose within the binding site is required; if successful, it could allow an accurate description of the FES and be informative of multiple ligand binding poses in a more rigorous way compared to both docking or BPMD and ultimately, it could provide information regarding the binding/unbinding mechanism. However, the setup of such simulations is far from trivial and involves a significant amount of trial and error.

It should be noted that a limiting step of the simulation are represented by the alignment of the complex to the funnel prior to every step of the simulation and by the fact that the PLUMED code is not yet supported to run on GPU; for these reasons and because of the parallel tempering scheme adopted, the performance of the simulation drops to 7 ns/day. Given the poor performance of the simulations, it was decided to focus on only one replica of **Ligand A** initiated from BM1.

The FM simulations with PT scheme started from BM1 were run for a total of 1.03 μs using as collective variables the vector connecting the binding pocket (defined as the centre of mass of the Cα of Asp106, Met107 and Phe133) and the nitrogen N1 atom of the ligand projected onto the Z-axis (dz) and XY-plane (dxy) of the funnel. The convergence of the simulation was monitored by plotting the evolution of the FES during the course of the simulation and by monitoring the evolution of the absolute protein-ligand binding free energy. Unfortunately, as will be discussed in detail below, no sign of convergence was observed despite >1μs of simulation time. Prior to the PT-MetaD scheme, two MetaD simulations with the funnel-like restrains were run and discussed in Appendix C.

The FES as function of dz and dxy obtained at the end of the simulation is illustrated in Figure 6-36. There are several aspects that indicate the simulation did not reach convergence by just looking at the generated FES. Firstly, it is important to note that the crystallographic positions (green triangles) have high free energy. Secondly, the deepest free energy minimum does not correspond to either BM1 or BM2. Thirdly, there are a lot of other energy minima which are located outside the ligand-binding site.

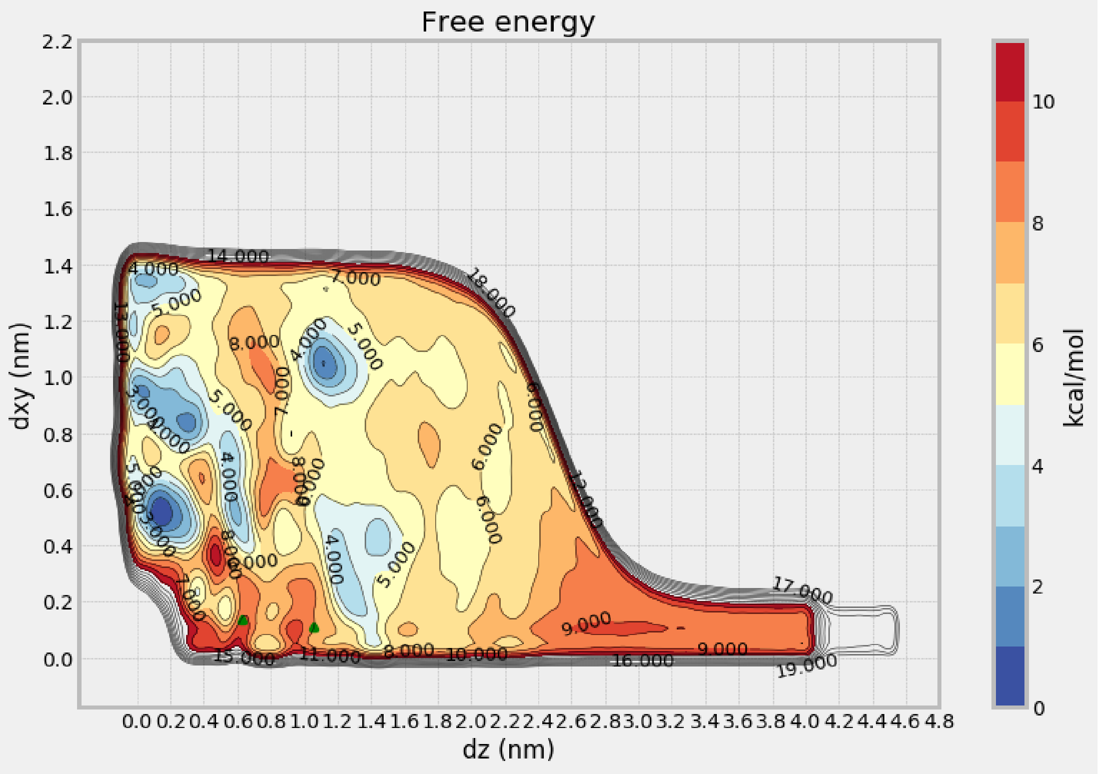


Figure 6-36 Projection of the FES using the collective variable (Z-projection and XY-projection). The green triangles are representative of BM1 and BM2.

The conformation of BM1 reported to be in the global minima is reported in Figure 6-37.

GSK CONFIDENTIAL INFORMATION – DO NOT COPY

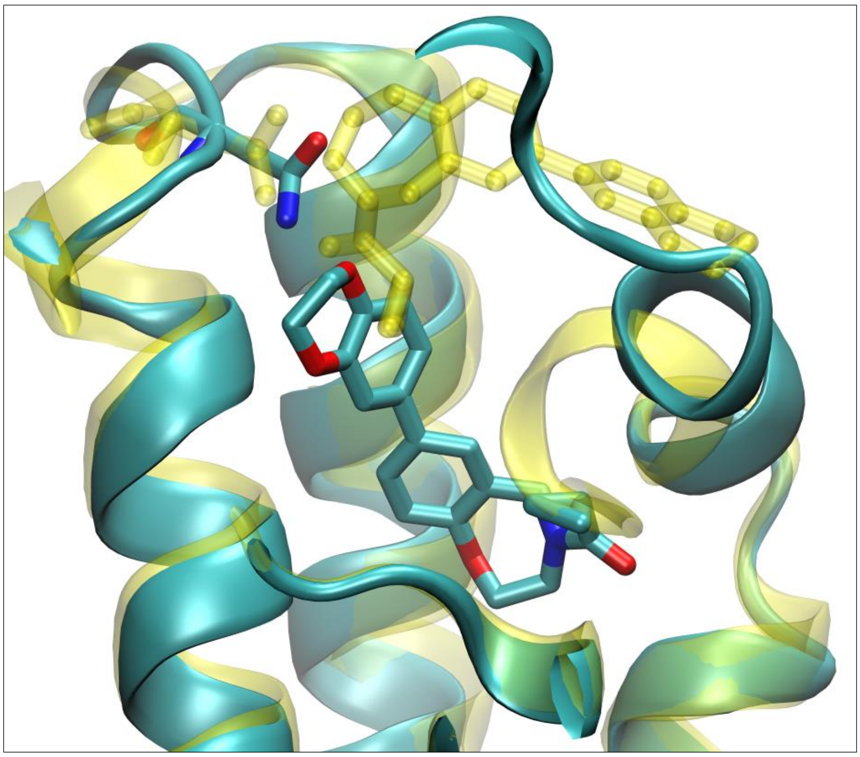


Figure 6-37 BM1 conformation in the global minima as reported in the FES overlaid to the BM1 crystal structure in yellow. The ZA loop is omitted for clarity.

From the analysis of the trajectories, it is observed that the ligand can easily escape from the binding pocket and it is interacting in other regions of the protein. Once the ligand leaves the binding pocket, the ZA loop rearranges in a closed-to-open conformation that is then kept till the end of the simulation (Figure 6-38). This conformation was observed in the absence of ligand by Kuang *et al.*[72] and was also reported by Heinzelmann *et al.*[200] and observed in the unbiased MD simulations (Section 6.4).

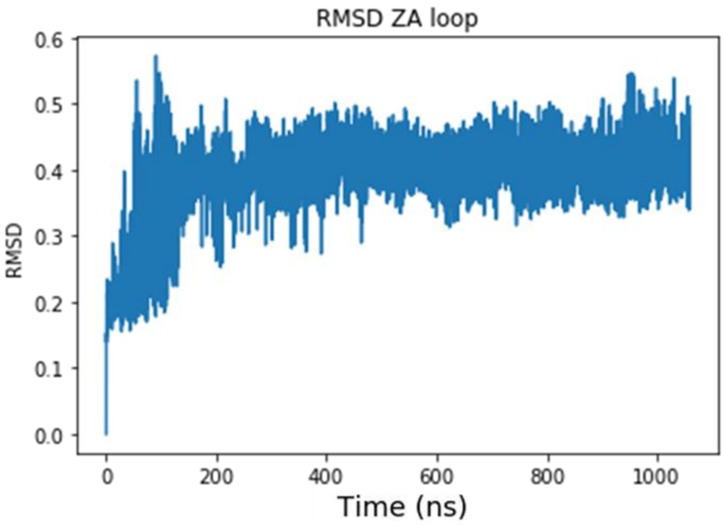


Figure 6-38 Evolution of the RMSD of the ZA loop (residue 76 to 106) during the course of the FM.

There are only a few binding/unbinding events. Once the ligand is outside the ligand-binding site, it is exploring other region of the protein surface and it goes back to the ligand-binding pocket with different conformations. This behaviour is observed when analysing the RMSD of the ligand by using either the ligand heavy atoms or the ligand center of mass (COM) as reported in Figure 6-39.

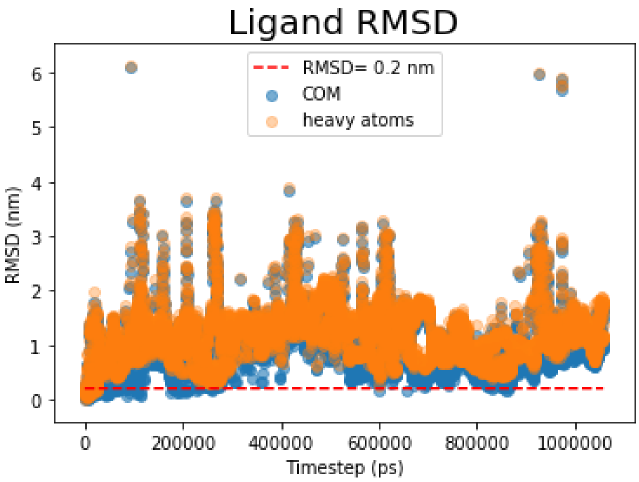


Figure 6-39 Evolution of the ligand RMSD using the COM of the ligand atoms excluding hydrogens as reference. The dashed red line indicates when the position of the COM is below 0.2 nm.

Moreover, to monitor the convergence of the absolute protein-ligand binding free energy, the free energies of the bound and unbound state were plotted with the simulation time as illustrated in Figure 6-40. The bound state was defined as the region including the dxy and dz of the two crystallographic poses, more specifically dxy between 0 nm and 0.2 nm and dz between 0.6 nm and 1.1 nm whereas the unbound state is considered in the region with dz > 3.4 nm.

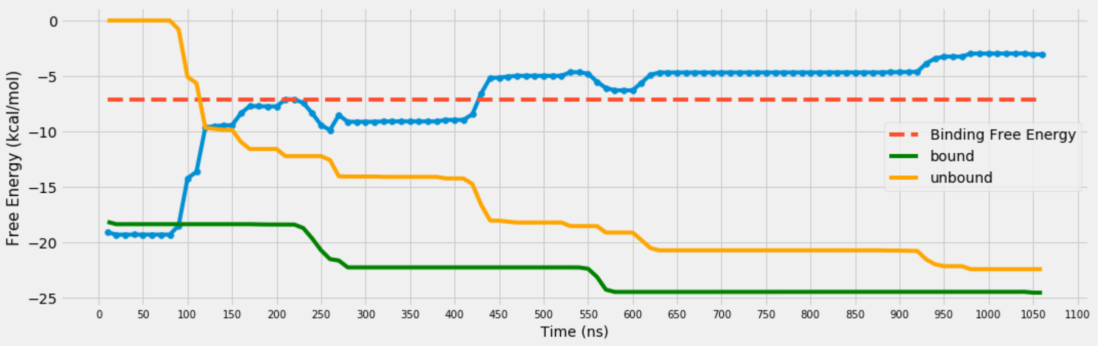


Figure 6-40 Absolute protein-ligand binding free energy (blue line) during the course of the MetaD simulation as compared to the experimental value (dashed red line). The free energy of the bound and unbound states are reported in green and orange line.

After around 600 ns the ligand is not exploring again the bound state as suggested by both the higher computed free energy with respect to the experimental value of 7.5 kcal/mol and the flat energy value of the bound state. For this reason, due to the lack of re-crossing events between different states by the system, this simulation cannot be considered to have reached convergence.

6.5.2.2.1 FM Discussion

FM has previously been shown to be a powerful method for the exploration of ligand binding events[195], but the computational cost can be significant. In this work it was not possible to obtain a converged FES even after > 1 μs of simulation. Typically, in a drug design project the time at disposal is very limited and project meeting tend to occur with a weekly basis; therefore, FM was not considered in line with the requested project-speed.

Given the available time as well as the computational resources it was not possible to investigate further the reason of such non-converged result. Nevertheless, in a different setting (e.g. more time and resources), several strategies could be undertaken to obtain a fully converged FES such as changes in the size of the funnel, exploration of different CVs (e.g. hydrogen bond between the ligand and protein).

Lastly, all of the parameters involved in the MetaD simulation: hills height, width, bias factor could also be varied. All of the proposed modifications cannot be tested without running lots of long (expensive) simulations. Therefore, for the time being, simulations with the FM which require a lot of trials and error are not yet ready to be implemented in an industry-based setting in which time and resources need to be carefully organised.

6.6 Summary and Conclusion

From the crystal structures of **Ligand A** in BRD4 two different binding modes were observed here called BM1 and BM2. In BM1, the water network at the bottom of the pocket is preserved while in BM2 the ligand is binding deeper such that the water molecules are not present. The key interaction with Asn140 is maintained in both BM1 and BM2 but the acceptor from the ligand is O1 (amide-like oxygen) in BM1 and O2 (oxygen of benzoxazepine ring) in BM2. Moreover, to accommodate **Ligand A** in BM1, Trp81 adopts a conformation that points out from the ligand binding site which was never observed in the available crystal structures from both the PDB and GSK databases.

By analysis of the static picture provided by the crystal structures, it could be hypothesised that Trp81 has a preferential conformation with χ_1~ 60° and χ_2~ -90°. However, by means of MD and MetaD simulations, it was observed that Trp81 has several minima in its free energy landscape which are separated by only ~1 or 2 kcal/mol. Therefore, it is believed that Trp81 does not have a strongly preferred orientation in solution as opposed to the preferred conformation in the crystal structures, which is influenced by the crystal packing. In light of those results, Trp81 can assume several conformations such as the ones observed in BM1 and BM2 without any significant energetic cost.

In the absence of experimental information, by using a strategy which consists of docking of **Ligand A** in the apo crystal structure of BRD4 followed by BPMD, the docking solutions with the closest RMSD to the crystal structures of BM1 and BM2 have been successfully identified as the most stable complexes. If the crystal structures of BM1 and BM2 are also scored by BPMD, the resulting *PoseScore* for both crystal structures is smaller than any docked pose; moreover, BM2 appears to be more stable than BM1 with a *PoseScore* of 1.3 and 1.7 respectively.

The ten MD replicas of 1μs each started from BM1 and BM2 are informative of the dynamics of the ligand binding pocket and of the stability of **Ligand A**. In the unbiased MD of BM1, the ligand is not as stable as in BM2: the hydrogen bond with Asn140 is lost in the majority of the replicas and the ligand is adopting several conformations, among which also BM2-like conformations are identified. The residues in the ligand binding pocket are also flexible, especially the ones located in the ZA loop. By comparison, the MD simulation started from BM2 showed that the portion of the ligand inside the binding pocket is well-anchored to Asn140. The solvent-exposed part of the ligand is flexible interacting with the lipophilic shelf and adopting several conformations. In the BM2 simulations the ligand does not assume a BM1-like conformation.

A final attempt to investigate all of the ligand binding modes of **Ligand A** was done with FM. The simulation was started with the ligand in BM1 but a final converged FES could not be obtained within a reasonable timeframe.

In conclusion, with the available computational methods, it was possible to identify both BM1 and BM2 by performing first docking of **Ligand A** in the apo X-ray crystal structure of BRD4 and then rescoring of the docking solutions with BPMD. The identification of BM1 by performing unbiased MD simulations from BM2 was not possible: **Ligand A** in BM2 was well anchored to Asn140 with O2 and no significative ligand reorganization was observed. From the unbiased MD simulations started from BM1 it was possible to identify a BM2-like ligand conformation among other conformations. In the BM1 simulations the ligand is not as stable as in BM2, in fact, several conformations are explored.

Lastly, from both IFD followed by BPMD and MD simulations it is observed that BM1 is less stable than BM2. This supports the experimental findings from which out of the 11 crystal structures only 2 were captured with a BM1 conformation

GSK CONFIDENTIAL INFORMATION – DO NOT COPY

Chapter 7 Conclusions & Future Perspective

The present work has been focused on several crucial aspects of SBDD with special focus on the BRDs protein. The water placement in the cavities of BRD4 protein was investigate with the 3D-RISM theory in Chapter 3. The failure of the Placevent algorithm in generating an optimized solution to the 3D-RISM density distribution function has been addressed with a new algorithm called GAsol. A comparison between the performance of Placevent and GAsol showed that the new algorithm is better at predicting water molecules in confined protein regions such as protein-ligand binding sites. The ability of GAsol to convert the continuous density distribution functions into explicit atoms could be further investigated by direct use of the electron density from X-ray crystallography.

Crystal structures have a central role in SBDD. The quality of the computational prediction depends also on the quality of the X-ray crystal structures used. The use of deposited crystal structures with incorrect ligand placement could lead to erroneous conclusions. The ability to identify such structures before undertaking any further studies would impact SBDD. Therefore, the use of the multiple replica MetaD simulations protocol called BPMD was validated to that purpose (Chapter 4). The great advantages of BPMD are that it allows a quantitative description of the dynamical behaviour of the ligands complexed with the protein and that it is fully automated for an industry-like setting. The results clearly show the ability of BPMD to discriminate between ligands in which the electron density supports their placement to those that are not supported. Therefore, BPMD can for example successfully guide the start of a SBDD when only one crystal structure is available and be informative on crystal structure whose ligand-binding mode is not properly modelled. The BPMD protocol could be further explored in term of simulation length, number of replicas and MetaD parameters (hills width and height) although the current protocol has already shown good performance. The enhanced sampling method was preferred to standard MD because by biasing the simulation, the time required to perturb a ligand in an unfavourable binding pose is expected to be lower. Nevertheless, it could be interesting to compare the performance of MD to the BPMD.

In Chapter 5, the design of a probe for BRD4-BD1 protein by using computational tools was presented. SAR analysis was used to generate a design strategy. MetaD simulations were used to test design hypotheses and provide valuable information that would have been missed by simple analysis of experimental data. In the attempt to rationalize the measured activity of the synthesised compounds, the effect of the water molecules and of the ligand dynamics were investigated. From the WaterMap simulations stable (negative free energy) water molecules were identified that could explain the measured potency trend in the case of BRD4-BD1 protein. At the same time, no clear conclusion could be drawn for the BD2 domain, which could reflect the limitations of the current methods for the prediction of the energetic contribution of the water molecules. In addition, several MD replicas of two newly synthesised ligands (**13** and **16**) in complex with both BD1 and BD2 were run with the intention to shed light on the origin of the selectivity. Given that multiple factors might contribute to the ligand affinity such as desolvation effect, protonation state, steric effects, kinetic rates, entropic-enthalpic compensation, it was not possible to draw a quantitative explanation on the selectivity from the MD simulations. More advanced techniques e.g. Free Energy Perturbation, or MetaD could be used to address this problem.

The key role played by crystal structures in SBDD was also investigated in Chapter 6. Proteins are not static, while X-ray structures represent a static picture, a snapshot, of the dynamic process of the ligand-protein binding. Hence, in some instances, crystal structures might not be able to fully represent the situation in solution and SAR inconsistencies could be experienced if for example protein and ligand flexibility is neglected. As an example of this problem, **Ligand A** (Figure 6-1) co-crystallized multiple times in the BRD4-BD1 protein showed the presence of multiple ligand binding modes called BM1 and BM2. This is a complex situation to investigate. In fact, not many similar cases are reported[201] but this does not imply that such phenomenon is of less importance but rather the observation of multiple ligand binding modes might happen more frequently than expected. Therefore, work was undertaken to investigate whether both binding modes could be predicted and identified with the current state-of-the-art computational tools.

One of the main differences between BM1 and BM2 is the Trp81 conformation. The tryptophan conformation in BM1 differs from the one observed in BM2 and from the available crystal structure (PDB + GSK). The reconstructed FES as a function of χ_1 and χ_2 of Trp81 via MetaD simulations revealed that in the solution-phase the global minimum of Trp81 does not correspond to the conformation observed in the majority of the crystal structures suggesting that such preferred state could be the result of the crystal packing. In addition, from the FES, Trp81 can assume multiple conformations which are not energetically prohibitive.

BPMD indicated that both ligand binding modes are stable. In addition, by using a strategy of docking followed by BPMD, not only both binding modes were reproduced but also, they were predicted by BPMD as the most stable among several different docking solutions. In this way, both ligand binding modes were correctly identified but some prior knowledge on the BRDs was needed. For example, the water molecules at the bottom of the pocket were removed in the docking experiment while in general they are retained in place. This is a limitation of the current docking procedure that does not allow to insert or delete water molecules during the calculations.

MD simulations and enhanced sampling techniques were used to evaluate if by starting from one binding mode, the second binding mode could be identified. From the MD simulations (in replicas of 5), the ligand in BM1 showed high fluctuations. It was possible to detect BM2-like conformations although other conformations were also accessed. The BM2 ligand portion inside the binding state was stable in the MD simulations while the solvent exposed portion is flexible and no changes from BM2 to BM1 were observed. A possible explanation could be that in order to observe rearrangements of BM2 to BM1, longer simulations might be necessary. Overall, the results of the MD simulations show that BM1 was consistently less stable than BM2 and that the ligand in BM2 despite maintaining in part the starting conformation, it was not establishing any strong interactions with the protein.

Funnel-shaped MetaD simulations were carried out with the aim to generate the full free energy landscape of **Ligand A** in the BRD4-BD1 protein. Although several long simulations (> 1μs) were performed, the free energy surface could not be satisfactorily converged, which meant that conclusions about ligand A in BRD4-BD1 could not be drawn. The method was found to be difficult and time-consuming to use and not yet suitable for routine applications in industrial settings. The difficulty in choosing a suitable CV with a trial and error strategy is one of the factors that limits the performance of the methodology. In addition, the alignment of the funnel to the complex at every step together with the fact that PLUMED code runs only on CPUs drastically reduces the performance of the method.

The application of several simulation-based methods to multiple drug discovery problems including, water placement, drug design and the identification of multiple ligand binding modes have been explored. It was possible to understand how those methods can assist and impact drug discovery and to identify gaps that could be addressed with further research. With the constant improvement in computer power and technologies, longer and more accurate simulations are going to be routinely obtained in a shorter time, especially in industry settings. In this way, simulation methods, together with general and robust analysis tools to extract quantitative information and derive new knowledge, are likely to play an increasingly important role in the development of new drugs.

GSK CONFIDENTIAL INFORMATION – DO NOT COPY

REFERENCES

1. Munos, B., Lessons from 60 years of pharmaceutical innovation. *Nat. Rev. Drug Discov.* **2009**, 8, 959-68.

2. Paul, S. M.; Mytelka, D. S.; Dunwiddie, C. T.; Persinger, C. C.; Munos, B. H.; Lindborg, S. R.; Schacht, A. L., How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat. Rev. Drug Discov.* **2010**, 9, 203-214.

3. Anderson, A. C., The Process of Structure-Based Drug Design. *Chem. Biol.* **2003**, 10, 787-797.

4. Lounnas, V.; Ritschel, T.; Kelder, J.; McGuire, R.; Bywater, R. P.; Foloppe, N., Current progress in Structure-Based Rational Drug Design marks a new mindset in drug discovery. *Comput Struct Biotechnol J* **2013**, 5.

5. Karplus, M.; McCammon, J. A., Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* **2002**, 9, 646-52.

6. Ferreira, L. G.; Dos Santos, R. N.; Oliva, G.; Andricopulo, A. D., Molecular docking and structure-based drug design strategies. *Molecules* **2015**, 20, 13384-421.

7. Roberts, N. A.; Martin, J. A.; Kinchington, D.; Broadhurst, A. V.; Craig, J. C.; Duncan, I. B.; Galpin, S. A.; Handa, B. K.; Kay, J.; Krohn, A.; et al., Rational design of peptide-based HIV proteinase inhibitors. *Science* **1990**, 248, 358-61.

8. Pastorekova, S.; Parkkila, S.; Pastorek, J.; Supuran, C. T., Review Article. *J. Enzyme Inhib. Med. Chem.* **2004**, 19, 199-229.

9. Zheng, H.; Handing, K. B.; Zimmerman, M. D.; Shabalin, I. G.; Almo, S. C.; Minor, W., X-ray crystallography over the past decade for novel drug discovery - where are we heading next? *Expert. Opin. Drug Discov.* **2015**, 10, 975-89.

10. Danley, D. E., Crystallization to obtain protein-ligand complexes for structure-aided drug design. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2006**, 62, 569-75.

11. Davis, A. M.; Teague, S. J.; Kleywegt, G. J., Application and limitations of X-ray crystallographic data in structure-based ligand and drug design. *Angew. Chem. Int. Ed. Engl.* **2003**, 42, 2718-36.

12. Zheng, H.; Hou, J.; Zimmerman, M. D.; Wlodawer, A.; Minor, W., The future of crystallography in drug discovery. *Expert. Opin. Drug Discov.* **2014**, 9, 125-37.

13. Fischer, E., Einfluss der Configuration auf die Wirkung der Enzyme. *Berichte der deutschen chemischen Gesellschaft* **1894**, 27, 2985-2993.

14. Koshland, D. E., Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proc. Natl. Acad. Sci. U.S.A.* **1958**, 44, 98-104.

15. Foote, J.; Milstein, C., Conformational isomerism and the diversity of antibodies. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, 91, 10370-4.

16. Stein, A.; Rueda, M.; Panjkovich, A.; Orozco, M.; Aloy, P., A Systematic Study of the Energetics Involved in Structural Changes upon Association and Connectivity in Protein Interaction Networks. *Structure* **2011**, 19, 881-889.

17. Alder, B. J.; Wainwright, T. E., Phase Transition for a Hard Sphere System. *J. Chem. Phys.* **1957**, 27, 1208-1209.

18. McCammon, J. A.; Gelin, B. R.; Karplus, M., Dynamics of folded proteins. *Nature* **1977**, 267, 585-590.

19. Lin, F.-Y.; MacKerell, A. D. Force Fields for Small Molecules. In *Biomolecular Simulations: Methods and Protocols*, Bonomi, M.; Camilloni, C., Eds.; Springer New York: New York, NY, 2019, pp 21-54.

20. Ponder, J. W.; Case, D. A. Force Fields for Protein Simulations. In *Adv. Protein Chem.*; Academic Press: 2003; Vol. 66, pp 27-85.

21. Robustelli, P.; Piana, S.; Shaw, D. E., Developing a molecular dynamics force field for both folded and disordered protein states. *Proc. Natl. Acad. Sci. U.S.A.* **2018**, 115, E4758.

22. Duan, Y.; Kollman, P. A., Pathways to a Protein Folding Intermediate Observed in a 1-Microsecond Simulation in Aqueous Solution. *Science* **1998**, 282, 740-744.

23. Shaw, D. E.; Chao, J. C.; Eastwood, M. P.; Gagliardo, J.; Grossman, J. P.; Ho, C. R.; Lerardi, D. J.; Kolossváry, I.; Klepeis, J. L.; Layman, T.; McLeavey, C.; Deneroff, M. M.; Moraes, M. A.; Mueller, R.; Priest, E. C.; Shan, Y.; Spengler, J.; Theobald, M.; Towles, B.; Wang, S. C.; Dror, R. O.; Kuskin, J. S.; Larson, R. H.; Salmon, J. K.; Young, C.; Batson, B.; Bowers, K. J., Anton, a special-purpose machine for molecular dynamics simulation. *Commun. ACM* **2008**, 51, 91.

24. Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wriggers, W., Atomic-level characterization of the structural dynamics of proteins. *Science* **2010**, 330, 341-6.

25. Harvey, M. J.; Giupponi, G.; Fabritiis, G. D., ACEMD: Accelerating Biomolecular Dynamics in the Microsecond Time Scale. *J. Chem. Theory Comput.* **2009**, 5, 1632-1639.

26. Buch, I.; Giorgino, T.; De Fabritiis, G., Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, 108, 10184.

27. Shan, Y.; Kim, E. T.; Eastwood, M. P.; Dror, R. O.; Seeliger, M. A.; Shaw, D. E., How Does a Drug Molecule Find Its Target Binding Site? *J. Am. Chem. Soc.* **2011**, 133, 9181-9183.

28. Michel, J., Current and emerging opportunities for molecular simulations in structure-based drug design. *Phys. Chem. Chem. Phys.* **2014**, 16, 4465-77.

29. Knapp, B.; Ospina, L.; Deane, C. M., Avoiding False Positive Conclusions in Molecular Simulation: The Importance of Replicas. *J. Chem. Theory Comput.* **2018**, 14, 6127-6138.

30. Ladbury, J. E., Just add water! The effect of water on the specificity of protein-ligand binding sites and its potential application to drug design. *Chem. Biol.* **1996**, 3, 973-980.

31. Homans, S. W., Water, water everywhere--except where it matters? *Drug Discov. Today* **2007**, 12, 534-9.

32. Lu, Y.; Wang, R.; Yang, C. Y.; Wang, S., Analysis of ligand-bound water molecules in high-resolution crystal structures of protein-ligand complexes. *J Chem Inf Model* **2007**, 47, 668-75.

33. Poornima, C. S.; Dean, P. M., Hydration in drug design. 1. Multiple hydrogen-bonding features of water molecules in mediating protein-ligand interactions. *J. Comput. Aided Mol. Des.* **1995**, 9, 500-512.

34. de Beer, S. B.; Vermeulen, N. P.; Oostenbrink, C., The role of water molecules in computational drug design. *Curr. Top. Med. Chem.* **2010**, 10, 55-66.

35. Spyrakis, F.; Ahmed, M. H.; Bayden, A. S.; Cozzini, P.; Mozzarelli, A.; Kellogg, G. E., The Roles of Water in the Protein Matrix: A Largely Untapped Resource for Drug Discovery. *J. Med. Chem.* **2017**.

36. Dunitz, J. D., The Entropic Cost of Bound Water in Crystals and Biomolecules. *Science* **1994**, 264, 670.

37. Stephanie, B. A. d. B.; Nico, P. E. V.; Chris, O., The Role of Water Molecules in Computational Drug Design. *Curr. Top. Med. Chem.* **2010**, 10, 55-66.

38. Sondergaard, C. R.; Garrett, A. E.; Carstensen, T.; Pollastri, G.; Nielsen, J. E., Structural artifacts in protein-ligand X-ray structures: implications for the development of docking scoring functions. *J. Med. Chem.* **2009**, 52, 5673-84.

39. Bodnarchuk, M. S., Water, water, everywhere... It's time to stop and think. *Drug Discov. Today* **2016**, 21, 1139-46.

40. Abel, R.; Young, T.; Farid, R.; Berne, B. J.; Friesner, R. A., Role of the Active-Site Solvent in the Thermodynamics of Factor Xa Ligand Binding. *J. Am. Chem. Soc.* **2008**, 130, 2817-2831.

41. Young, T.; Abel, R.; Kim, B.; Berne, B. J.; Friesner, R. A., Motifs for molecular recognition exploiting hydrophobic enclosure in protein–ligand binding. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, 104, 808-813.

42. Nguyen, C. N.; Cruz, A.; Gilson, M. K.; Kurtzman, T., Thermodynamics of Water in an Enzyme Active Site: Grid-Based Hydration Analysis of Coagulation Factor Xa. *J. Chem. Theory Comput.* **2014**, 10, 2769-2780.

43. Ross, G. A.; Bodnarchuk, M. S.; Essex, J. W., Water Sites, Networks, And Free Energies with Grand Canonical Monte Carlo. *J. Am. Chem. Soc.* **2015**, 137, 14930-43.

44. Hu, B.; Lill, M. A., WATsite: hydration site prediction program with PyMOL interface. *J. Comput. Chem.* **2014**, 35, 1255-60.

45. Chandler, D.; Andersen, H. C., Optimized cluster expansions for classical fluids. 2. Theory of molecular liquids. . *J. Chem. Phys.* **1972**, 57, 1930−1937.

46. Brodney, M. A.; Barreiro, G.; Ogilvie, K.; Hajos-Korcsok, E.; Murray, J.; Vajdos, F.; Ambroise, C.; Christoffersen, C.; Fisher, K.; Lanyon, L.; Liu, J.; Nolan, C. E.; Withka, J. M.; Borzilleri, K. A.; Efremov, I.; Oborski, C. E.; Varghese, A.; O'Neill, B. T., Spirocyclic sulfamides as beta-secretase 1 (BACE-1) inhibitors for the treatment of Alzheimer's disease: utilization of structure based drug design, WaterMap, and CNS penetration studies to identify centrally efficacious inhibitors. *J. Med. Chem.* **2012**, 55, 9224-39.

47. Horbert, R.; Pinchuk, B.; Johannes, E.; Schlosser, J.; Schmidt, D.; Cappel, D.; Totzke, F.; Schachtele, C.; Peifer, C., Optimization of potent DFG-in inhibitors of platelet derived growth factor receptorbeta (PDGF-Rbeta) guided by water thermodynamics. *J. Med. Chem.* **2015**, 58, 170-82.

48. Robinson, D.; Bertrand, T.; Carry, J. C.; Halley, F.; Karlsson, A.; Mathieu, M.; Minoux, H.; Perrin, M. A.; Robert, B.; Schio, L.; Sherman, W., Differential Water Thermodynamics Determine PI3K-Beta/Delta Selectivity for Solvent-Exposed Ligand Modifications. *J Chem Inf Model* **2016**, 56, 886-94.

49. Myrianthopoulos, V.; Kritsanida, M.; Gaboriaud-Kolar, N.; Magiatis, P.; Ferandin, Y.; Durieu, E.; Lozach, O.; Cappel, D.; Soundararajan, M.; Filippakopoulos, P.; Sherman, W.; Knapp, S.; Meijer, L.; Mikros, E.; Skaltsounis, A. L., Novel Inverse Binding Mode of Indirubin Derivatives Yields Improved Selectivity for DYRK Kinases. *ACS Med Chem Lett* **2013**, 4, 22-26.

50. Arrowsmith, C. H.; Bountra, C.; Fish, P. V.; Lee, K.; Schapira, M., Epigenetic protein families: a new frontier for drug discovery. *Nat. Rev. Drug Discov.* **2012**, 11, 384-400.

51. Marushige, K., Activation of chromatin by acetylation of histone side chains. *Proc. Natl. Acad. Sci. U.S.A.* **1976**, 73, 3937.

52. Zhao, D.; Li, Y.; Xiong, X.; Chen, Z.; Li, H., YEATS Domain-A Histone Acylation Reader in Health and Disease. *J. Mol. Biol.* **2017**, 429, 1994-2002.

53. Haynes, S. R.; Dollard, C.; Winston, F.; Beck, S.; Trowsdale, J.; Dawid, I. B., The bromodomain: a conserved sequence found in human, Drosophila and yeast proteins. *Nucleic Acids Res.* **1992**, 20, 2603-2603.

54. Owen, D. J.; Ornaghi, P.; Yang, J. C.; Lowe, N.; Evans, P. R.; Ballario, P.; Neuhaus, D.; Filetici, P.; Travers, A. A., The structural basis for the recognition of acetylated histone H4 by the bromodomain of histone acetyltransferase Gcn5p. *EMBO J.* **2000**, 19, 6141.

55. Sanchez, R.; Meslamani, J.; Zhou, M. M., The bromodomain: from epigenome reader to druggable target. *Biochim. Biophys. Acta* **2014**, 1839, 676-85.

56. Filippakopoulos, P.; Picaud, S.; Mangos, M.; Keates, T.; Lambert, J. P.; Barsyte-Lovejoy, D.; Felletar, I.; Volkmer, R.; Muller, S.; Pawson, T.; Gingras, A. C.; Arrowsmith, C. H.; Knapp, S., Histone recognition and large-scale structural analysis of the human bromodomain family. *Cell* **2012**, 149, 214-31.

57. Padmanabhan, B.; Mathur, S.; Manjula, R.; Tripathi, S., Bromodomain and extra-terminal (BET) family proteins: New therapeutic targets in major diseases. *J. Biosci. (Bangalore)* **2016**, 41, 295-311.

58. Fu, L. L.; Tian, M.; Li, X.; Li, J. J.; Huang, J.; Ouyang, L.; Zhang, Y.; Liu, B., Inhibition of BET bromodomains as a therapeutic strategy for cancer drug discovery. *Oncotarget* 6, 5501-5516.

59. Barbieri, I.; Cannizzaro, E.; Dawson, M. A., Bromodomains as therapeutic targets in cancer. *Brief Funct Genomics* **2013**, 12, 219-30.

60. Vidler, L. R.; Brown, N.; Knapp, S.; Hoelder, S., Druggability analysis and structural classification of bromodomain acetyl-lysine binding sites. *J. Med. Chem.* **2012**, 55, 7346-59.

61. Hewings, D. S.; Rooney, T. P.; Jennings, L. E.; Hay, D. A.; Schofield, C. J.; Brennan, P. E.; Knapp, S.; Conway, S. J., Progress in the development and application of small molecule inhibitors of bromodomain-acetyl-lysine interactions. *J. Med. Chem.* **2012**, 55, 9393-413.

62. Jeanmougin, F.; Wurtz, J. M.; Le Douarin, B.; Chambon, P.; Losson, R., The bromodomain revisited. *Trends Biochem. Sci* **1997**, 22, 151-153.

63. Crawford, T. D.; Tsui, V.; Flynn, E. M.; Wang, S.; Taylor, A. M.; Cote, A.; Audia, J. E.; Beresini, M. H.; Burdick, D. J.; Cummings, R.; Dakin, L. A.; Duplessis, M.; Good, A. C.; Hewitt, M. C.; Huang, H. R.; Jayaram, H.; Kiefer, J. R.; Jiang, Y.; Murray, J.; Nasveschuk, C. G.; Pardo, E.; Poy, F.; Romero, F. A.; Tang, Y.; Wang, J.; Xu, Z.; Zawadzke, L. E.; Zhu, X.; Albrecht, B. K.; Magnuson, S. R.; Bellon, S.; Cochran, A. G., Diving into the Water: Inducible Binding Conformations for BRD4, TAF1(2), BRD9, and CECR2 Bromodomains. *J. Med. Chem.* **2016**, 59, 5391-402.

64. Prinjha, R. K.; Witherington, J.; Lee, K., Place your BETs: the therapeutic potential of bromodomains. *Trends Pharmacol. Sci.* **2012**, 33, 146-153.

65. Steiner, S.; Magno, A.; Huang, D.; Caflisch, A., Does bromodomain flexibility influence histone recognition? *FEBS Lett.* **2013**, 587, 2158-2163.

66. Magno, A.; Steiner, S.; Caflisch, A., Mechanism and Kinetics of Acetyl-Lysine Binding to Bromodomains. *J. Chem. Theory Comput.* **2013**, 9, 4225-32.

67. Jennings, L. E.; Schiedel, M.; Hewings, D. S.; Picaud, S.; Laurin, C. M. C.; Bruno, P. A.; Bluck, J. P.; Scorah, A. R.; See, L.; Reynolds, J. K.; Moroglu, M.; Mistry, I. N.; Hicks, A.; Guzanov, P.; Clayton, J.; Evans, C. N. G.; Stazi, G.; Biggin, P. C.; Mapp, A. K.; Hammond, E. M.; Humphreys, P. G.; Filippakopoulos, P.; Conway, S. J., BET bromodomain ligands: Probing the WPF shelf to improve BRD4 bromodomain affinity and metabolic stability. *Biorg. Med. Chem.* **2018**, 26, 2937-2957.

68. Genheden, S.; Ryde, U., The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert. Opin. Drug Discov.* **2015**, 10, 449-461.

69. Wang, Q.; Li, Y.; Xu, J.; Wang, Y.; Leung, E. L.; Liu, L.; Yao, X., Selective inhibition mechanism of RVX-208 to the second bromodomain of bromo and extraterminal proteins: insight from microsecond molecular dynamics simulations. *Sci. Rep.* **2017**, 7, 8857.

70. Cheng, C.; Diao, H.; Zhang, F.; Wang, Y.; Wang, K.; Wu, R., Deciphering the mechanisms of selective inhibition for the tandem BD1/BD2 in the BET-bromodomain family. *Phys. Chem. Chem. Phys.* **2017**, 19, 23934-23941.

71. Muvva, C.; Singam, E. R.; Raman, S. S.; Subramanian, V., Structure-based virtual screening of novel, high-affinity BRD4 inhibitors. *Mol. Biosyst.* **2014**, 10, 2384-97.

72. Kuang, M.; Zhou, J.; Wang, L.; Liu, Z.; Guo, J.; Wu, R., Binding Kinetics versus Affinities in BRD4 Inhibition. *J Chem Inf Model* **2015**, 55, 1926-35.

73. Aldeghi, M.; Heifetz, A.; Bodkin, M. J.; Knapp, S.; Biggin, P. C., Accurate calculation of the absolute free energy of binding for drug molecules. *Chem Sci* **2016**, 7, 207-218.

74. Wan, S.; Bhati, A. P.; Zasada, S. J.; Wall, I.; Green, D.; Bamborough, P.; Coveney, P. V., Rapid and Reliable Binding Affinity Prediction of Bromodomain Inhibitors: A Computational Study. *J. Chem. Theory Comput.* **2017**, 13, 784-795.

75. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A., A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, 117, 5179-5197.

76. MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M., All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, 102, 3586-3616.

77. Jorgensen, W. L.; Tirado-Rives, J., The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.* **1988**, 110, 1657-1666.

78. Oostenbrink, C.; Villa, A.; Mark, A. E.; Gunsteren, W. F. V., A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6. *J. Comput. Chem.* **2004**, 25, 1656-1676.

79. Leach Andrew, R.; Leach, A. R., *Molecular Modelling: Principles and Applications*. Prentice Hall: Harlow, UK.

80. Frenkel, D.; Frenkel, D., *Understanding Molecular Simulation: From Algorithms to Applications*. Academic: San Diego, Calif.

81. Verlet, L., Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Phys. Rev.* **1967**, 159, 98-103.

82. Swope, W. C.; Andersen, H. C.; Berens, P. H.; Wilson, K. R., A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *J. Chem. Phys.* **1982**, 76, 637-649.

83. Hockney, R. W., The potential calculation and some applications. *Methods Comput. Phys.* **1970**, 9, 136.

84. Evans, D. J.; Holian, B. L., The Nose–Hoover thermostat. *J. Chem. Phys.* **1985**, 83, 4069-4074.

85. Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R., Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, 81, 3684-3690.

86. Darden, T.; York, D.; Pedersen, L., Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, 98, 10089-10092.

87. Senftle, T. P.; Hong, S.; Islam, M. M.; Kylasa, S. B.; Zheng, Y.; Shin, Y. K.; Junkermeier, C.; Engel-Herbert, R.; Janik, M. J.; Aktulga, H. M.; Verstraelen, T.; Grama, A.; van Duin, A. C. T., The ReaxFF reactive force-field: development, applications and future directions. *npj Computational Materials* **2016**, 2, 15011.

88. Warshel, A.; Levitt, M., Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J. Mol. Biol.* **1976**, 103, 227-249.

89. Lane, T. J.; Shukla, D.; Beauchamp, K. A.; Pande, V. S., To milliseconds and beyond: challenges in the simulation of protein folding. *Curr. Opin. Struct. Biol.* **2013**, 23, 58-65.

90. Henzler-Wildman, K.; Kern, D., Dynamic personalities of proteins. *Nature* **2007**, 450, 964-972.

91. Laio, A.; Parrinello, M., Escaping free-energy minima. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, 99, 12562-6.

92. Barducci, A.; Bussi, G.; Parrinello, M., Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. *Phys. Rev. Lett.* **2008**, 100, 020603.

93. Yuriev, E.; Agostino, M.; Ramsland, P. A., Challenges and advances in computational docking: 2009 in review. *J Mol Recognit* **2011**, 24, 149-64.

94. Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S., Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, 47, 1739-1749.

95. Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L., Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. *J. Med. Chem.* **2004**, 47, 1750-1759.

96. *Schrödinger Release 2018-4: ConfGen, Schrödinger, LLC, New York, NY, 2018*, 2018.

97. Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P., Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput. Aided Mol. Des.* **1997**, 11, 425-445.

98. Farid, R.; Day, T.; Friesner, R. A.; Pearlstein, R. A., New insights about HERG blockade obtained from protein modeling, potential energy mapping, and docking studies. *Biorg. Med. Chem.* **2006**, 14, 3160-3173.

99. Sherman, W.; Beard, H. S.; Farid, R., Use of an Induced Fit Receptor Structure in Virtual Screening. *Chem. Biol. Drug Des.* **2006**, 67, 83-84.

100. Sherman, W.; Day, T.; Jacobson, M. P.; Friesner, R. A.; Farid, R., Novel Procedure for Modeling Ligand/Receptor Induced Fit Effects. *J. Med. Chem.* **2006**, 49, 534-553.

101. Andersen, H. C.; Chandler, D., Optimized Cluster Expansions for Classical Fluids. I. General Theory and Variational Formulation of the Mean Spherical Model and Hard Sphere Percus-Yevick Equations. *J. Chem. Phys.* **1972**, 57, 1918-1929.

102. Chandler, D.; Andersen, H. C., Optimized Cluster Expansions for Classical Fluids. II. Theory of Molecular Liquids. *J. Chem. Phys.* **1972**, 57, 1930-1937.

103. Chandler, D., Cluster diagrammatic analysis of the RISM equation. *Mol. Phys.* **1976**, 31, 1213-1223.

104. Chandler, D.; Pratt, L. R., Statistical mechanics of chemical equilibria and intramolecular structures of nonrigid molecules in condensed phases. *J. Chem. Phys.* **1976**, 65, 2925-2940.

105. Chandler, D., The dielectric constant and related equilibrium properties of molecular fluids: Interaction site cluster theory analysis. *J. Chem. Phys.* **1977**, 67, 1113-1124.

106. Chandler, D., Structures of Molecular Liquids. *Annu. Rev. Phys. Chem.* **1978**, 29, 441-471.

107. Beglov, D.; Roux, B., An Integral Equation To Describe the Solvation of Polar Molecules in Liquid Water. *J. Phys. Chem. B* **1997**, 101, 7821-7826.

108. Imai, T.; Kovalenko, A.; Hirata, F., Hydration structure, thermodynamics, and functions of protein studied by the 3D-RISM theory. *Mol. Simulat.* **2006**, 32, 817-824.

109. Misin, M.; Fedorov, M. V.; Palmer, D. S., Hydration Free Energies of Molecular Ions from Theory and Simulation. *J. Phys. Chem. B* **2016**, 120, 975-83.

110. Nikolic, D.; Blinov, N.; Wishart, D.; Kovalenko, A., 3D-RISM-Dock: A New Fragment-Based Drug Design Protocol. *J. Chem. Theory Comput.* **2012**, 8, 3356-72.

111. Palmer, D. S.; Misin, M.; Fedorov, M. V.; Llinas, A., Fast and General Method To Predict the Physicochemical Properties of Druglike Molecules Using the Integral Equation Theory of Molecular Liquids. *Mol. Pharm.* **2015**, 12, 3420-32.

112. Genheden, S.; Luchko, T.; Gusarov, S.; Kovalenko, A.; Ryde, U., An MM/3D-RISM Approach for Ligand Binding Affinities. *J. Phys. Chem. B* **2010**, 114, 8505-8516.

113. Ratkova, E. L.; Palmer, D. S.; Fedorov, M. V., Solvation thermodynamics of organic molecules by the molecular integral equation theory: approaching chemical accuracy. *Chem. Rev.* **2015**, 115, 6312-56.

114. Sindhikara, D. J.; Yoshida, N.; Hirata, F., Placevent: An algorithm for prediction of explicit solvent atom distribution—Application to HIV-1 protease and F-ATP synthase. *J. Comput. Chem.* **2012**, 33, 1536-1543.

115. Lazaridis, T., Inhomogeneous Fluid Approach to Solvation Thermodynamics. 1. Theory. *J. Phys. Chem. B* **1998**, 102, 3531-3541.

116. Lazaridis, T., Inhomogeneous Fluid Approach to Solvation Thermodynamics. 2. Applications to Simple Fluids. *J. Phys. Chem. B* **1998**, 102, 3542-3550.

117. Beuming, T.; Che, Y.; Abel, R.; Kim, B.; Shanmugasundaram, V.; Sherman, W., Thermodynamic analysis of water molecules at the surface of proteins and applications to binding site prediction and characterization. *Proteins* **2012**, 80, 871-83.

118. Amadasi, A.; Surface, J. A.; Spyrakis, F.; Cozzini, P.; Mozzarelli, A.; Kellogg, G. E., Robust Classification of "Relevant" Water Molecules in Putative Protein Binding Sites. *J. Med. Chem.* **2008**, 51, 1063-1067.

119. Ross, G. A.; Morris, G. M.; Biggin, P. C., Rapid and accurate prediction and scoring of water molecules in protein binding sites. *PLoS One* **2012**, 7, e32036.

120. D.A. Case, R. M. B., D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, T.J. Giese, H. Gohlke,; A.W. Goetz, N. H., S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T.S. Lee, S. LeGrand, P. Li, C.; Lin, T. L., R. Luo, B. Madej, D. Mermelstein, K.M. Merz, G. Monard, H. Nguyen, H.T. Nguyen, I.; Omelyan, A. O., D.R. Roe, A. Roitberg, C. Sagui, C.L. Simmerling, W.M. Botello-Smith, J. Swails,; R.C. Walker, J. W., R.M. Wolf, X. Wu, L. Xiao and P.A. Kollman, *AMBER 2016, University of California, San Francisco.* 2016.

121. Dolinsky, T. J.; Czodrowski, P.; Li, H.; Nielsen, J. E.; Jensen, J. H.; Klebe, G.; Baker, N. A., PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Res.* **2007**, 35, W522-W525.

122. Sousa da Silva, A. W.; Vranken, W. F., ACPYPE - AnteChamber PYthon Parser interfacE. *BMC Res Notes* **2012**, 5, 367.

123. Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J., The Amber biomolecular simulation programs. *J. Comput. Chem.* **2005**, 26, 1668-1688.

124. Luchko, T.; Gusarov, S.; Roe, D. R.; Simmerling, C.; Case, D. A.; Tuszynski, J.; Kovalenko, A., Three-Dimensional Molecular Theory of Solvation Coupled with Molecular Dynamics in Amber. *J. Chem. Theory Comput.* **2010**, 6, 607-624.

125. Liebeschuetz, J.; Hennemann, J.; Olsson, T.; Groom, C. R., The good, the bad and the twisted: a survey of ligand geometry in protein crystal structures. *J. Comput. Aided Mol. Des.* **2012**, 26, 169-183.

126. Pozharski, E.; Weichenberger, C. X.; Rupp, B., Techniques, tools and best practices for ligand electron-density analysis and results from their application to deposited crystal structures. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2013**, 69, 150-167.

127. Reynolds, C. H. Protein-ligand cocrystal structures: we can do better. ACS medicinal chemistry letters, 5, 727-729, 2014.

128. Deller, M. C.; Rupp, B., Models of protein-ligand crystal structures: trust, but verify. *J. Comput. Aided Mol. Des.* **2015**, 29, 817-36.

129. Smart, O. S.; Horský, V.; Gore, S.; Svobodová Vařeková, R.; Bendová, V.; Kleywegt, G. J.; Velankar, S., Validation of ligands in macromolecular structures determined by X-ray crystallography. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2018**, 74, 228-236.

130. Carlomagno, T.; Blommers, M. J.; Meiler, J.; Jahnke, W.; Schupp, T.; Petersen, F.; Schinzer, D.; Altmann, K. H.; Griesinger, C., The high-resolution solution structure of epothilone A bound to tubulin: an understanding of the structure-activity relationships for a powerful class of antitumor agents. *Angew. Chem. Int. Ed. Engl.* **2003**, 42, 2511-5.

131. Forli, S., Epothilones: From discovery to clinical trials. *Curr. Top. Med. Chem.* **2014**, 14, 2312-2321.

132. Nettles, J. H.; Li, H.; Cornett, B.; Krahn, J. M.; Snyder, J. P.; Downing, K. H., The Binding Mode of Epothilone A on α,ß-Tubulin by Electron Crystallography. *Science* **2004**, 305, 866.

133. Prota, A. E.; Bargsten, K.; Zurwerra, D.; Field, J. J.; Diaz, J. F.; Altmann, K. H.; Steinmetz, M. O., Molecular mechanism of action of microtubule-stabilizing anticancer agents. *Science* **2013**, 339, 587-90.

134. Navratna, V.; Nadig, S.; Sood, V.; Prasad, K.; Arakere, G.; Gopal, B., Molecular basis for the role of Staphylococcus aureus penicillin binding protein 4 in antimicrobial resistance. *J. Bacteriol.* **2010**, 192, 134-144.

135. Xu, G.; Lo, Y.-C.; Li, Q.; Napolitano, G.; Wu, X.; Jiang, X.; Dreano, M.; Karin, M.; Wu, H., Crystal structure of inhibitor of κB kinase β. *Nature* **2011**, 472, 325.

136. Bränd´en, C.-I.; Alwyn Jones, T., Between objectivity and subjectivity. *Nature* **1990**, 343, 687-689.

137. Jones, T. A.; Zou, J. Y.; Cowan, S. W.; Kjeldgaard, M., Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr., Sect. A: Found. Crystallogr.* **1991**, 47, 110-119.

138. Weichenberger, C. X.; Pozharski, E.; Rupp, B., Visualizing ligand molecules in Twilight electron density. *Acta Cryst., Sect. F: Struct. Biol. Crystal. Comm.* **2013**, 69, 195-200.

139. Emsley, P.; Lohkamp, B.; Scott, W. G.; Cowtan, K., Features and development of Coot. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2010**, 66, 486-501.

140. Terwilliger, T., Using prime-and-switch phasing to reduce model bias in molecular replacement. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2004**, 60, 2144-2149.

141. Murshudov, G. N.; Skubák, P.; Lebedev, A. A.; Pannu, N. S.; Steiner, R. A.; Nicholls, R. A.; Winn, M. D.; Long, F.; Vagin, A. A., REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2011**, 67, 355-367.

142. Madhavi Sastry, G.; Adzhigirey, M.; Day, T.; Annabhimoju, R.; Sherman, W., Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *J. Comput. Aided Mol. Des.* **2013**, 27, 221-234.

143. Schrödinger Release 2018-4: Maestro, S., LLC, New York, NY, 2018.

144. Harder, E.; Damm, W.; Maple, J.; Wu, C.; Reboul, M.; Xiang, J. Y.; Wang, L.; Lupyan, D.; Dahlgren, M. K.; Knight, J. L.; Kaus, J. W.; Cerutti, D. S.; Krilov, G.; Jorgensen, W. L.; Abel, R.; Friesner, R. A., OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. *J. Chem. Theory Comput.* **2016**, 12, 281-96.

145. Clark, A. J.; Tiwary, P.; Borrelli, K.; Feng, S.; Miller, E. B.; Abel, R.; Friesner, R. A.; Berne, B. J., Prediction of Protein-Ligand Binding Poses via a Combination of Induced Fit Docking and Metadynamics Simulations. *J. Chem. Theory Comput.* **2016**, 12, 2990-8.

146. Bowers, A. K. J.; Chow, E.; Xu, H.; Dror, R. O.; Eastwood, M. P.; Gregersen, B. A.; Klepeis, J. L.; Kolossvary, I.; Moraes, M. A.; ., F. D. S.; et al., In *Proceedings of the 2006 ACM/IEEE conference on Supercomputing*; Association for Computing Machinery: Tampa, Florida, 2006, pp 84–es.

147. Weichenberger, C. X.; Afonine, P. V.; Kantardjieff, K.; Rupp, B., The solvent component of macromolecular crystals. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2015**, 71, 1023-1038.

148. Reynolds, C. H., Protein–Ligand Cocrystal Structures: We Can Do Better. *ACS Medicinal Chemistry Letters* **2014**, 5, 727-729.

149. Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E. M.; Mittal, J.; Feig, M.; MacKerell, A. D., Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting

Improved Sampling of the Backbone ϕ, ψ and Side-Chain χ1 and χ2 Dihedral Angles. *J. Chem. Theory Comput.* **2012**, 8, 3257-3273.

150. Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C., ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **2015**, 11, 3696-3713.

151. Harder, E.; Damm, W.; Maple, J.; Wu, C.; Reboul, M.; Xiang, J. Y.; Wang, L.; Lupyan, D.; Dahlgren, M. K.; Knight, J. L.; Kaus, J. W.; Cerutti, D. S.; Krilov, G.; Jorgensen, W. L.; Abel, R.; Friesner, R. A., OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. *Journal of Chemical Theory and Computation* **2016**, 12, 281-296.

152. Kettle, J. G.; Ballard, P.; Bardelle, C.; Cockerill, M.; Colclough, N.; Critchlow, S. E.; Debreczeni, J.; Fairley, G.; Fillery, S.; Graham, M. A.; Goodwin, L.; Guichard, S.; Hudson, K.; Ward, R. A.; Whittaker, D., Discovery and Optimization of a Novel Series of Dyrk1B Kinase Inhibitors To Explore a MEK Resistance Hypothesis. *J. Med. Chem.* **2015**, 58, 2834-2844.

153. Bender, A. T.; Gardberg, A.; Pereira, A.; Johnson, T.; Wu, Y.; Grenningloh, R.; Head, J.; Morandi, F.; Haselmayer, P.; Liu-Bujalski, L., Ability of Bruton's Tyrosine Kinase Inhibitors to Sequester Y551 and Prevent Phosphorylation Determines Potency for Inhibition of Fc Receptor but not B-Cell Receptor Signaling. *Mol. Pharmacol.* **2017**, 91, 208-219.

154. Yun, C.-H.; Boggon, T. J.; Li, Y.; Woo, M. S.; Greulich, H.; Meyerson, M.; Eck, M. J., Structures of Lung Cancer-Derived EGFR Mutants and Inhibitor Complexes: Mechanism of Activation and Insights into Differential Inhibitor Sensitivity. *Cancer Cell* **2007**, 11, 217-227.

155. Lietha, D.; Eck, M. J., Crystal structures of the FAK kinase in complex with TAE226 and related bis-anilino pyrimidine inhibitors reveal a helical DFG conformation. *PLoS One* **2008**, 3, e3800.

156. Estébanez-Perpiñá, E.; Arnold, L. A.; Nguyen, P.; Rodrigues, E. D.; Mar, E.; Bateman, R.; Pallai, P.; Shokat, K. M.; Baxter, J. D.; Guy, R. K.; Webb, P.; Fletterick, R. J., A surface on the androgen receptor that allosterically regulates coactivator binding. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, 104, 16074.

157. Filippakopoulos, P.; Picaud, S.; Mangos, M.; Keates, T.; Lambert, J.-P.; Barsyte-Lovejoy, D.; Felletar, I.; Volkmer, R.; Müller, S.; Pawson, T.; Gingras, A.-C.; Arrowsmith, Cheryl H.; Knapp, S., Histone Recognition and Large-Scale Structural Analysis of the Human Bromodomain Family. *Cell* **2012**, 149, 214-231.

158. Mahindroo, N.; Peng, Y.-H.; Lin, C.-H.; Tan, U.-K.; Prakash, E.; Lien, T.-W.; Lu, I. L.; Lee, H.-J.; Hsu, J. T.-A.; Chen, X.; Liao, C.-C.; Lyu, P.-C.; Chao, Y.-S.; Wu, S.-Y.; Hsieh, H.-P., Structural Basis for the Structure−Activity Relationships of Peroxisome Proliferator-Activated Receptor Agonists. *J. Med. Chem.* **2006**, 49, 6421-6424.

159. Mahindroo, N.; Wang, C.-C.; Liao, C.-C.; Huang, C.-F.; Lu, I. L.; Lien, T.-W.; Peng, Y.-H.; Huang, W.-J.; Lin, Y.-T.; Hsu, M.-C.; Lin, C.-H.; Tsai, C.-H.; Hsu, J. T. A.; Chen, X.; Lyu, P.-C.; Chao, Y.-S.; Wu, S.-Y.; Hsieh, H.-P., Indol-1-yl Acetic Acids as

Peroxisome Proliferator-Activated Receptor Agonists: Design, Synthesis, Structural Biology, and Molecular Docking Studies. *J. Med. Chem.* **2006**, 49, 1212-1216.

160. Jang, J. Y.; Koh, M.; Bae, H.; An, D. R.; Im, H. N.; Kim, H. S.; Yoon, J. Y.; Yoon, H.-J.; Han, B. W.; Park, S. B.; Suh, S. W., Structural basis for differential activities of enantiomeric PPARγ agonists: Binding of S35 to the alternate site. *Biochim. Biophys. Acta* **2017**, 1865, 674-681.

161. Filippakopoulos, P.; Qi, J.; Picaud, S.; Shen, Y.; Smith, W. B.; Fedorov, O.; Morse, E. M.; Keates, T.; Hickman, T. T.; Felletar, I.; Philpott, M.; Munro, S.; McKeown, M. R.; Wang, Y.; Christie, A. L.; West, N.; Cameron, M. J.; Schwartz, B.; Heightman, T. D.; La Thangue, N.; French, C. A.; Wiest, O.; Kung, A. L.; Knapp, S.; Bradner, J. E., Selective inhibition of BET bromodomains. *Nature* **2010**, 468, 1067-73.

162. Nicodeme, E.; Jeffrey, K. L.; Schaefer, U.; Beinke, S.; Dewell, S.; Chung, C. W.; Chandwani, R.; Marazzi, I.; Wilson, P.; Coste, H.; White, J.; Kirilovsky, J.; Rice, C. M.; Lora, J. M.; Prinjha, R. K.; Lee, K.; Tarakhovsky, A., Suppression of inflammation by a synthetic histone mimic. *Nature* **2010**, 468, 1119-23.

163. Romero, F. A.; Taylor, A. M.; Crawford, T. D.; Tsui, V.; Cote, A.; Magnuson, S., Disrupting Acetyl-Lysine Recognition: Progress in the Development of Bromodomain Inhibitors. *J. Med. Chem.* **2016**, 59, 1271-98.

164. Smith, S. G.; Zhou, M. M., The Bromodomain: A New Target in Emerging Epigenetic Medicine. *ACS Chem Biol* **2016**, 11, 598-608.

165. Filippakopoulos, P.; Knapp, S., Targeting bromodomains: epigenetic readers of lysine acetylation. *Nat. Rev. Drug Discov.* **2014**, 13, 337-56.

166. Theodoulou, N. H.; Tomkinson, N. C.; Prinjha, R. K.; Humphreys, P. G., Clinical progress and pharmacology of small molecule bromodomain inhibitors. *Curr. Opin. Chem. Biol.* **2016**, 33, 58-66.

167. Cochran, A. G.; Conery, A. R.; Sims, R. J., 3rd, Bromodomains: a new target class for drug development. *Nat. Rev. Drug Discov.* **2019**, 18, 609-628.

168. Chung, C. W.; Coste, H.; White, J. H.; Mirguet, O.; Wilde, J.; Gosmini, R. L.; Delves, C.; Magny, S. M.; Woodward, R.; Hughes, S. A.; Boursier, E. V.; Flynn, H.; Bouillot, A. M.; Bamborough, P.; Brusq, J. M.; Gellibert, F. J.; Jones, E. J.; Riou, A. M.; Homes, P.; Martin, S. L.; Uings, I. J.; Toum, J.; Clement, C. A.; Boullay, A. B.; Grimley, R. L.; Blandel, F. M.; Prinjha, R. K.; Lee, K.; Kirilovsky, J.; Nicodeme, E., Discovery and characterization of small molecule inhibitors of the BET family bromodomains. *J. Med. Chem.* **2011**, 54, 3827-38.

169. Galdeano, C.; Ciulli, A., Selectivity on-target of bromodomain chemical probes by structure-guided medicinal chemistry and chemical biology. *Future Med Chem* **2016**, 8, 1655-80.

170. Zaware, N.; Zhou, M. M., Bromodomain biology and drug discovery. *Nat. Struct. Mol. Biol.* **2019**, 26, 870-879.

171. Gacias, M.; Gerona-Navarro, G.; Plotnikov, A. N.; Zhang, G.; Zeng, L.; Kaur, J.; Moy, G.; Rusinova, E.; Rodriguez, Y.; Matikainen, B.; Vincek, A.; Joshua, J.; Casaccia, P.; Zhou, M. M., Selective chemical modulation of gene transcription favors oligodendrocyte lineage progression. *Chem. Biol.* **2014**, 21, 841-854.

172. Hugle, M.; Lucas, X.; Weitzel, G.; Ostrovskyi, D.; Breit, B.; Gerhardt, S.; Einsle, O.; Gunther, S.; Wohlwend, D., 4-Acyl Pyrrole Derivatives Yield Novel Vectors for Designing Inhibitors of the Acetyl-Lysine Recognition Site of BRD4(1). *J. Med. Chem.* **2016**, 59, 1518-30.

173. Zhang, G.; Plotnikov, A. N.; Rusinova, E.; Shen, T.; Morohashi, K.; Joshua, J.; Zeng, L.; Mujtaba, S.; Ohlmeyer, M.; Zhou, M. M., Structure-guided design of potent diazobenzene inhibitors for the BET bromodomains. *J. Med. Chem.* **2013**, 56, 9251-64.

174. Law, R. P.; Atkinson, S. J.; Bamborough, P.; Chung, C.-W.; Demont, E. H.; Gordon, L. J.; Lindon, M.; Prinjha, R. K.; Watson, A. J. B.; Hirst, D. J., Discovery of Tetrahydroquinoxalines as Bromodomain and Extra-Terminal Domain (BET) Inhibitors with Selectivity for the Second Bromodomain. *J. Med. Chem.* **2018**, 61, 4317-4334.

175. Picaud, S.; Wells, C.; Felletar, I.; Brotherton, D.; Martin, S.; Savitsky, P.; Diez-Dacal, B.; Philpott, M.; Bountra, C.; Lingard, H.; Fedorov, O.; Muller, S.; Brennan, P. E.; Knapp, S.; Filippakopoulos, P., RVX-208, an inhibitor of BET transcriptional regulators with selectivity for the second bromodomain. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, 110, 19754-9.

176. Baud, M. G.; Lin-Shiao, E.; Zengerle, M.; Tallant, C.; Ciulli, A., New Synthetic Routes to Triazolo-benzodiazepine Analogues: Expanding the Scope of the Bump-and-Hole Approach for Selective Bromo and Extra-Terminal (BET) Bromodomain Inhibition. *J. Med. Chem.* **2016**, 59, 1492-500.

177. Raux, B.; Voitovich, Y.; Derviaux, C.; Lugari, A.; Rebuffet, E.; Milhas, S.; Priet, S.; Roux, T.; Trinquet, E.; Guillemot, J. C.; Knapp, S.; Brunel, J. M.; Fedorov, A. Y.; Collette, Y.; Roche, P.; Betzi, S.; Combes, S.; Morelli, X., Exploring Selective Inhibition of the First Bromodomain of the Human Bromodomain and Extra-terminal Domain (BET) Proteins. *J. Med. Chem.* **2016**, 59, 1634-41.

178. Schrödinger Release 2018-4: LigPrep, S., LLC, New York, NY, 2018.

179. Schrödinger Release 2018-4: Glide, S., LLC, New York, NY, 2020.

180. Humphrey, W.; Dalke, A.; Schulten, K., VMD: Visual molecular dynamics. *J. Mol. Graphics* **1996**, 14, 33-38.

181. Rodriguez, Y.; Gerona-Navarro, G.; Osman, R.; Zhou, M. M., In silico design and molecular basis for the selectivity of Olinone toward the first over the second bromodomain of BRD4. *Proteins* **2020**, 88, 414-430.

182. Geschwindner, S.; Ulander, J.; Johansson, P., Ligand Binding Thermodynamics in Drug Discovery: Still a Hot Tip? *J. Med. Chem.* **2015**, 58, 6321-35.

183. Martin, S. F.; Clements, J. H., Correlating structure and energetics in protein-ligand interactions: paradigms and paradoxes. *Annu. Rev. Biochem* **2013**, 82, 267-93.

184. Chodera, J. D.; Mobley, D. L., Entropy-enthalpy compensation: role and ramifications in biomolecular ligand recognition and design. *Annu Rev Biophys* **2013**, 42, 121-42.

185. Dunitz, J. D., Win some, lose some: enthalpy-entropy compensation in weak intermolecular interactions. *Chem. Biol.* **1995**, 2, 709-712.

186. Nittinger, E.; Gibbons, P.; Eigenbrot, C.; Davies, D. R.; Maurer, B.; Yu, C. L.; Kiefer, J. R.; Kuglstatter, A.; Murray, J.; Ortwine, D. F.; Tang, Y.; Tsui, V., Water molecules in protein–ligand interfaces. Evaluation of software tools and SAR comparison. *J. Comput. Aided Mol. Des.* **2019**, 33, 307-330.

187. Wlodawer, A.; Vondrasek, J., Inhibitors of HIV-1 protease: a major success of structure-assisted drug design. *Annu. Rev. Biophys. Biomol. Struct.* **1998**, 27, 249-84.

188. von Itzstein, M.; Wu, W. Y.; Kok, G. B.; Pegg, M. S.; Dyason, J. C.; Jin, B.; Van Phan, T.; Smythe, M. L.; White, H. F.; Oliver, S. W.; et al., Rational design of potent sialidase-based inhibitors of influenza virus replication. *Nature* **1993**, 363, 418-23.

189. Capdeville, R.; Buchdunger, E.; Zimmermann, J.; Matter, A., Glivec (STI571, imatinib), a rationally developed, targeted anticancer drug. *Nat. Rev. Drug Discov.* **2002**, 1, 493-502.

190. Blum, A.; Bottcher, J.; Dorr, S.; Heine, A.; Klebe, G.; Diederich, W. E., Two solutions for the same problem: multiple binding modes of pyrrolidine-based HIV-1 protease inhibitors. *J. Mol. Biol.* **2011**, 410, 745-55.

191. Mangani, S.; Cancian, L.; Leone, R.; Pozzi, C.; Lazzari, S.; Luciani, R.; Ferrari, S.; Costi, M. P., Identification of the binding modes of N-phenylphthalimides inhibiting bacterial thymidylate synthase through X-ray crystallography screening. *J. Med. Chem.* **2011**, 54, 5454-67.

192. Daura, X.; Gademann, K.; Jaun, B.; Seebach, D.; van Gunsteren, W. F.; Mark, A. E., Peptide Folding: When Simulation Meets Experiment. *Angew. Chem. Int. Ed.* **1999**, 38, 236-240.

193. Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E., GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, 1-2, 19-25.

194. Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G., PLUMED 2: New feathers for an old bird. *Comput. Phys. Commun.* **2014**, 185, 604-613.

195. Limongelli, V.; Bonomi, M.; Parrinello, M., Funnel metadynamics as accurate binding free-energy method. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, 110, 6358-63.

196. Eyal, E.; Gerzon, S.; Potapov, V.; Edelman, M.; Sobolev, V., The limit of accuracy of protein modeling: influence of crystal packing on protein structure. *J. Mol. Biol.* **2005**, 351, 431-42.

197. Aldeghi, M.; Ross, G. A.; Bodkin, M. J.; Essex, J. W.; Knapp, S.; Biggin, P. C., Large-scale analysis of water stability in bromodomain binding pockets with grand canonical Monte Carlo. *Commun Chem* **2018**, 1.

198. Huang, D.; Rossini, E.; Steiner, S.; Caflisch, A., Structured Water Molecules in the Binding Site of Bromodomains Can Be Displaced by Cosolvent. *ChemMedChem* **2014**, 9, 573-579.

199. Deng, Z.; Chuaqui, C.; Singh, J., Structural interaction fingerprint (SIFt): a novel method for analyzing three-dimensional protein-ligand binding interactions. *J. Med. Chem.* **2004**, 47, 337-44.

200. Heinzelmann, G.; Henriksen, N. M.; Gilson, M. K., Attach-Pull-Release Calculations of Ligand Binding and Conformational Changes on the First BRD4 Bromodomain. *J. Chem. Theory Comput.* **2017**, 13, 3260-3275.

201. Kuhnert, M.; Diederich, W. E., Structure-Based Drug Design in Medicinal Chemistry: The Devil is in the Detail. *Synlett* **2016**, 27, 641-649.

Appendix A Chapter 4

MD Simulations of 16 with HIE, HID and HIP

The His437 in the BD2 domain has been modelled with the HIE tautomeric state (Figure 5-14). This seems to be the preferred state when the histidine is in the closed conformation as explained in Section 5.6.2 (Figure 5-14). When the histidine is in an open conformation and the ligand is present, the same protonation state was modelled (HIE) but it is possible that the histidine residue could exist in a number of different tautomeric (HIE, HID and HIP) and rotameric (flipped HIE, flipped HID, flipped HIP) states at physiological pH (Figure A. 1).

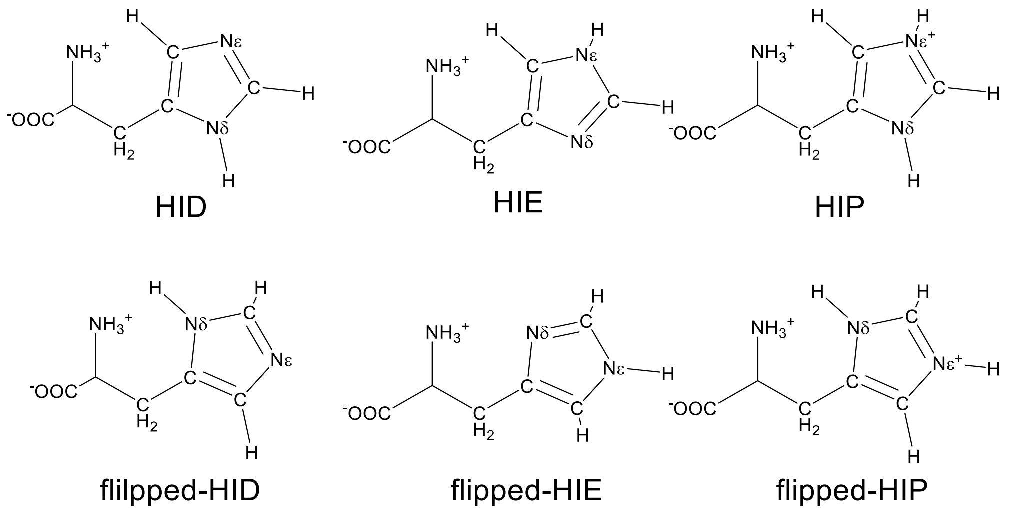


Figure A. 1 The histidine residue can exist in six different protonation and rotameric states. The Nδ and ε are marked.

The MD simulations using the same procedure as described in the main text (Section 5.5.3) were run for compound **16** with the three different protonation states, HIE, HID and HIP of the histidine residue (His437). The ligand RMSD and the distance between the piperazine nitrogen of the ligand and the Nδ of the histidine residue were monitored during the course of the 3 replicas (Figure A. 2).

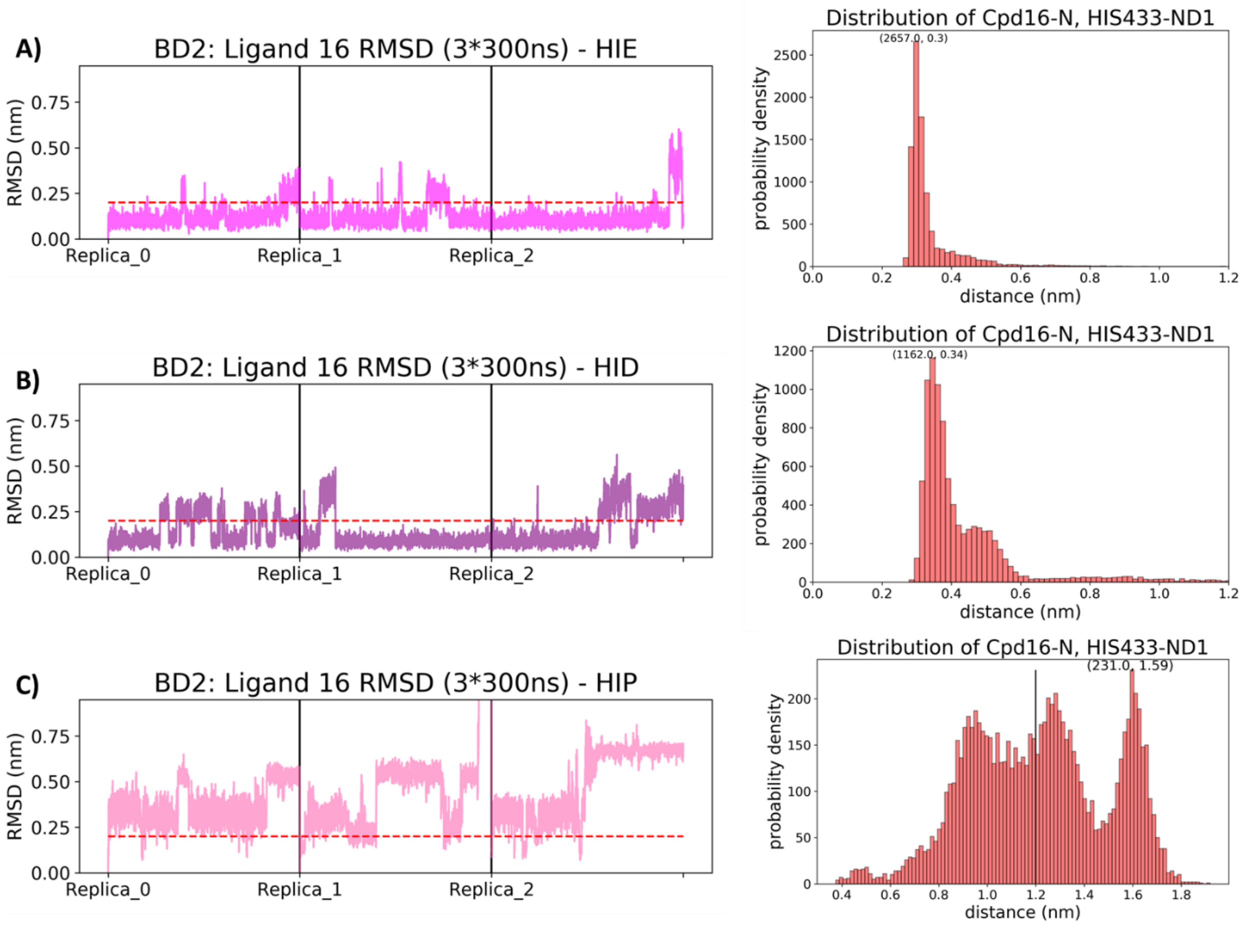


Figure A. 2 On the right, RMSD of the ligand heavy atoms of **16** during the course of the 3 MD replicas when His437 is modelled in the HIE (A), HID (B) and HIP (C) tautomeric states. On the left, distance distribution between the protonated nitrogen of the piperidine ring in the ligand and the Nδ of the His347. The distance distribution in A and B arrives at a maximun of 1.2 nm whereas in C it is up to 1.8 nm. A black vertical line in C indicates the 1.2 nm threshold observed in A and B.

It is reasonable to assume that His437 is best modelled in the HIE tautomeric state in lieu of HID and HIP but this is not clearly explicated by the analysis carried out int the MD simulations. The ligand is either trapped in the starting conformation with the His347 (HIE) or it appears to quickly loose its original conformation to escape from the two positively charged moieties that would otherwise be unfavourable. Therefore, there is scope for further studies with more advanced techniques aimed to identify which would be the most favourable state of the histidine residue when in the open conformation. In summary, it is believed that the insertion of a larger group in the region close to the histidine is unfavourable because of both lack of flexibility and poor shape complementarity with the binding site.

Summary of interactions of 13 and 16 in BD1 and BD2

The interactions in Figure A. 3 and Figure A. 4 were monitored as follow:

Hydrogen bonds: distance between the donor and acceptor atoms ≤ 2.5; a donor angle ≥ 120° between the donor-hydrogen-acceptor atoms and an acceptor angle of ≥ 90° between the hydrogen-acceptor-bonded_atom atoms.

Hydrophobic contacts: fall into three subtypes π-cation, π-π and other non-specific interactions. Generally involve a hydrophobic amino acid and an aromatic or aliphatic group of the ligand and it also include the π-Cation interactions. The criteria are π-cation – aromatic and charged group within a distance of 4.5 Å; π-π – two aromatic groups stacked face-to-face or face-to-edge; other – a non-specific hydrophobic sidechain within 3.6 Å of a ligand's aromatic or aliphatic carbons.

Ionic interactions: between two oppositely charged atoms that are within 3.7 Å of each other and do not involve hydrogen bond.

Water bridges: hydrogen-bonded protein-ligand interactions mediated by a water molecule. The criteria are: distance of 2.8 Å between donor and acceptor atoms; a donor angle of ≥ 110° between the donor-hydrogen-acceptor atoms and an acceptor angle of ≥ 90° between the hydrogen-acceptor-bonded_atom atoms.

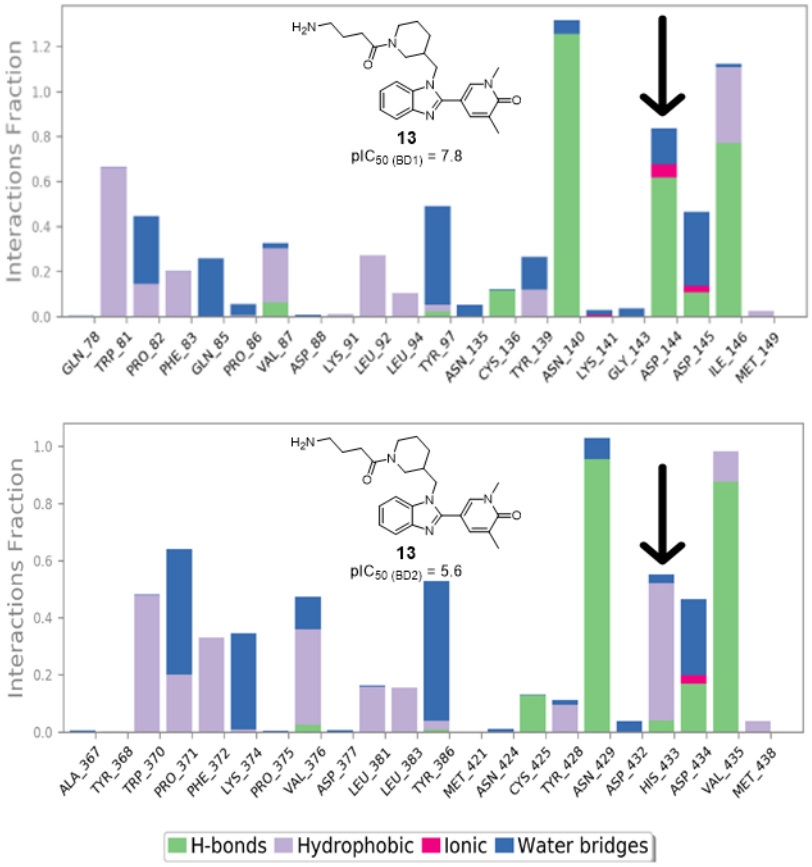


Figure A. 3 Interactions between **13** and the BD1 domain (upper plot) or BD2 domain (lower plot) during the course of the MD simulations (total of 3 replicas). The hydrogen bond formed in BD1 with Asp144 is not formed in BD2 (black arrows).

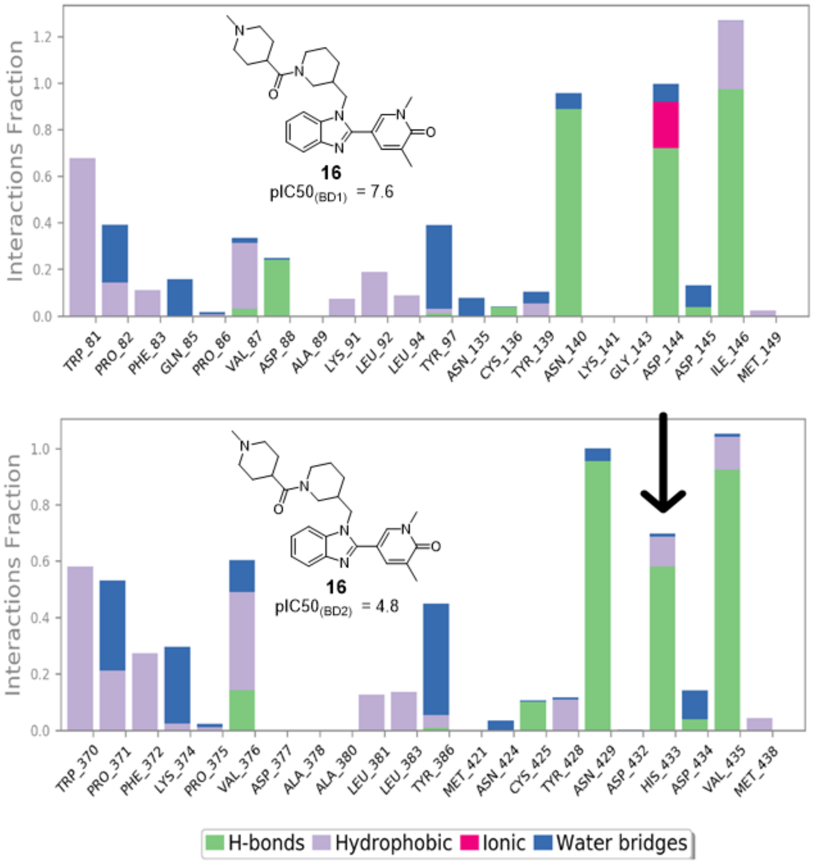


Figure A. 4 Interactions between **16** and the BD1 domain (upper plot) or BD2 domain (lower plot) during the course of the MD simulations (total of 3 replicas). In BD2 a hydrogen bond between the ligand and His437 is identified (black arrow).

Appendix B Trp81: MD simulations

During the MD simulations of BM1, Trp81 is flexible and adapts several conformations. In replica 3, the conformation observed for BM1 is the most populated, in fact this is the only replica in which the ligand appears to be stable (Section 6.5.2.1.1 infra). Moreover, the conformation observed in the most populated cluster is observed in the replica in which the ligand is adopting a BM2-like conformation, replica 0 and replica 1.

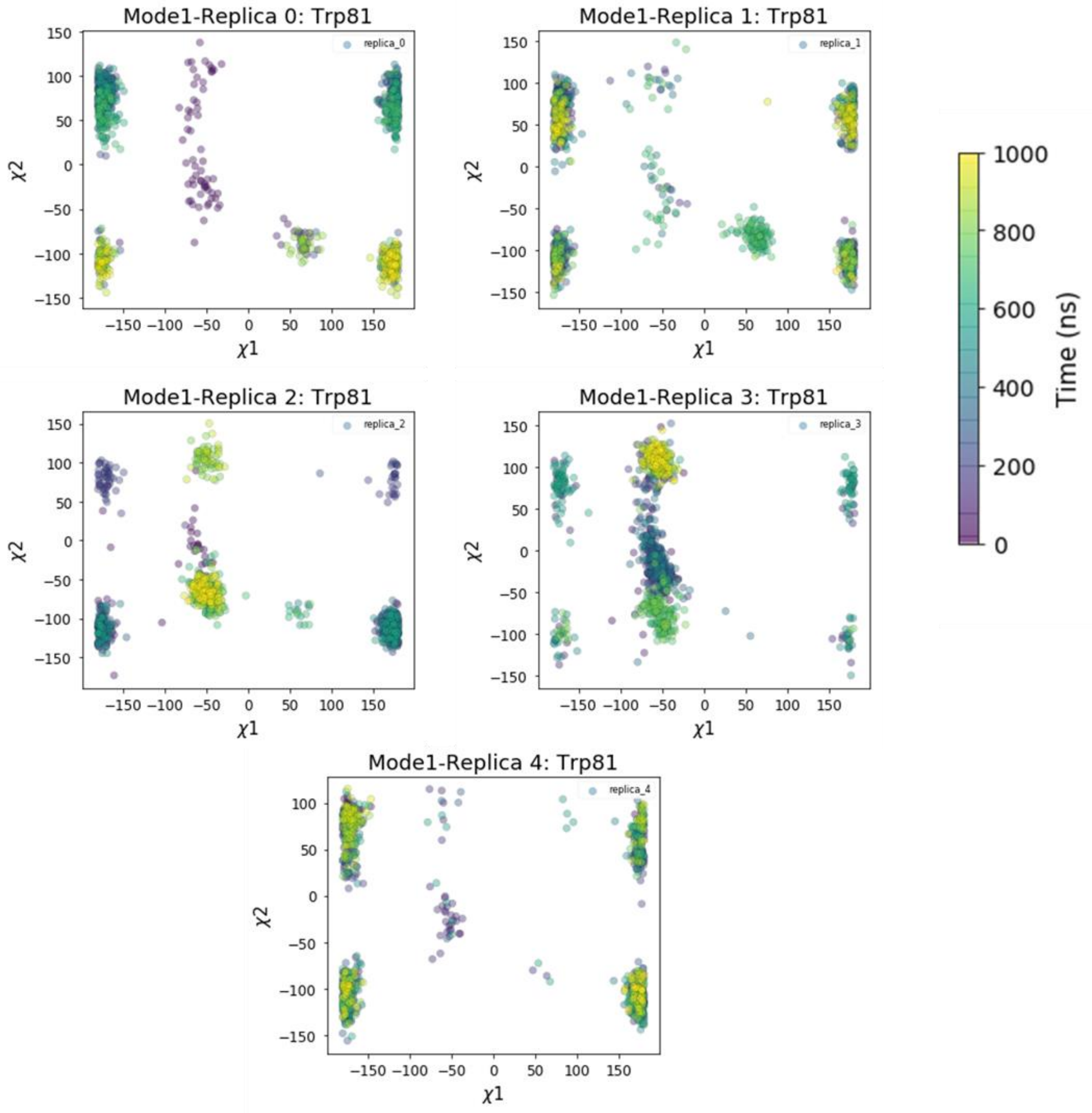


Figure A. 5 χ_1 and χ_2 conformations of Trp81 during the course of the MD simulations started from BM1.

During the MD simulations of BM2, Trp81 is also showing flexibility and the starting conformation is not maintained during the course of the simulations. This enhanced flexibility compared to the conformations explored in the BM1 simulations, could be attributed to the solvent exposed part of the ligand and/or because the residue in the crystal packing are not present. The movement of this portion of the ligand is correlated with the movement of Trp81. The conformations of Trp81 observed most frequently in the BM2 simulations have -180° < χ_1 < -160° and either -150° < χ_2 < -70° or 30° < χ_2 < 110°.

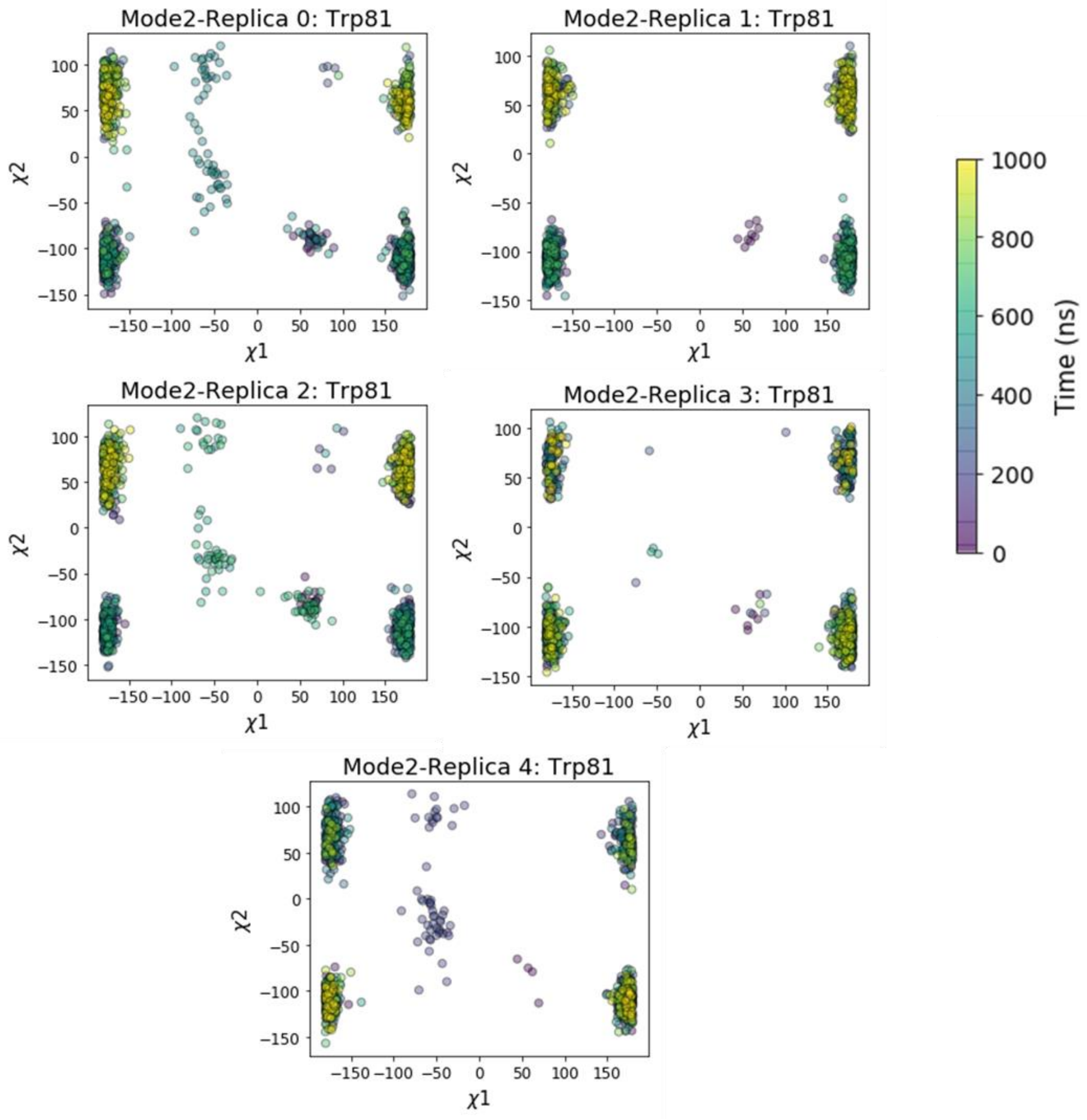


Figure A. 6 χ_1 and χ_2 conformations of Trp81 during the course of the MD simulations started from BM2.

Appendix C Additional FM simulations

Two additional FM simulations were run from BM1 and BM2 by defining the funnel as a sigmoid function as explained in Section 6.2.2 with funnel width *h*= 1.8 nm instead of 1.2 nm and the steepness of the function s= 3 nm instead of 5 nm. In this way, the funnel was covering a bigger surface around the ligand-binding domain. The FES was studied by using the same CVs as the one reported in the main text (see Section 6.5.2.2). A total of 2.2 μs and 2.1 μs were accumulated but the obtained FES did not reach convergence.

For this reason, it was decided to carry out the simulations as reported in the text with a smaller funnel and applying the PT-MetaD scheme with the idea to increase the exploration of the bound and unbound state. The resulting FES of the FM started from BM1 and BM2 are reported in Figure A. 7 and Figure A. 8.

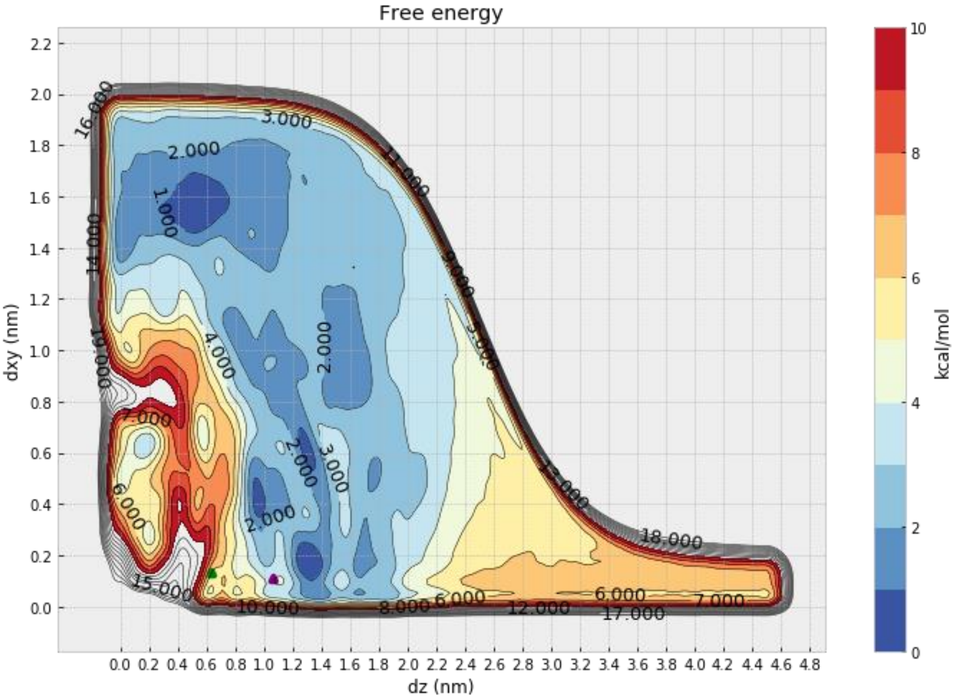


Figure A. 7 FES landscape of the BM1 using the dxy and dz as CVs.

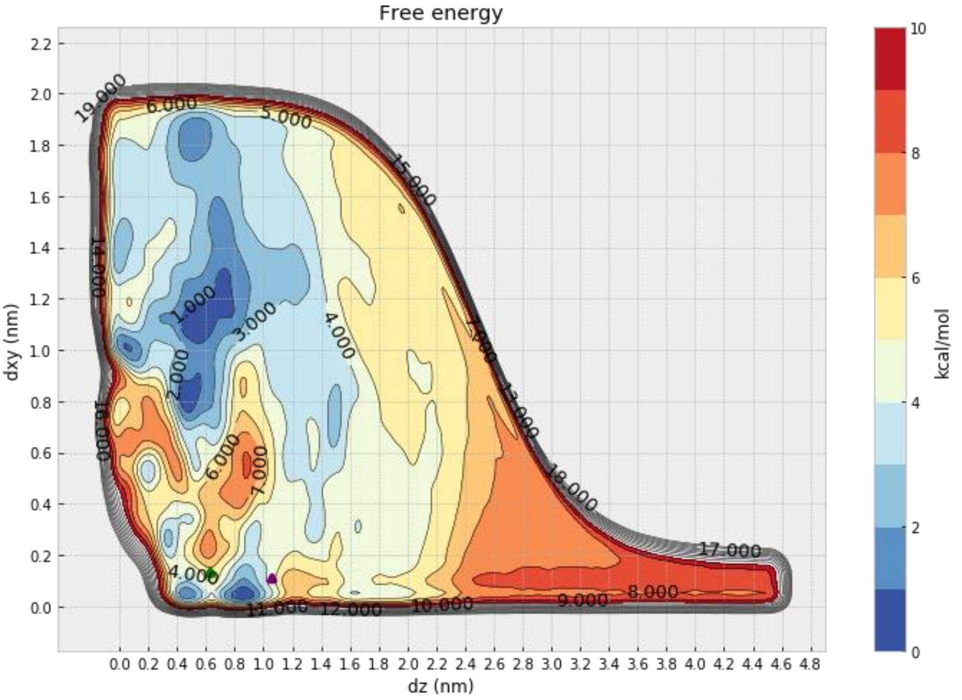


Figure A. 8 FES landscape of the BM2 using the dxy and dz as CVs.

The two FES are different if compared with each other. In the BM1-FES, the BM2 position is barely explored and the free energy minima are populated all around the protein and not in the binding site.

In the BM2-FES, the free energy minima are close to the crystallographic poses but, as reported from the absolute binding free energy, there are not enough events where the ligand moves from bound to unbound state and vice versa.

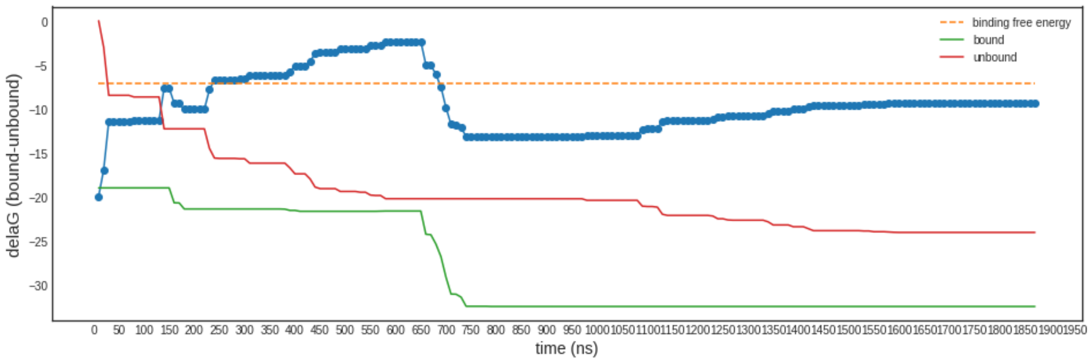


Figure A. 9 Evolution of the computed binding free energy over the time of the simulation in comparison with experimental value (dotted orange line).